2013-06-19

Computational Techniques for Public Health Surveillance

Scott H. Burton
Brigham Young University - Provo

Computational Techniques for Public Health Surveillance

Scott H. Burton

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Christophe G. Giraud-Carrier, Chair
Dan A. Ventura
Carl L. Hanson
William A. Barrett
Charles D. Knutson

Department of Computer Science

Brigham Young University

June 2013

ABSTRACT

Computational Techniques for Public Health Surveillance

Scott H. Burton
Department of Computer Science, BYU
Doctor of Philosophy

Public health surveillance is a critical part of understanding, and ultimately influencing, health behaviors. Traditional methods, such as questionnaires and focus groups have significant limitations including cost, delay, and size. Online social media data has the potential to overcome many of the challenges of traditional methods, but its exploitation is not trivial. We develop and apply computational techniques to enable public health surveillance in novel ways and on a larger scale than currently performed.

In this regard, we present techniques for mining the *who*, *what*, and *where* of public health surveillance in social media. We show how computational methods can identify health content and conversations in social media, and that people do in fact speak openly about health topics, including those that might be considered private. In addition, we demonstrate how location information can be mined and used to study distributions of various conditions. Finally, and perhaps most importantly, we develop techniques to identify and leverage pertinent social network relationships in public health surveillance. We demonstrate each of these approaches in large data sets of actual social networks spanning blogs, micro-blogs, and video-sharing sites.

Keywords: Public Health Surveillance, Community Mining, Social Media

ACKNOWLEDGMENTS

I would like to thank the faculty and students at Brigham Young University for all of their help. I am especially grateful to my advisor, Christophe Giraud-Carrier, who has been an incredible influence to me, personally and academically, and without whom I could not have produced this research. Finally, I would like to thank my amazing wife, Amber, for her continued support and encouragement.

Table of Contents

List of Figures

List of Tables

Part I

Introduction

Health is a topic of individual and global interest, as each year, millions of lives are affected by health challenges and death. In addition, the total health expenditures in the United States in 2010 were estimated at $2.6 trillion or 17.9% of the nation's Gross Domestic Product [170]. Public health surveillance is a key to understanding, and ultimately improving, health. Defined by the World Health Organization as "the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice," [249] public health surveillance is a core component of areas such as epidemiology, health promotion, substance abuse prevention and treatment, and public policy. Tracking the spread of diseases and observing behaviors, attitudes, and beliefs of individuals regarding health issues is critical to influencing health behaviors and measuring relevant outcomes, making "monitoring health status to identify community health problems" one of the 10 Essential Public Health Services that comprise the framework for the National Public Health Performance Standards Program [173].

Traditional methods of health surveillance include quantitative and qualitative techniques such as questionnaires, focus groups, and clinical trials, as well as health department laboratory reporting. These methods have many strengths and have been successful in observing elements of public health, but they have significant limitations. For example, outbreak data from health department labs, by nature, lags weeks behind symptoms' on-

set [198, 248], and can be costly to determine. Similarly, there are large costs associated with running effective questionnaires and trials, which, in many instances, necessitates small sample sizes, usually composed of isolated individuals, as opposed to studying the context of their associations. In addition, there are delays to when people answer questions, requiring responses about previous events, as well as delays to the availability of results. Furthermore, there may be differences in reported versus actual behavior, whether intentional or not, such as people thinking they behave differently than they actually do, or the case of the Hawthorne effect [1], where the mere presence of the investigators causes unintended influence on the responses people give or the way they act, because they know they are being observed.

The recent explosion of popularity of online social media provides unprecedented opportunities for public health surveillance and the exploitation of social media data has the potential to address many of the challenges of traditional methods. For example, the cost of downloading and analyzing data is significantly cheaper than administering physical tests or trials. Delays are eliminated because people post in real-time about events as they occur, and computational methods enable near real-time analysis. Also, because social media captures natural interactions between people, it can reveal true feelings and behaviors. Furthermore, data is available from hundreds of millions of people throughout the world, including data not available through traditional channels such as from developing countries [31]. In addition to the textual content, social media is also a rich source of relational data that provides a view of people in the context of friendships, communities, and other social structures.

While the possibilities are very attractive, exploiting online social media data for public health surveillance is not trivial and requires expertise from both health and computer sciences. In this regard, we have established the Computational Health Science Research Group at Brigham Young University[1] to bring together researchers from health and computer science to address these problems from a trans-disciplinary perspective. Through this collaboration, as shown in this dissertation, we have been able to make contributions to both fields.

[1] `http://dml.cs.byu.edu/chs`

Organization of the Dissertation

In this dissertation we develop and apply computational methods for mining the *who*, *what*, and *where* of public health surveillance in social media. This document is divided into two main parts: Part II demonstrates mining the *what* and *where*, and Part III focuses on the *who*.

Part II is composed of four papers, comprising Chapters 1–4. In Chapter 1 (published in the *Journal of Medical Internet Research*), we demonstrate the different types of location information that can be mined from Twitter. We show that there are significant differences between the amount of users who say they provide GPS data (according to a questionnaire) and the number that actually do, which provides further justification for our empirical approach to observing behavior. In addition, we show that the proportion of tweets per state strongly correlates with the proportion of the census population in that state, establishing that while Twitter users may not be a representative sample of the population as a whole, their distribution is not biased geographically between states, and is therefore valid for geographical surveillance (e.g., epidemiology).

Suicide is a significant problem in the United States, where nearly the same number of people die from suicide as from breast cancer, and suicide is the third leading cause of death among adolescents [8, 41]. Unfortunately, this is also a local problem, as Utah consistently ranks in the top 10 among states for highest rates of suicide. Social media could be an ideal setting to observe suicide risk factors as people may exhibit them to peers before contacting professionals. Building on the location results of Chapter 1, in Chapter 2 (*in submission*), we explore the use of social media for suicide detection and intervention. In particular, we discover discussion of suicide risk factors in the United States and demonstrate that the *discussion* of these factors per state correlates with actual rates of suicide.

Prescription drug abuse is another growing problem nationally where more people now die from prescription drug overdoses than from car accidents, and unfortunately Utah is also consistently among the states having the worst problem [211]. In addition, it is an

interesting research question in its own regard to determine if potentially private topics, such as prescription drug abuse, are discussed in public social media channels. In Chapter 3 (published in the *Journal of Medical Internet Research*), we identify discussion of prescription drug abuse, specifically focusing on abuse of the medication *Adderall*. We show that discussion of Adderall is often related to alternative motives (e.g., as a study aid), and that it is discussed with abnormally high frequency during traditional college final exam periods. Using computational methods, we demonstrate that students that discuss Adderall can be connected to clusters of nearby universities, highlighting regions that have the highest incidence. Being able to simultaneously study students from different universities highlights a key advantage of social media data, compared to using focus groups or questionnaires which are often limited to one or a small group of locations to study. In addition, this study demonstrates that social media users do, in fact, openly discuss topics that some may consider taboo or private.

In Chapter 4 (published in *Network Modeling in Health Informatics and Bioinformatics*), we develop a process for identifying health advice-seeking questions and discovering their responses. We show that users with larger numbers of followers (an evidence of social capital) can leverage their networks to receive advice more frequently and more quickly. Demonstrating that users seek and receive health advice from peers provides further validation for a lay health advisor model, wherein users can become advisors to their peers.

Part III builds on the value of the network expressed in Chapter 4 to focus on the *social* component of social media, and demonstrates how computational techniques can leverage the inherent relational structure of social media for public health surveillance. Indeed, social network analysis and mining of large networks is one of the major strengths offered by the computational side of the computational health science collaboration. This part is composed of four papers (comprising Chapters 5–8).

In Chapter 5 (published in *Proceedings of the 2nd ACM International Health Informatics Symposium*), we demonstrate ways of mining relationships in YouTube. While on the surface YouTube is known as a site to post and watch online videos, it is surprisingly rich in

relational structures existing among authors, subscribers, commenters, and even among videos themselves. We seek to identify communities using these relations, but find that existing community discovery methods include many nodes that are unrelated. We introduce a new community mining algorithm, and demonstrate its effectiveness in identifying communities of users and videos of interest to public health researchers.

As Chapter 5 exposed the need for algorithms that can effectively identify a set of nodes surrounding an initial query set, in Chapter 6 we further explore this area to develop a more robust algorithm for discovering social circles in directed graphs (*in submission*). In this work, we show that while there are a number of community mining algorithms, many are not suitable for use in the local context, where the entire graph cannot be feasibly known *a priori*. Furthermore, those algorithms that are designed for local community mining may discover sets *containing* initial query nodes, but they do not adequately find the social circle *surrounding* the query set. We introduce a new local algorithm to identify these social circles in directed graphs and demonstrate its effectiveness on standard benchmarks, large networks with ground-truth communities, and real-world social networks.

Using the social circle discovery algorithm of Chapter 6, in Chapter 7 (to appear in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*), we discover communities of mothers in social media. Communities of mothers are particularly relevant in this context because of the active nature of mothers in social media and because of the important role of mothers in the health decisions of the home [57]. We discover social circles of mothers in Twitter and the blogosphere and link accounts across platforms to observe similarities and differences. Using both directed and unsupervised methods, we find that mothers do indeed discuss many health topics, and yet there are also important topics that are seldom mentioned. The fact that mothers do openly discuss health matters suggests that social media could be an excellent mechanism to help raise awareness of those that are not discussed. We also identify implicit affinity networks

defined by the latent topics of interest to each mother, and highlight the opportunity for additional connections among mothers of similar interests.

Where Chapter 3 identifies that prescription drug abuse is discussed in social media, and Chapter 7 establishes the health discussion that exists among users' networks, in Chapter 8 (*in submission*), we discover social circles around likely prescription drug abusers to identify the social engagement among them. We show that prescription drug abuse is not discussed in isolation, but rather that users connect with others who discuss prescription drug abuse, which confirms theoretical expectation and has health implications in terms of social norms, where peer "support" can be enabling to unhealthy behavior. We additionally show that the level of social engagement around prescription drugs varies across social circles and higher levels of engagement about prescription drug abuse correlates with higher levels of abusiveness.

Finally, Part IV concludes the dissertation and highlights future work.

Development of the Experimental Framework

An important contribution of this dissertation is the development of an experimental framework that will continue to enable future research in these areas. We have written an extensible library in Microsoft C#.NET to handle collection, storage, and analysis of the social media data, which is composed of the following projects:

- *Twitter Miner.* Twitter provides a robust API to access both tweets and user profile data. We have written a library to interface with their Streaming API (to obtain Tweets as they occur), Search API (to find past tweets matching certain criteria, such as those mentioning a particular user), and their detailed REST API (to get detailed user information, and to get historical tweets from a user). In addition, we have written classes to wrap these functions so that properties of "users" and their "profiles" automatically load data from cache or the API appropriately. Also, we have written monitors to send email messages and restart services if problems occur during streaming.

- *Blog Miner.* We have written a custom web crawler to identify links to RSS Feeds, and download their blog entries via the unofficial Google Reader API. This tool enables us to identify neighboring blogs, access their content, and build communities.
- *YouTube Miner.* Using the Google's official YouTube API and their provided .NET wrapper, we have further wrapped YouTube objects to easily obtain user and video information, including relations, and add them to expanding communities.
- *Community Discovery.* We have developed a custom implementation for weighted, directed graphs that connects nodes of a generic type. These nodes can implement interfaces to hold additional context-specific properties and to discover their own neighbors in the appropriate context (e.g., YouTube video nodes discover their related list, Twitter user nodes discover friends/followers, etc.). These graphs can be serialized to and deserialized from the graph exchange xml format (.gexf) used by open source visualization tools such as Gephi. In addition, we have implemented several community mining algorithms (existing and our own) that can be applied to these different graphs.
- *Social Media Utilities and Experiments.* In addition to the platform specific projects, we have also written utilities that are common to all of them, and can link nodes across them. In particular, we have written methods to categorize entries according to any or all keywords for a category, including exclusion terms, and subcategories. We have also written methods to interface with Google and Yahoo Maps APIs to handle GPS lookups and reverse-lookups, and have developed a sophisticated caching mechanism to cache the social media data to the file system or a database.

These projects interface with a Microsoft SQL Server database, using the .NET Entity Framework for object-relational mapping.

Part II

Validating Social Media for Surveillance: Exploring the "What" and "Where"

S.H. Burton, K.W. Tanner, C.G. Giraud-Carrier, J.H. West, and M.D. Barnes. "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. *Journal of Medical Internet Research* **14**(6):e156, 2012.

J. Jashinsky, S. Burton, C.L. Hanson, J. West, C. Giraud-Carrier, M. Barnes, and T. Argyle. Tracking Suicide Risk Factors through Twitter in the U.S. *In Submission*, 2013.

C.L. Hanson, S. Burton, C. Giraud-Carrier, J. West, M. Barnes, and B. Hansen. Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) Among College Students. *Journal of Medical Internet Research* **15**(4):e62, 2013.

S.H. Burton, K.W. Tanner, and C.G. Giraud-Carrier. Leveraging Social Networks for Anytime-Anyplace Health Information. *Network Modeling Analysis in Health Informatics and Bioinformatics* **1**(4):173-181, 2012.

Chapter 1

"Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information

Abstract

- **Background**: Twitter provides various types of location data, including exact Global Positioning System (GPS) coordinates, which could be used for infoveillance and infodemiology (ie, the study and monitoring of online health information), health communication, and interventions. Despite its potential, Twitter location information is not well understood or well documented, limiting its public health utility.
- **Objective**: The objective of this study was to document and describe the various types of location information available in Twitter. The different types of location data that can be ascertained from Twitter users are described. This information is key to informing future research on the availability, usability, and limitations of such location data.
- **Methods**: Location data was gathered directly from Twitter using its application programming interface (API). The maximum tweets allowed by Twitter were gathered (1% of the total tweets) over 2 separate weeks in October and November 2011. The final dataset consisted of 23.8 million tweets from 9.5 million unique users. Frequencies for each of the location options were calculated to determine the prevalence of the various location data options by region of the world, time zone, and state within the United States. Data from the US Census Bureau were also compiled to determine population proportions in each state, and Pearson correlation coefficients were used to compare

each state's population with the number of Twitter users who enable the GPS location option.

- **Results**: The GPS location data could be ascertained for 2.02% of tweets and 2.70% of unique users. Using a simple text-matching approach, 17.13% of user profiles in the 4 continental US time zones were able to be used to determine the user's city and state. Agreement between GPS data and data from the text-matching approach was high (87.69%). Furthermore, there was a significant correlation between the number of Twitter users per state and the 2010 US Census state populations ($r \geq 0.97$, $P < .001$).
- **Conclusions**: Health researchers exploring ways to use Twitter data for disease surveillance should be aware that the majority of tweets are not currently associated with an identifiable geographic location. Location can be identified for approximately 4 times the number of tweets using a straightforward text-matching process compared to using the GPS location information available in Twitter. Given the strong correlation between both data gathering methods, future research may consider using more qualitative approaches with higher yields, such as text mining, to acquire information about Twitter users' geographical location.

1.1 Introduction

People's daily use of technology creates "digital breadcrumbstiny records of [their] daily experiences" that, when mined and analyzed, can provide insight into health behavior and health outcomes [196]. Traditional behavioral assessments rely on self-report or observation, but increased use of mobile communication devices linked to the Internet and social media applications (apps) are creating unprecedented opportunities for collecting real-time health data and delivering health innovations. For example, mHealth represents a new form of health care delivery and treatment where patients are able to interact with their health care providers through mobile devicesproviding additional "breadcrumbs" for studying/mining health behaviors and health outcomes [67].

Although some researchers have expressed concerns about the use of social media in public health [65], an increasing number of researchers welcome the novel opportunities offered by social media to complement (and partially replace in some cases) existing practices in public health and health communication [72, 107, 160]. A number of recent studies have demonstrated the value of online information for understanding public health problems and their determinants in areas as diverse as influenza and cholera outbreaks [49, 188], tobacco-related issues [17, 34, 201], problem drinking [244], dental pain [96], breastfeeding [243], and others [192, 214]. This new real-time observation and analysis of user-generated health content in social media has given rise to the terms infoveillance and infodemiology (the study and monitoring of online health information) [70].

Social media connect a wide variety of individuals around many topics and provide a new way for them to share information, reach out, and exchange ideas. As recently editorialized by Ratzan [205], "This change to the way people learn, think, and communicate has revolutionized the context in which health information...needs to be communicated." Not only is the context different, so is the sheer volume and scale. Millions of individuals worldwide can be reached almost instantaneously with textual, pictorial, and video messages that could alter health behaviors. Additionally, the distribution of social media usage suggests that health disparities may be reduced, and traditionally underrepresented groups and low-income populations may be reached more effectively [83].

As a kind of "listening ear" to the conversations of the world, social media enable health surveillance in completely novel ways. Whereas researchers have relied on questionnaires and focus groups to understand the opinions and behaviors of the public in the past, by using social media they can now observe Internet postings about users' attitudes and behaviors, many of which can be accessed in real time. These approaches are optimistic because they are typically less expensive and may better reflect the real-life context of behavioral indicators as part of everyday living than traditional assessments of health behaviors. Further, online surveillance enables researchers to study trends as they happen, removing the delay that

often arises from designing, administering, and collecting questionnaire-style responses. In addition, by observing users as they interact naturally with one another and their environment, researchers can study true feelings and avoid the Hawthorne effect where the investigator's presence can cause unintended influence [1]. Thus, these social media channels "are quickly becoming dominant sources of information on emerging diseases" [29].

In addition to using social media for surveillance, these technologies could also be harnessed for health communication and intervention. Although still largely underutilized, social media provide the ability to communicate with people in a completely tailored manner, which has been shown to significantly improve the chances of affecting actual behavior change [129, 226, 227]. Furthermore, the real-time nature of the data and the location information of social media provide the opportunity for truly "right time, right place" communication where a person receives the message exactly when and where it is needed. Consider, for example, the possibilities of direct intervention with a potential drunk driver before leaving a party or of a diet reminder reaching a person as they walk into a fast food restaurant. Identifying—and reacting to—health needs in such a timely manner is consistent with Patrick et al. [191] and Heron and Smyth [99] who referred to this process as "ecological momentary interventions" or as Intille et al. [102] call it, "just-in-time."

Despite its promise, location in social media is not well understood or well documented. Although proponents of research using social media have pointed to the geolocation information provided by many platforms, such as Twitter, as a means of pinpointing the exact location of users [132], others have cautioned that location information may be underspecified and that location "based on user-identified location or the time zone" could be of questionable quality [65]. The exact Global Positioning System (GPS) coordinates available in some social media platforms could help mitigate this risk because they are direct measures and more difficult to misrepresent. However, unless GPS use is widespread, this does not address the problem of underspecification. Until research is conducted to assess location availability,

usability, and the limitations of this data, health practitioners may have limited capacity to observe time- and place-based interventions for determining risks or health conditions.

The objective of this study is to fill this gap in our understanding of location information in social media, especially as it relates to Twitter. The major contribution of this work is to present the different types of location that can be ascertained from Twitter users and to document the prevalence of each type in an attempt at informing future infoveillance, infodemiology, and health communication research of the availability, usability, and limitations of such location data.

1.2 Methods

Twitter is a social network in which users post status updates, or tweets, that are restricted to 140 characters in length. Users can "follow" others to be notified of their updates, but tweets are also generally available to the public. Because of the public nature of the tweets, users do not have any expectation of privacy, so researchers may openly observe the content. Additionally, Twitter provides a rich application programming interface (API) that enables programmatic searching and retrieval of the data. Twitter users tend to be young and affluent [217]; therefore, one could conclude that they are not representative of entire populations. However, this should not diminish perceptions of Twitter's utility as a public health tool because it may be an appropriate mechanism for studying attitudes and behaviors of the demographic most represented among its users (ie, young and affluent individuals).

1.2.1 Location Indicators in Twitter

Twitter users provide varying degrees of information about their thoughts, attitudes, and behaviors in their profile description and through their tweets. Similarly, they may or may not provide information about their location. When they wish to provide location information, Twitter users have 4 options: (1) exact GPS coordinates associated with a tweet, (2) GPS coordinates of a place (eg, a city or metropolitan region), (3) free-text location information

listed in the public profile description, and (4) time zone associated with the user account. Options 1 and 2 are combined into a single setting, the Twitter Location feature, which is disabled by default so that a user must opt-in to use it. Further details about each option and its functionality follow.

Many users post to Twitter from smartphones or other GPS-enabled devices, and have the ability to broadcast their exact GPS coordinates alongside the text of their tweet. This setting is disabled by default, but when used, this GPS information provides reliable and accurate data about a Twitter user's location.

Users posting from their computers and other devices without GPS via the Twitter website can still broadcast their location by providing a GPS "place." This place is defined by a bounding box of GPS coordinates and often refers to a city or a metropolitan area. This place is inferred by Web browsers, such as Firefox and Google Chrome, and on other browsers through the use of extensions or add-ons. In the case of a GPS-enabled device, this place can be determined directly by the GPS coordinates.

When users create accounts on Twitter they can fill out a public profile that includes personal information, such as their name, website, bio, picture, and location. Location is an optional text field in which users can enter anything they want. Many users provide their geographical position, such as a city and state/country, but many opt to specify something humorous (eg, "somewhere in my imagination :)" or "a cube world in Minecraft"), sarcastic (eg, "in yhur [bleep!!!] face" or "Here...obvious!"), or just leave the field blank. The free-text nature of the user-specified location field poses serious challenges. First and most obvious, humorous, sarcastic, and missing entries do not correspond to any identifiable physical location. Second, the entry requires some amount of text processing to correct spelling errors, interpret "textese" and emoticons, and handle abbreviations. Third, the information may be incomplete or ambiguous, such as when a city name is given, but no state or country is provided. Finally, even if the location field can be recognized as a specific location, it is still

possible that users chose to provide a location different from where they actually are or that the information is not up-to-date.

Twitter automatically infers a time zone when a user account is created, probably from the local time on the user's computer or device, and selects it for the user by default. The user can subsequently change this default value, if desired. Although time zones do not denote specific locations, they can still be used to distinguish between major world regions, such as North America and Europe, or the East and West Coasts of the United States. This time zone information could also be helpful in resolving ambiguous city names from profile descriptions.

In addition to these mechanisms supported directly by Twitter, users can also provide location context indirectly in the text of their tweets (eg, "My plane just landed at JFK") or through third-party applications, such as foursquare (https://foursquare.com/). In some cases, these applications will broadcast GPS coordinates via the standard Twitter mechanisms. In other cases, they may broadcast text or links that would point users elsewhere to see the location. For clarity and to avoid the bias of catering to specific conventions or applications, this study focuses exclusively on the mechanisms supported directly and explicitly by Twitter.

1.2.2 Data Collection Methodology

The Twitter streaming API provides the ability to receive a portion of the real-time stream of all tweets. This stream can be filtered by certain criteria, such as keywords or a bounding box of GPS coordinates. If no filtering criteria are used (or if the criteria are too general and more than 1% of the tweet stream would be retrieved), the streaming API will return 1% of the total tweets sampled by taking every 100th tweet. As of June 2011 (3 months prior to our data collection), Twitter estimated that approximately 200 million tweets were posted every day [235], resulting in a daily sample of approximately 2 million tweets when using the streaming API with no filter.

Using the Twitter streaming API, we observed the stream of tweets for 2 weeks: October 1-7 and November 7-14, 2011 (approximately 6 hours of the early morning on November 14 were not observed due to a server error). We did not find significant differences between the data of the 2 weeks; therefore, the results presented here are an aggregation of the 2 weeks' data. By not applying a filter, we received the maximum random sample of 1% of all tweets, yielding a total of 23.8 million tweets posted by 9.5 million unique users. Additionally, because we did not use a filter, our results are not biased by a choice of language or any other artificial means. For each tweet, we recorded the associated location information, both from the tweet itself and from the corresponding user's profile when applicable.

Frequencies for each of the location options were calculated to determine the prevalence of the various location data by region of the world, time zone, and state within the United States. Furthermore, data from the US Census Bureau were compiled to determine the proportion of the total United States population living in each state. Pearson correlation coefficients were used to compare states' populations with the prevalence of Twitter users who enable the GPS location option.

1.3 Results

Table 1.1 shows the total number of tweets and users, and their distribution over 3 types of location information: exact GPS coordinates (GPS-exact), GPS coordinates of a place (GPS-place), and time zone. In addition, the table shows the percentage of those who had either type of GPS coordinates, which is less than their sum because many users who supplied one also supplied the other. This aggregate value gives a more accurate picture of the amount of reliable (although less specific) location information directly available from tweets.

There was an average of 2.5 tweets per user. The extremely rapid rate of posting on Twitter (200 million posts per day amounts to more than 2000 tweets per second) and the streaming API's sampling mechanism (every 100th tweet) mean that it is unlikely that any user is overrepresented or underrepresented. Indeed, the probability that a user could post

Table 1.1: Tweets and users providing location indicators

Location indicator	Tweets		Users	
	n	%	n	%
Total (with and without location)	23,830,273	100	9,496,448	100
GPS-exact	216,900	0.91	140,451	1.48
GPS-place	458,295	1.92	241,010	2.54
GPS-exact or GPS-place	481,179	2.02	256,059	2.70
Time zone	18,347,947	76.99	6,831,414	71.94

in exact sync with the streaming API's sampling is virtually zero. The larger proportions of users who have enabled GPS as opposed to tweets containing GPS information may be explained by the fact that user accounts that run automated applications (ie, bots) are less likely to be GPS-enabled, but may post more frequently and account for more tweets in the sample than regular users. We have not attempted to identify such users here. The remainder of our results are based on unique users identified by their tweets during the 2-week time period.

1.3.1 Worldwide Distribution

To see whether the number of users and their location information varied across the world, we used the time zone information to overlay these values on a map of the world. The result is displayed in Figure 1.1[1], which shows the number of unique users in each time zone who enabled GPS, including the percentage of GPS-exact and GPS-place data. Although the time zones of North and South America have a high number of tweets, European time zones have a higher proportion of tweets that provide GPS information.

1.3.2 Profile Description Location Information

To parse the free text of the user-supplied information, we used a simple method of looking for text followed by a comma and a state name or abbreviation (ie, "text, state name" or "text, state abbreviation"). This simple parsing method could be improved, yet it provides a useful conservative estimate in its simplicity and efficiency. This method is inherently biased

[1]Time zones are aligned with longitudes not accounting for deviations based on country borders.

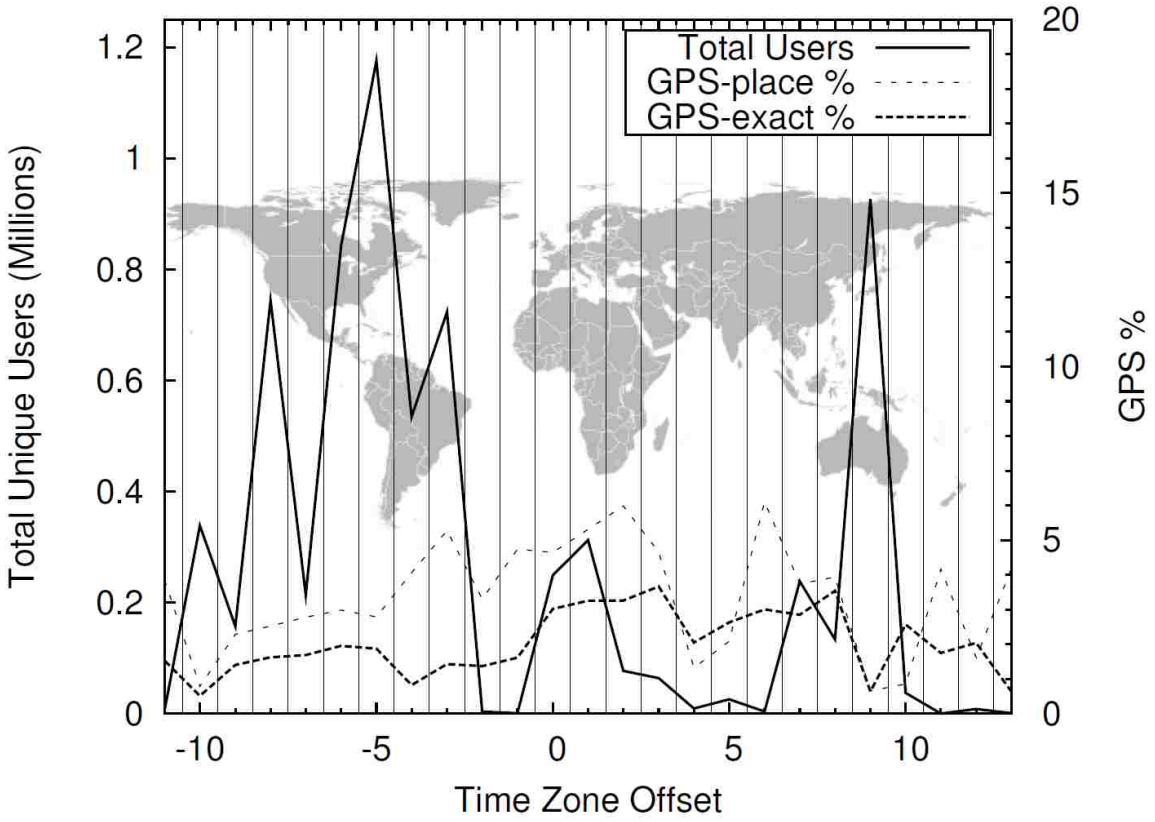

Figure 1.1: Distribution of Twitter users by time zone.

Table 1.2: Location of Twitter users within the time zones of the United States

Location indicator	Labeled "US & Canada"		Time zones GMT 5:00 to 8:00	
	n	%	n	%
Total (with and without location)	2,117,064	100	2,904,103	100
GPS-exact	41,416	1.96	53,997	1.86
GPS-place	60,979	2.88	82,322	2.83
Parsed state	315,819	14.92	379,576	13.07
Any (GPS-exact, GPS-place, or parsed state)	362,663	17.13	445,800	15.35

toward English-speaking locations and locations within the United States; therefore, results are shown only for users with time zones listed as one of the US time zones. As a matter of interest, the top 10 pairs parsed (with number of users) are Atlanta, Georgia (10,935); Los Angeles, California (10,244); Chicago, Illinois (8980); Houston, Texas (8147); New York, New York (7804); Washington, District of Columbia (6751); Miami, Florida (5734); Dallas, Texas (5688); Boston, Massachusetts (5562); and Austin, Texas (4678).

Table 1.2 and Figure 1.2 show the number of users who matched our parsing criteria for the 4 continental US time zones, specifically Greenwich Mean Time (GMT) 5:00 to 8:00. When restricting to the US time zones, there is ambiguity about whether to include those that are specifically labeled as a US time zone, such as "Pacific Time (US & Canada)", or simply those that contain a time zone offset that falls within the range of continental US time zones. For example, the time zone "Mexico City" is not labeled as a US time zone, yet its offset of GMT 6:00 is the same as Central Standard Time in the United States. Because the time zone may be automatically inferred by the user's local time when creating an account, many users in the United States may have their time zone set to a different zone with the same offset. Thus, focusing on those specifically labeled as "US and Canada" is likely to miss some users, but focusing on those within the offset range is likely to include many Central and South American users. We have included results for both cases.

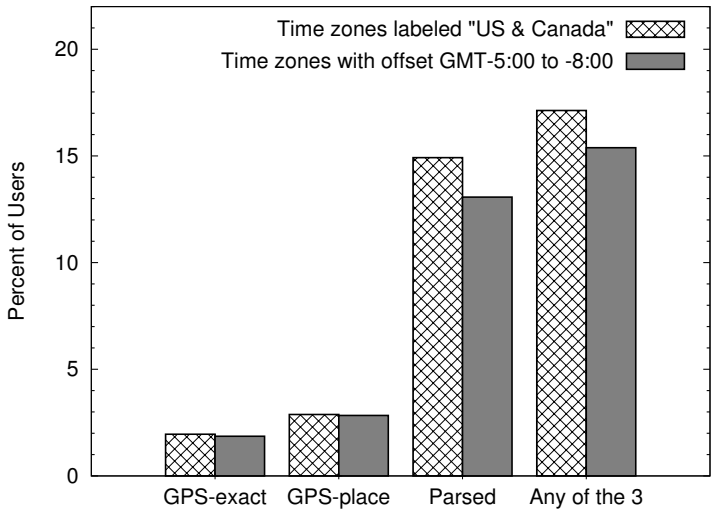

Figure 1.2: Twitter users providing location indicators in the US time zones.

1.3.3 Accuracy of User-Supplied Data

Users enter their location information in their user profiles themselves; thus, there is potential for inaccuracy. To evaluate the accuracy of the user-supplied profile location, we compared parsed state data and GPS coordinate data when both were available. City data may be too difficult to parse because individuals may live in one city and work or go to school in another. Therefore, a comparison of state data is more appropriate provided the same individuals are less likely to cross state boundaries repeatedly on a daily basis.

When GPS-exact data were available, we used the Yahoo! Place Finder API [252] to determine the state's identity through a reverse GPS lookup service. When GPS-place data were available, we extracted the state name based on the Twitter Place Type (directly, when supplied, or using a reverse GPS lookup as described previously). We compared the state name obtained by these methods with the state name parsed from the user-supplied location information. Table 1.3 shows the results.

Table 1.3: Comparison of GPS location data to parsed location data

GPS location indicator	State name parsed from user profile (n)	Matching parsed and GPS data	
		n	%
GPS-exact	16,009	13,935	87.04
GPS-place	21,092	18,599	88.18
Total	37,101	32,534	87.69

1.3.4 Distribution in the United States

With the parsing method in place, we extended our analysis of location information in Twitter to include parsed state data for the United States. Parsing international location data is a complex task, requiring such tools as standardization, place authority, and handling diverse conventions and languages. Figure 1.3 shows the proportions of users with parsed state data, with GPS-exact data and with GPS-place data in each state, and the proportions of 2010 US census population in each state. All of the location indicators correlate strongly ($P < .001$) with the population data (GPS-exact $r = 0.97$, GPS-place $r = 0.97$, and parsed $r = 0.98$).

Figure 1.4 complements Figure 1.3 by showing the number of Twitter users in each state per capita (ie, divided by the census population) and the median value (0.0015) for the states identified through parsing. This does not represent the total number of registered Twitter users, but rather the number of unique users who posted during our sample period. The relatively high number of Twitter users in the District of Columbia, compared to its population is likely because users identify with and tweet from the metropolitan area, but actually reside in outlying suburbs in different states.

1.4 Discussion

The purpose of this study was to document the prevalence of the location identification options available through Twitter and to present an estimate of the usability of each option. We have shown that there are several location indicators in Twitter and, when taken together,

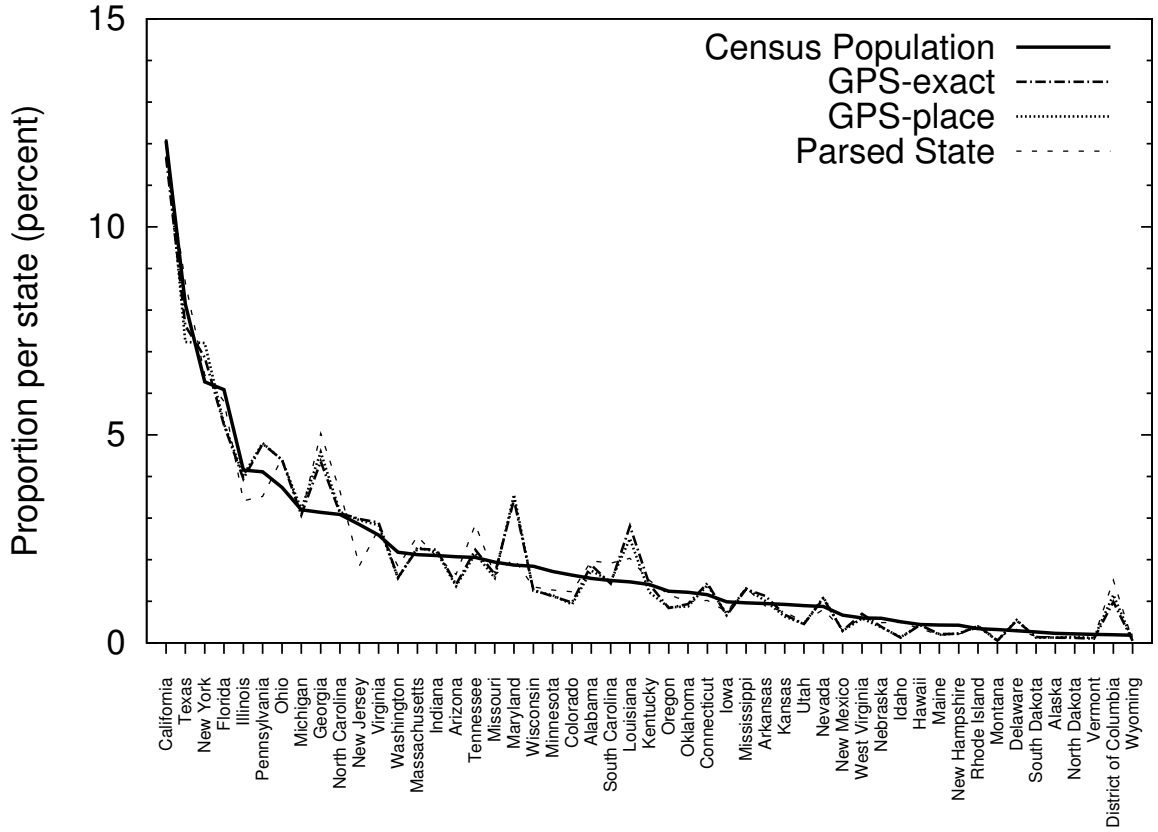

Figure 1.3: The proportion of Twitter users identified in each state and the proportion of the 2010 US census population in each state, ordered by census population.

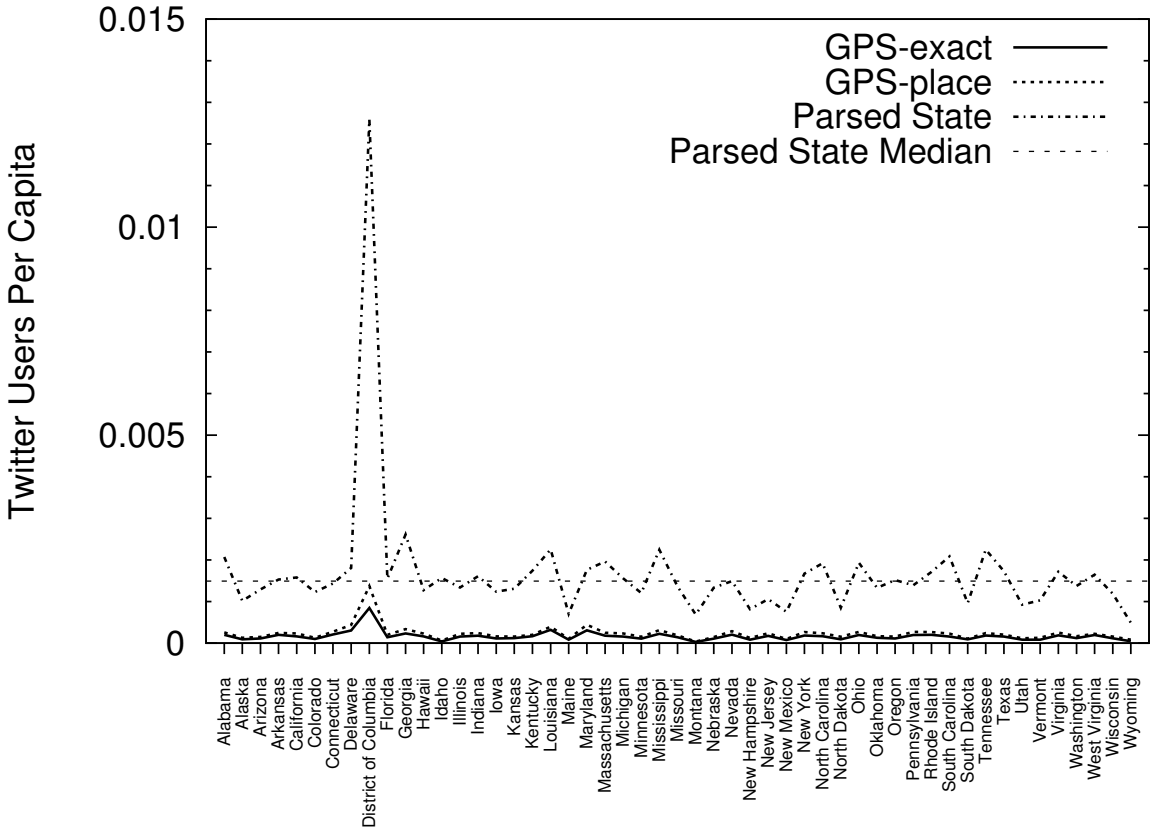

Figure 1.4: The number of geolocated Twitter users per capita in each state.

they offer a sizable sample of individuals whose location can be accurately inferred. This has clear implications for infoveillance, infodemiology, and "right time, right place" health communication.

Although only a small percentage of Twitter users provide reliable GPS coordinates (2.70%), there is actually a large number of users (and tweets) with GPS data because of the size of the overall data set. In the 2-week period of this study, the 2.02% of tweets that contained GPS information corresponded to 481,179 tweets. Because the sample is only 1% of the overall traffic on Twitter during those 2 weeks, if the same proportion were to hold true in the larger sample, we could infer that there were about 48 million tweets with GPS information posted during that period. With 2.5 tweets per user, this would correspond to approximately 19 million individuals. Furthermore, we saw that user-supplied location information matched GPS data in 87.69% of cases (in the United States). Hence, one could reliably use location information for between 15.35% and 17.13% of users. Interestingly, Keeter and colleagues compared the results of a 5-day survey employing the Pew Research Center's methodology (with a 25% response rate) to those from a more rigorous survey conducted over a much longer field period and achieved a higher response rate of 50% [119]. In 77 of 84 comparisons, the 2 surveys yielded results that were statistically indistinguishable. Thus, it appears that surveys with lower response rates (20%) were only minimally less accurate. As a result, researchers can have additional confidence to value the location information available from Twitter, a real-time and real-place benefit of social media over traditional survey methodology.

Table 1.1 also shows that 2.70% of Twitter users broadcast their GPS location. Interestingly, this is a significantly lower figure than the 14% of social media users who use automatic location tagging on posts reported in a recent publication by the Pew Internet and American Life Project [261]. An obvious difference is that the Pew research considers all social media, whereas we have focused exclusively on Twitter. Additionally, some of the respondents represented in the Pew report could be using third-party location-tagging

applications (eg, foursquare), which data may not appear in our sample, or they may be tweeting so infrequently that they would be underrepresented in our tweet-based sample. However, even considering these possibilities, the magnitude of the difference suggests that there may be additional factors. This difference warrants future research to determine the extent to which users are even aware that broadcasting GPS location is possible. It is plausible that users are largely unaware of such features or have minimal understanding with respect to how they function, both of which may attribute to this discrepancy.

An additional explanation for this (rather significant) discrepancy between 2.70% and 14% may be the distinct data collection approaches employed: questionnaires administered via phone interviews versus direct observation of user behavior on Twitter. Questionnaires can only report on what people perceive as opposed to what may actually be happening. For example, the question asked in the Pew questionnaire leading to the above result was "Thinking about the ways people might use social networking sites...Do you ever...Set up your account so that it automatically includes your location on your posts?" The answers included "Yes, do this: 14%" and "No, do not do this/have not done this: 84%."

It is possible that respondents believed that GPS location was a default setting. This would lead to the conclusion that they had enabled location tagging for social media on their device, although GPS coordinates were not broadcast. From our own experience with the iPad 1, we found that the device itself may be GPS-enabled, yet the Twitter application on the device is not. Furthermore, the application could be GPS-enabled, yet coordinates are not broadcast because the location setting is not activated in the Twitter profile. In that sense, it is possible that someone may think that their tweets are location-tagged when, in fact, they are not. In this way, public health may benefit from eliciting additional location information that can be provided in the actual tweet. Twitter users who are otherwise willing to reveal their location, but are unaware of the default privacy settings, could be encouraged to provide such information. For example, followers of Twitcident (http://twitcident.com), a Dutch-based system for filtering emergency-related tweets, may feel inclined to tweet the

location of emergency situations in an effort to assist emergency responders. Twitter prompts that ask users to tweet about their favorite locations to exercise may be useful in helping authorities allocate resources for promoting active lifestyles in areas where they are most likely to be successful.

As observed in this study, the parsed state data matched the GPS-derived state data 87.69% of the time. A mismatch does not necessarily mean that the user-supplied location was inaccurate or purposefully misleading, but it could represent a user tweeting from a business trip or vacation, or working in a metropolitan area across state lines. In this regard, the percentages in Table 1.3 are a lower bound and validate that the majority of the user-supplied locations are accurate for those users who provide GPS data and have profiles that can be parsed with our method. However, there is a potential bias in that users who are willing to broadcast their location might be more likely to tell the truth in their profile. Also, users who are unwilling to give an accurate profile location may be more likely to leave it empty or provide a non-descriptive location, as opposed to supplying an inaccurate, yet well-formed, location. This could be the focus of future research aimed at determining the extent to which Twitter users enable/disable GPS broadcasting and their reasons for doing so. For example, Twitter users vacationing in an exotic location may wish to enable GPS broadcasting, whereas others may disable broadcasting if their desire is to remain anonymous. This assumes that these users are aware of the toggle settings available for GPS broadcasting. Studies of this nature could establish the basis for determining the representativeness of GPS-enabled tweets. Moreover, this finding may question Twitter's utility as a means for providing "right time, right place" tailored interventions, considering the location may not reflect the user's actual setting, provided he or she knowingly deactivates location.

As presented in this study, there is a significant level of consistency between the proportion of location-tagged tweets and state populations in the United States. This finding indicates that, at least within the United States, there is no evidence of disproportionate GPS enabling among states. Although much more information is needed to assess the true

qualitative representativeness of Twitter (eg, ethnicity, age, and gender), this quantitative consistency is promising. Whereas it was beyond the purview of the current study to assess the validity of social media data, for public health researchers and communicators to dismiss such data sources without further consideration would be premature because it may miss an opportunity to observe, reach, and communicate with people in unprecedented ways. And although it is unlikely that social media could ever completely replace more traditional research methods (eg, questionnaires), it can certainly complement them and add a further dimension to research.

In conclusion, we note that we have focused our attention on what users can do explicitly to specify their own location information. Although Twitter's opt-in policy for location information is ethically sound, it would be interesting to study what could be done to encourage increased opt-in, for example, by working on dispelling concerns about how information is used or by demonstrating how information can be used for the good of all (eg, the Twitcident app). Furthermore, recent studies have demonstrated that it may be possible to infer location information based on either the words appearing in a user's tweets [45] or the location of a user's friends [18]. Further exploration of these ideas and other means of geographical prediction could augment the amount of location information available in social media.

Acknowledgments

We are grateful to Twitter for the availability of their data and associated API.

Chapter 2

Tracking Suicide Risk Factors through Twitter in the U.S.

2.1 Introduction

Suicide is a leading public health concern in the United States. As the tenth leading cause of death in 2009, the most recently compiled year for national suicide statistics, suicide resulted in 36,909 deaths [128]. When accounting for the age at death, suicide becomes the fifth leading cause of years of potential life lost in America [40]. Non-fatal forms of self-inflicted violence further burden the nation with mounting emergency department visits; a total of 472,000 visits were seen in 2007 alone [184]. Furthermore, surviving family and friends who have to endure the outcomes of fatal or nonfatal self-directed violent behavior also shoulder the burden of suicide.

While suicide poses a major community health risk, research in this area remains difficult. Several barriers in regards to suicide studies include the lack of organized surveillance (specifically concerning suicide attempts), relatively low base-rate of suicide, issues concerning ethics and safety, and the difficulty of ascertaining information after the death of an individual (who may not have shared pertinent information with those around him or her). Each of these barriers complicates the gathering of suicide data, thereby slowing the pace of understanding suicide through research [86].

To aid in suicide prevention, public health and mental health officials need data that is collected in real time in order to intervene before people actually take their own lives. Crosby, Han, Ortega, Parks, and Gfroerer stated, "Public health surveillance with timely and consistent exchange of data between data collectors and prevention program

implementers allows prevention program practitioners to implement effective prevention and control activities" [55, p. 1].

Social media is an emerging tool that may assist research in this area, as there exists the possibility of passively surveying and then subsequently influencing large groups of people in real time. Recent studies have shown "that social media feeds can be effective indicators of real-world performance" [15, p.1], box office predictions [15] and stock market forecasts [260]. Twitter has also been used to "estimate disease activity in real time, i.e., 12 weeks faster than current practice allows" in a study tracking the spread of Influenza A H1N1 in 2009 [216, p. 7]. Furthermore, Ruder, Hatch, Ampanozi, Thali, and Fischer have shown that some Facebook users do, in fact, post suicide notes on their profiles, exposing the potential for suicide research in social media [210].

The amount of publicly available information spread across the realm of social media is extensive. Twitter is of interest due to its greater public availability of data, larger user base, and it being a platform of personal expression. Twitter is a social media platform wherein users ("tweeters") post status updates, or "tweets," that are distributed to others that "follow" them, and are also made available to the public. Emerging from its beginning in 2006 [14], "Twitter is now playing a major role in our society, with over 200 million users already and estimates of 500,000 new accounts being added each day" [90]. Together these users generate 400 million tweets per day [24]. This large reservoir of information regarding people's daily lives and behaviors, if handled correctly, can be used to study suicide and possibly intervene.

The recent live Twitter feed of a pending suicide demonstrates that at-risk tweets about suicide can lead to suicidal behavior in this case fatal [150]. While suicidal risk factors may or may not be a direct cause, they are important characteristics associated with suicide and can be observed through conversation. Research regarding these risk factors is well established and provides a framework for further research and intervention [156, 237].

The purpose of this study was to determine if at-risk suicide Twitter conversations are related to actual suicide rates. If so, Twitter could serve as an important portal for future research and a potential platform for public health interventions to prevent suicide.

2.2 Method

The following subsections define the methodology.

2.2.1 Twitter Data

Twitter provides an application-programming interface (API) that enables programmatic consumption of the data. The Twitter Streaming API provides means of obtaining tweets as they occur, filtered by specific criteria, such as a list of keywords. While some tweets/accounts are marked private, most are openly available to the public and authored without expectation of privacy, making them an accessible data source for researchers. We received an exemption from the university's internal review board to monitor these publicly available tweets.

To identify potential suicide-related tweets, a list of search terms was created based on various risk factors and warning signs linked to suicide. These risk factors and warning signs included depression and other psychological disorders [142], prior suicide attempts [142], family violence, family history of drug abuse, firearms in the home, and exposure to the suicidal behavior of others [171]. Other search terms included common antidepressants, as well as phrases that indicated suicide [94], ideation [9], deliberate self-harm [258], bullying [126], feelings of isolation [40], and impulsiveness [10].

The researchers employed a 2-part process to identify keywords, or search terms that represented each risk factor. First, the researchers jointly generated multiple search terms for each risk factor by simply identifying phrases or keywords that appeared to be related to the risk factor. Second, the researchers pilot tested each search term. Those terms that appeared in tweets, accompanied by the expected suicide risk context, were retained. Search terms

Table 2.1: Twitter Search Terms and Statements for Suicide Risk Factors

Suicide Risk Factor	Search Terms and Statements
Depressive Feelings	me abused depressed, me hurt depressed, feel hopeless depressed, feel alone depressed, I feel helpless, I feel worthless, I feel sad, I feel empty, I feel anxious
Depression Symptoms	sleeping 'a lot' lately, I feel irritable, I feel restless
Drug Abuse	depressed alcohol, sertraline, Zoloft, Prozac, pills depressed
Prior Suicide Attempts	suicide once more, me abused suicide, pain suicide, I've tried suicide before
Suicide Around Individual	mom suicide tried, sister suicide tried, brother suicide tried, friend suicide, suicide attempted sister
Suicide Ideation	suicide thought about before, thought suicide before, had thoughts suicide, had thoughts killing myself, used thoughts suicide, once thought suicide, past thoughts suicide, multiple thought suicide
Self-harm	stop cutting myself
Bullying	I'm being bullied, I've been cyber bullied, feel bullied I'm, stop bullying me, keeps bullying me, always getting bullied
Gun Ownership	gun suicide, shooting range went, gun range my
Psychological Disorders	I was diagnosed schizophrenia, been diagnosed anorexia, diagnosed bulimia, I diagnosed OCD, I diagnosed bipolar, I diagnosed PTSD, diagnosed borderline personality disorder, diagnosed panic disorder, diagnosed social anxiety disorder
Family Violence/Discord	dad fight again, parents fight again
Impulsivity	I impulsive, I'm impulsive

that did not appear in the initial search were deleted from the list. These terms are listed in Table 2.1.

Using the Twitter Streaming API filtering by terms listed in Table 2.1, tweets were collected and stored in a database categorized as potential "at-risk" tweets, or tweets that seemed indicative of a potential risk factor of the tweeter. To focus on those tweets that were most relevant to the purpose of the study and also those tweets that were geolocated, this set was further refined in two ways. First, only those tweets where the user's state name could be easily identified were used. These states were identified by either the user-provided direct GPS information, or by parsing the user's profile "location" field for either a state name or abbreviation, or text followed by a comma and a state name or abbreviation.

The second way the at-risk dataset was filtered was through a process aimed at removing tweets that were either jokes, non-pertinent, or sarcastic in nature. A manual inspection of sample tweets collected resulted in identified words or phrases that could be used to filter out irrelevant tweets. For example, the tweets obtained through the Twitter

Streaming API included those with the words "stop", "cutting", and "myself" (key words indicative of self-harm), which would seem to be related to a risk factor, but not if they also contained words such as "shaving," "accidentally," and "slack." Thus, by using a list of exclusion terms in combination with each inclusion term phrase, the number of sarcastic tweets was reduced. The list of terms used as exclusion criteria can be found in Table 2.2[1]. It was not feasible to manually inspect all of the at-risk tweets in this sample to determine the extent to which these exclusion terms refined the study sample. However, a review of the content of a sample of study tweets revealed that this process worked as expected.

Using the user's state information, the Twitter users that posted these at-risk tweets were grouped by state for further analysis. Rather than rely on raw numbers of tweets, which vary greatly over time, we focused on proportions. A baseline was first established using the results of Burton et al. [35]. In that study, the default random sample of 1% of all tweets provided by the Twitter API was observed during two separate weeks in October and November of 2011. Unique users were identified and classified according to state using the same process as described above. The proportions of tweeters per state with respect to the total number of tweeters were then computed. These baseline values, one for each state s, are referred to here as $\alpha_b(s)$. Similarly, the proportions of at-risk tweets per state with respect to the total number of at-risk tweets were also computed. The resulting values, one for each state s, are referred to here as $\alpha_r(s)$. In the absence of other information, the simplest hypothesis, in a Bayesian sense, is to assume that the distributions of these quantities over states are the same, i.e., for all states s, $\alpha_r(s) = \alpha_b(s)$. It is therefore possible to design a natural, unit-free measure of departure from this expectation, namely the ratio $d_\alpha(s) = \alpha_r(s)/\alpha_b(s)$. A value of d_α greater than 1 for a given state suggested that there were proportionally more at-risk tweeters in that state than expected, whereas a value of d_α smaller than 1 suggested the opposite. We do realize that the collection of at-risk tweeters lagged behind the collection of all tweets by approximately 6 months. While the raw numbers of accounts and tweets would

[1]Any search terms and statements not found in this table did not undergo a filtering process because they were found to produce sufficiently positive results.

Table 2.2: Exclusion Filter Terms used for Search Terms and Statements

Search Terms and Statements	Exclusion Filter Terms
feel alone depressed	cockroach, 364
I feel helpless	when, without, girl
I feel sad	episode, when, lakers, about, game, you, sorry, for, bad, bieber
I feel empty	stomach, phone, hungry, food
sleeping 'a lot' lately	'haven't been'
I feel irritable	was
depressed alcohol	ronan
sertraline	"special class", viagra, study, clinical, http
Zoloft	toma, para, necesito, siempre, gracioso, desde, decirle, palabra, vida, sabor, aborto, gusta
Prozac	toma, para, necesito, siempre, gracioso, desde, decirle, palabra, vida, sabor, aborto, gusta
pills depressed	http
suicide once more	will, by, live
pain suicide	http
mom suicide tried	dog, cat, fish, who
sister suicide tried	dog, cat, fish
brother suicide tried	dog, cat, fish, big brother
friend suicide	hold still
suicide attempted sister	paperback
thought suicide before	http
had thoughts suicide	http, never
had thoughts killing myself	not
stop cutting myself	off, shaving, hair, shave, slack, accidentally
I'm being bullied	straightophobic
feel bullied I'm	lol
stop bullying me	#stop
always getting bullied	lol
gun suicide	zimmerman, news, you, water, nerf
been diagnosed anorexia	http
I diagnosed OCD	never, CDO, check
I diagnosed bipolar	n't
dad fight again	food
parents fight again	sartan, bradley, pacquiao, gas
I impulsive	clementine
I'm impulsive	clementine

have certainly changed over that period (see above about the estimated 500,000 accounts being added each day), there is no reason to expect the distribution of tweeters across states to have varied significantly, thus further validating our use of d_α.

2.2.2 Vital Statistics Data

Geographic, state-by-state, suicide rates from 2009 were based on age-adjusted data. These data were taken from the National Vital Statistics System as reported in the Center for Disease Control and Prevention report "Death: Final Statistics for 2009." This report provides the total number of deaths, the death rate, and age-adjusted death rate for intentional self-harm (suicide) for all 50 states and the District of Columbia. Data are gathered from death certificates as completed by funeral directors, physicians, medical examiners, and coroners [128]. As with the Twitter data, we also transformed the death data into departure from expectation values $d_\beta(s) = \beta_r(s)/\beta_b(s)$, where $\beta_r(s)$ is the ratio of the proportion of deaths by suicide per state with respect to the total number of deaths, and $\beta_b(s)$ is the proportion of the US population per state with respect to the total US population. Again, as with tweeters, variations in population distribution across states are slow so that d_β is valid.

2.2.3 Analysis

Using Microsoft Excel we calculated a Spearman's rank correlation coefficient between the d_α's (observations on Twitter) and the d_β's (observations in the real world), and a corresponding p-value to verify statistical significance.Geographic maps of the d_α's and actual suicide rates were created using ESRI ArcMap 10 geographic information system software.

2.3 Results

Using the Twitter Streaming API filtering by the inclusion terms listed in Table 2.1, tweets were collected from May 15, 2012 to August 13, 2012, totaling 1,659,274 tweets from 1,208,809 unique users throughout the world. Applying the exclusion terms in Table 2.2 resulted in

Table 2.3: Example Tweets for Suicide Risk Factors

Suicide Risk Factor	Example Twitter Posts
Depressive Feelings	I feel so worthless today.
Depression Symptoms	I've been sleeping a lot lately. I take like 6 hour naps.
Drug Abuse	Dear Prozac, time for a upping in your dosage!
Prior Suicide Attempts	I tried to commit suicide before .. Several times.
Suicide Around Individual	I have a friend that comitted suicide :'(.While hate may run deep love runs even deeper.
Suicide Ideation	I have had thoughts on suicide and running away from home....and sometimes I still do.
Self-harm	people say "stop cutting! be happy with who you are." its so much easier to say than do... i hate myself so much..
Bullying	I'm sick of being bullied. Everyone care about there problems and don't even bother to check on me. I'm going to kill myself!! ?
Gun Ownership	I need to get into da gun range I haven't fired my old gun in over 2 years now
Psychological Disorders	@antashafarhanah Idk what to say but yes, I've been diagnosed with anorexia since late 2009 and early 2010.
Family Violence/Discord	BIGGEST fight with dad EVER. Ended in a fist fight. I've packed my bags & I'm leaving. I hold a grudge so dunno how long b4 we talk again.
Impulsivity	I'm so impulsive. I don't think before I do things. That's why I make mistakes.

a set of 733,011 tweets from 594,776 users. Sample tweets for each risk factor are listed in Table 2.3 (original spelling and grammar preserved). Of these tweets, a specific state in the United States could be identified for 37,717 tweets from 28,088 unique users. This set of location-identified users is used for analysis and referred to as the at-risk tweeter set. Tweets indicative of suicide risk factors were varied in their seriousness and clarity. To verify the relevance of the set, we had two raters independently classify the same random sample of 1,000 tweets. They were in agreement 79.6% percent of the time. A Cohen's Kappa coefficient was calculated to measure the level of agreement between the two coders ($k = 0.48$), which is classified as moderate agreement [136]. Athird rater was then used to arbitrate those tweets that were in disagreement.Of the 1,000 tweets, 789 (78.9%) were found to be relevant, in that the keyword terms were being used to indicate the risk factor, as opposed to being out of context or in a completely sarcastic manner.

Table 2.4: Top 10 At-Risk States According to d_α

Rank	State	# At-Risk Suicide Twitter Users	At-Risk Suicide d_α
1.	Alaska	61	1.800
2.	New Mexico	136	1.683
3.	Idaho	72	1.617
4.	South Dakota	57	1.607
5.	Montana	27	1.557
6.	Utah	195	1.551
7.	Texas	3022	1.491
8.	Kansas	241	1.365
9.	Arizona	509	1.334
10.	Oklahoma	314	1.285

Table 2.5: Bottom 10 At-Risk States According to d_α

Rank	State	# At-Risk Suicide Twitter Users	At-Risk Suicide d_α
42.	Vermont	26	0.814
43.	New York	1548	0.771
44.	Hawaii	90	0.749
45.	Connecticut	280	0.729
46.	New Jersey	595	0.728
47.	District of Columbia	215	0.706
48.	Delaware	104	0.673
49.	Pennsylvania	902	0.661
50.	Maryland	606	0.606
51.	Louisiana	435	0.590

Table 2.4 lists the top 10 states with the highest values of the d_α[2]. States with the highest values of d_α tended to be in the midwestern and western states such as Alaska (1.800), New Mexico (1.683), Idaho (1.617), South Dakota (1.607), and Montana (1.557). Table 2.5 lists the bottom 10 states with the lowest values of the d_α. States with the lowest values of d_α tended to be in the south and eastern states such as Louisiana (0.590), Maryland (0.606), Pennsylvania (0.661), Delaware (0.673), and the District of Columbia (0.706).

Results revealed a Spearman's rank correlation coefficient of $r = 0.53$ ($p < 0.001$) when comparing the Twitter-generated d_α values with the age-adjusted d_β values computed from the National Vital Statistics System, (Spearman's $r = 0.53$, $p < 0.001$).Figure 2.1 illustrates the d_α values for all states while Figure 2.2 illustrates the U.S. age-adjusted suicide rates.

[2]The number of Twitter users indicates the number of suicide risk factor Twitter users for a 3 month period.

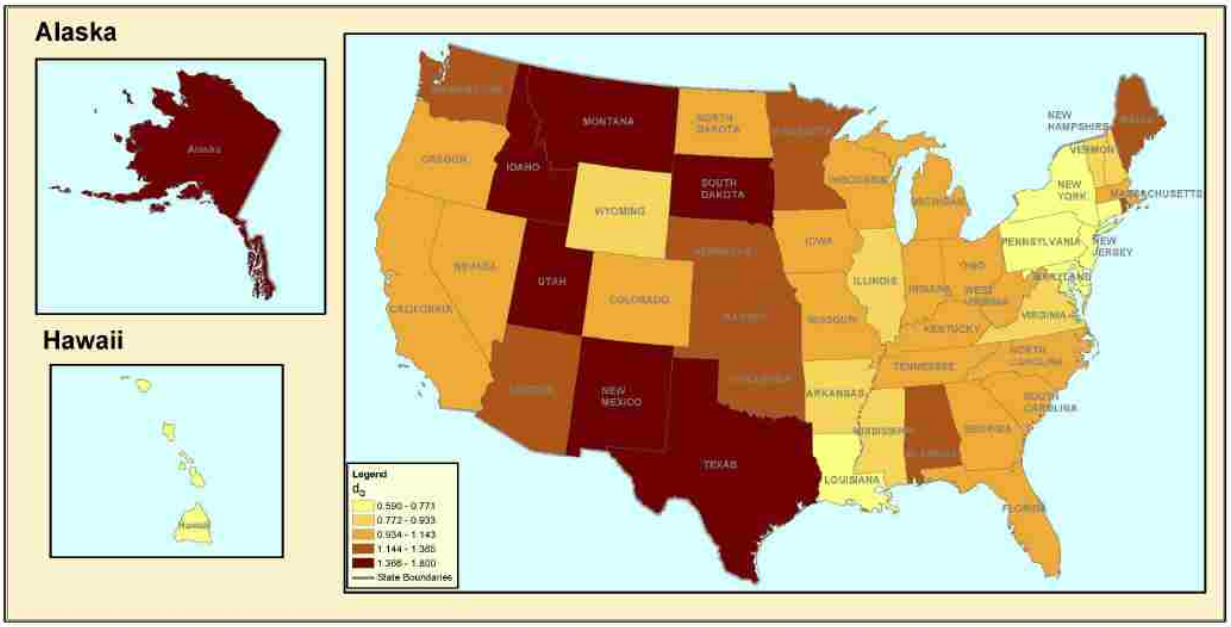

Figure 2.1: Risk factor tweet d_α values in the U.S.

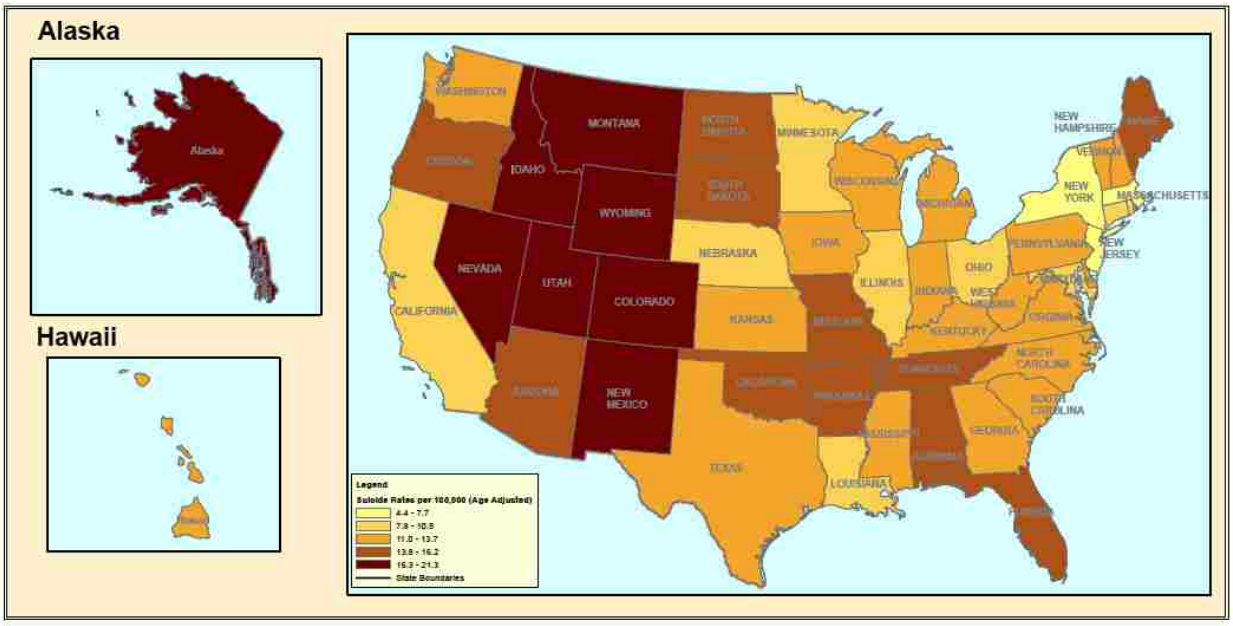

Figure 2.2: Age-adjusted suicide rates in the U.S.

2.4 Discussion

The purpose of this study was to demonstrate that Twitter conversations indicative of suicide risk factors are related to geographic-specific suicide rates from traditional data sources. These findings provide initial validation for Twitter as a potential dataset for future suicide research and a platform for public health and social service interventions. Findings indicate that there is an association between rates of tweets by users determined to be at-risk for suicide and actual suicide rates. States in the midwest-western US region, and Alaska, were observed to have a higher proportion than expected of suicide-related tweeters (i.e., $d_\alpha > 1$). These states also have the highest actual rates of suicide. To our knowledge this is the first study of its kind to attempt to compare tweets containing suicide-related content to actual rates of suicide. Whereas these findings do not extend our understanding of human behavior per se, this sort of validity testing of an emerging data source provides preliminary confirmation of its potential value in monitoring and understanding suicide related risks. Had the findings from this study been inconsistent with the study hypotheses, concerns would have been raised about the utility of microblogging and social media content as a surveillance tool for social scientists.

Suicide assessment and subsequent intervention are among the most important roles of mental health professionals [151]. With rising healthcare costs and the expense associated with collecting and analyzing data for entire populations, this task of assessing patients at-risk for suicide is challenging. Lamberg calls attention to the gravity of the current situation by recalling that: "It used to be that a patient talking about suicide was always hospitalized. Today the patient has to come in with a gun to his head or to your head to get hospitalized. We have to deal with suicidal patients more in the community" [130, p. 687]. Indeed, projects like the current one that employ innovative methods may play an increasingly important role in strengthening the link between primary and secondary prevention efforts, both of which have been identified as necessary components of a comprehensive prevention effort [124]. Such primary prevention efforts are underway among adolescents and have involved instructing

teachers to be aware of verbal manifestations of suicide risk factors. King identifies phrases such as, "my family would be better without me," and "I can't stand living anymore," as characteristic vernacular for patients at-risk for suicide [124]. For obvious reasons this is more challenging with adults. Hence, results from the current study are promising as they suggest a potential mechanism for identifying adults at-risk for suicide to the extent that they tweet, and make publicly available suicide-related statements. While this study is not an attempt for intervention, it may be an important surveillance tool to detect suicidal patterns and create a potential mechanism for a directed tweet response. Further, as identified in Table 2.3, a selection of qualitative statements identifies a representative listing of tweeted messages about suicide, which corroborates the phrases similarly noted above by King [124]. A systematic analysis of these qualitative data may be helpful for future research as to the severity of risk as well as the social responses that emerge from social media.

Twitter users tweet about a variety of subjects, the content of which may not be truly reflective of their feelings about a given topic. Indeed, much of Twitter content has been labeled meaningless discussion [120]. Despite that claim, the fact that users tweet suicide-related content likely suggests that they are at least thinking about the topic and are comfortable sharing this information to such a broad audience. Perhaps users open up in online settings more than in face-to-face settings, where research has shown individuals require a commitment to confidentiality in order to share sensitive information. Such confidentiality is antithetical to the concept of Twitter where tweets are publicly available. West et al. [244] showed that Twitter users readily share information about their problem drinking. Another recent study of social media showed that women on blogs readily discuss challenges to breastfeeding, which is a topic of potential embarrassment [243]. Humphreys, Gill and Krishnamurthy, found in their content analysis of Twitter messages that the majority of users do tweet about themselves [101]; however, they overwhelmingly take care to protect privacy by not providing personal information such as phone numbers, email, or home addresses. Whereas it is unknown to what degree people tweet about their feelings related

specially to suicide, future research might focus on this question. In addition, while this study demonstrates the efficacy of Twitter for surveillance purposes, the feasibility of using this channel of communication to intervene among those at-risk will likely depend on whether privacy can be ensured. That said, the current study provides promising evidence of a new way of collecting data to help advance intervention possibilities.

Provided that additional research studies corroborate the findings from the current study, public health priorities in suicide prevention should consider creating profiles of individuals that might lead to earlier detection of suicide ideation. These profiles might include characteristics such as common discussion topics, frequency of tweets, gender, etc. Users that are flagged early as at-risk for suicide could be engaged in Twitter conversations with professionally trained practitioners that may be effective at convincing the tweeters to seek medical attention, or the user could simply be referred to web-based resources. As an example, Twitcident[3] is a Dutch-based system for filtering emergency related tweets and may be used as a mode for public health. Twitcident uses Twitter data to engage emergency services personnel by monitoring tweets that discuss local emergencies. These retrospective profiles built from potential suicidal users' tweets may allow a coordinated public health and mental health response to preventing suicide. In this way, the public health response can more squarely address secondary prevention opportunities in addition to its existing primary prevention priorities.

While there have not been many applications of real-time data collection and prevention strategies within the realm of suicide research, there have been new utilizations amongst depression researchers. A new smartphone app called Mobilyze [32] has been created that uses data collected by an individual's smart phone (such as location, social contexts, and recent activities) to assess the current level of depression within that person. After installing the app, the user answers a series of surveys that the app uses to determine whether or not the owner is depressed. When the smartphone detects activities or contexts that equate with

[3]http://twitcident.com

high attitudes of depression, it sends messages to select family or friends, alerting them of the individual's depression status.

A similar app could be created that measures a user's online activities, gathered in real time, to assess the level of suicide risk of the app user and alert family, friends, or a professional counselor of the elevated risk of the individual. If a patient gives consent to a counselor, that counselor can then monitor their patient's social media mood and collect important data (e.g., disrupted relationships, loss of a job, online suicide threats) within minutes and hours, rather than having to wait until their next appointment to gather such information. With this real-time data, counselors and family will be able to reach out to these at-risk individuals in the moment of need. Policy-makers and those who fund research projects should consider next steps for studying and supporting more social media based efforts for public health and social service interventions.

2.5 Limitations

Findings from this study should be interpreted in the context of several key limitations. First, the search filters allowed for a proportion of unrelated tweets to be coded as at-risk. Moreover, the search terms may have been insufficient to capture all instances of at-risk tweets. However, previous research was consulted to compile a list of keywords and search terms in an effort to reduce the number of false positives and false negatives. There is undoubtedly a balance that must be achieved; a sufficient number of search terms to identify risk, but not too many so as not to falsely determine risk. This balance is likely needed in face-to-face settings as well. Second, identifying tweet location in some states was challenging, which led to a smaller number of tweets. Smaller samples introduce inherent challenges related to generalizability. Notwithstanding this limitation, the trends were largely consistent with those from states with larger samples within the same general geographical region. Difficulties in ascertaining location information were not limited to this study and have been the focus of previous research [35]. Efforts to detect levels of suicidal intent could not be assessed. As a result, the

study findings cannot differentiate between persons who are contemplating versus those who are preparing to take immediate action. However, since the rates between actual suicides and Twitter discussion were so highly correlated, it can be presumed that the identified Twitter users are at least at-risk for suicide in some regards. Third, actual suicide rates in the current study reflect 2009 values, while tweets came from 2012. The extent to which this impacted the findings of this study is unclear, especially considering that there is very little variance from year to year in suicide rates. Nevertheless, more definitive conclusions about the association between twitter content and actual rates should be reserved for comparisons in future studies that feature data comparisons from common years. Lastly, findings from this study should be interpreted in the context of what is known about important social and cultural demographic characteristics of Twitter users. The Twitter community consists largely of young adults. In fact, 26% of Internet users aged 18-29 use Twitter compared to 14% of those aged 30-49 and 9% of those aged 50-64 [217]. In addition, more black Internet users use Twitter (28%) compared to Hispanics (14%) and whites (12%). Due to the nature of the social media, Twitter provides users with a platform to engage with other users online. Engaging and associating with others is a characteristic not expected of one at-risk for suicide that may be experiencing depression and its associated symptoms of social isolation and withdrawal. The degree to which social isolation and withdrawal occur within social media communities such as Twitter is less understood and warrants further research.

2.6 Conclusions

An association exists between the proportion of Twitter users determined to be at-risk for suicide and actual suicide rates. States in the midwest-western U.S. region, and Alaska, were observed to have the highest d_α values (i.e., proportions of at-risk tweeters much larger than expected). These states also have the highest actual rates of suicide. Twitter may be an effective and valuable tool for gathering data in real time and on a large scale, which has not been conducted for suicide before. Suicide data gathered from Twitter is comparable to

data gathered through other means and is less costly. Using social media, researchers and practitioners may be one more step toward affordably and rapidly detecting individuals with suicidal intentions and may subsequently provide a platform to improve suicide prevention strategies through timely intervention.

Chapter 3

Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) Among College Students

Abstract

- **Background**: Adderall is the most commonly abused prescription stimulant among college students. Social media provides a real-time avenue for monitoring public health, specifically for this population.
- **Objective**: This study explores discussion of Adderall on Twitter to identify variations in volume around college exam periods, differences across sets of colleges and universities, and commonly mentioned side effects and co-ingested substances.
- **Methods**: Public-facing Twitter status messages containing the term "Adderall" were monitored from November 2011 to May 2012. Tweets were examined for mention of side effects and other commonly abused substances. Tweets from likely students containing GPS data were identified with clusters of nearby colleges and universities for regional comparison.
- **Results**: 213,633 tweets from 132,099 unique user accounts mentioned "Adderall." The number of Adderall tweets peaked during traditional college and university final exam periods. Rates of Adderall tweeters were highest among college and university clusters in the northeast and south regions of the United States. 27,473 (12.9%) mentioned an alternative motive (eg, study aid) in the same tweet. The most common substances

mentioned with Adderall were alcohol (4.8%) and stimulants (4.7%), and the most common side effects were sleep deprivation (5.0%) and loss of appetite (2.6%).

- **Conclusions**: Twitter posts confirm the use of Adderall as a study aid among college students. Adderall discussions through social media such as Twitter may contribute to normative behavior regarding its abuse.

3.1 Introduction

The mixed salt amphetamine Adderall, commonly prescribed as a treatment of Attention Deficit Hyperactivity Disorder (ADHD), is the most commonly abused prescription stimulant among college students [74]. Colleges, as well as medical and dental schools, report abuse rates of stimulant ADHD medications [97, 157] ranging from a low of 8.1% to a high of 43% [2, 154]. According to the National Survey on Drug Use and Health, 6.4% of college students aged 18-22 abused Adderall in the past year [228]. Given high academic expectations and competition in college settings, some students turn to prescription stimulants like Adderall as a study aid to improve concentration and increase mental alertness [16, 147, 232]. Rates of nonmedical use or abuse of ADHD drugs tend to be higher at colleges and universities where admission standards are higher [153]. A contributing factor to abuse of ADHD drugs is attention difficulties and the notion that these drugs can help with academic success [203]. DeSantis confirmed this finding and reported a higher tendency toward abuse among fraternity members during periods of high academic stress [62].

Other studies have affirmed racial and gender discrepancies in stimulant drug abuse as well as a correlation between prescription drug abuse and other illicit drug use among college students [62, 154, 232, 245]. Nonacademic motivations are also common and include, but are not limited to, counteracting the effects of other drugs, feeling a high, or as an appetite suppressor [232] as well as self-diagnosis of ADHD [113, 203]. A contributing factor for illicit drug abuse and prescription stimulant abuse among college students is the misperception that the vast majority of their peers use drugs [143, 197]. Elevated misperceptions about the

prevalence of drug use among peers are attributed to the traditional media's (eg, popular television depictions of college students using Adderall to gain academic advantages) portrayal of abuse. Misperception of reality is believed to be a leading contributor to increased levels of acceptance of abusive drug behavior, community norms for abuse, and higher levels of abuse [197]. Additional misperceptions such as the lack of danger of abusing prescription stimulants have also been found to contribute to justifications for illicit use [62].

Social media provides a relatively new and untapped resource for monitoring and understanding public health problems. As a surveillance tool, real-time data obtained through social media can be collected and analyzed quicker than traditional public health assessment tools such as questionnaires. In addition, research using social media provides an avenue for observing discussion between people in their natural interactions with one another, eliminating the Hawthorne Effect, where the presence of the researchers biases the response. Likewise, because people make statements as they occur, memory recall biases common with cross-sectional surveys or questionnaires are reduced. With the expansion of the Internet and social media, new fields of study such as infodemiology and infoveillance have emerged and represent "the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy" [70, p. 3].

Studies have demonstrated the utility of online information for understanding public health problems and their determinants. Using information obtained on trends in Internet searches, researchers have predicted outbreaks of influenza [68, 84, 198], listeriosis from contaminated foods [248], and gastroenteritis and chickenpox [195]. The feasibility of using online information for epidemiological intelligence purposes has led to the creation of proprietary systems such as Google Flu Trends, which is an online search query system that has demonstrated the ability to track regional outbreaks of influenza 7-10 days in advance of conventional Centers for Disease Control and Prevention (CDC) mechanisms for reporting [38].

In addition, Healthmap represents a public system for aggregating large amounts of online information (eg, news sources) for the purpose of monitoring global disease activity [30].

Recognizing the wealth of user-generated information produced by people through their participation with social media, researchers have begun tapping or mining this information to better understand health outcomes and even health behavior. For example, Corley, Cook, Mikler, and Singh [54] mined text data in the blogosphere for "influenza" and "flu." Their findings revealed trends in posts about the flu that were consistent with CDC report data. Several studies mined YouTube content for information relative to immunizations [118], H1N1 influenza pandemic [188], smoking cessation [17], cardiopulmonary resuscitation [167], kidney stones [220], and prostate cancer [223]. To date, no identified study has analyzed user-generated content in social media to describe the nonmedical use of Adderall.

The purpose of this study was to leverage the power of social media (ie, Twitter) to better understand Adderall abuse as a study aid among college and university students. More specifically, the following research questions were examined: (1) When do Twitter users typically tweet about Adderall?, (2) To what extent do tweets about Adderall abuse differ among various college and university clusters in the United States?, (3) What, if any, substances do Twitter users tweet about commonly abusing in combination with Adderall?, and (4) What common side effects are mentioned? Twitter was selected as the social media application for data collection because of its appeal with young adults including the ubiquitous research design advantages identified above. Twenty-six percent of all Internet users age 18-29 and 31% of all Internet users age 18-24 are also Twitter users [217]. Finally, using Twitter as a data source affords the ability to observe nationwide (and even international) behaviors simultaneously, as opposed to arbitrarily restricting a study to only a few regions. The use of social media data, in particular tweets, remains largely a novel concept for public health researchers. Questions surrounding the validity and utility of the data exist. Furthermore, little is known about the extent to which Twitter users might actually tweet about potentially sensitive health topics, such as Adderall abuse. Studies like the current one contribute to a

type of validity testing process whereby researchers can determine the extent to which trends in Twitter content coincide with documented patterns of behavior.

3.2 Methods

The following subsections define the methodology.

3.2.1 Procedures

Twitter is a popular online social media website in which users post status updates, or "tweets," that are limited to 140 characters. Public tweets are available and given without expectation of privacy. In addition, Twitter provides an Application Programing Interface (API), enabling programmatic consumption of the data. Specifically, the Twitter streaming API supplies tweets in real-time matching any given filter criteria. For example, using the keyword filter of "Adderall," all tweets mentioning the substance are collected.

In addition to the content of tweets, many users also provide location indicators [35]. Specifically in Twitter, users can potentially supply exact global positioning system (GPS) coordinates (eg, from a smart phone or other GPS-enabled device) or a GPS specified place (such as a neighborhood or city). Note that users providing only state or country level GPS were not included. Furthermore, tweets were excluded if they did not originate in the United States, based on GPS location. This GPS data can be used to associate a Twitter user with a nearby college/university. However, because many college campuses are within close geographic proximity, Twitter users may not necessarily have association with the campus to which they are physically nearest. Because of this proximity issue, rather than try to determine which of two nearby colleges should be used, the colleges were instead grouped into a cluster and treated as a single entity. Colleges and universities with a student population of 10,000 or more were identified using the National Center for Education Statistics database. Clusters were determined using hierarchical agglomerative clustering (HAC) [106, 110] with complete linkage with a cutoff distance of 150 miles. HAC produces a dendrogram of the

complete sequence of nested clusterings, as follows. HAC starts by assigning each college to its own cluster. Then, the two closest clusters are merged into a single new cluster. This pairwise merging process is repeated until a single cluster containing all of the colleges is obtained. Although we have a distance defined over colleges, HAC also needs a distance over clusters. Several distance measures may be considered, the most common of which are complete linkage, which uses the maximum distance between all pairs of objects across clusters, single linkage, which takes the minimum distance, and average linkage, which computes the average of all intercluster distances. We chose complete linkage here as it tends to create more compact, clique-like clusters [105]. Given the fully nested sequence of clusterings, the choice of a specific final grouping is typically made by selecting a level at which to cut through the dendrogram, and defining the clusters as the groups of elements hanging from the subtrees whose top branches intersect with the horizontal line corresponding to the chosen level. Our cut point of 150 miles means that pairs of colleges in a cluster were no more than 150 miles apart.

The student body population of a cluster was determined by summing the populations of each included college. Twitter users are then associated with the nearest college cluster if a college in that cluster is within 100 miles of the user's GPS location.

3.2.2 Measures

Keywords related to co-ingestion with other drugs, alternative motives, and possible side effects are shown in Table 3.1. A case-insensitive comparison was performed to count the number of tweets containing the keywords specified. Where multiple words are given as a single term, it was considered an exact phrase.

3.2.3 Data Analysis

After the tweets were obtained from the Twitter API, the data were imported to Microsoft Excel spreadsheets and then into SPSS version 20 for analysis. Frequencies, percentages,

Table 3.1: Search Terms for Alternative Motive, Co-ingestion, and Side Effects.

Topic	Subtopic	Search Terms
Alternative Motive (study aid)		test, final, finalsa, studya, studia, college, class, midterm, exam, homework, paper, essay, project, school, crama, quiz, assignment, all-nighta, allnighta
Co-ingestion	Alcohol-related	alcohol, wine, vodka, shots, patron, booza, margarita, mimosa, beer, drinka, bud
	Stimulants	coffee, caffeine, red bull, monster, no dose, no doze, 5 hour energy, five hour energy, rockstar
	Cocaine-related	cocaine, coke, crack, rock, freebase
	Marijuana	marijuana, MJ, pot, weed, grass, reefer, Mary Jane
	Anti-anxiety	xanax, tranquilizer, valium, beanies, ativan, benzoa
	Meth-related	crystal, meth, methamphetamine, amphetamine
Side Effects	Sleep deprivation	tired, awake, sleepa, slept, insomnia, restless, asleep, trouble sleeping,
	Anxiety	anxiety, anxious, antsy, jittera, shaka, nerva, nervous, uneasa, worry, tense, tension, dread, restlessa
	Teeth grinding	teeth, tooth, grinda, file, grata, grita, clencha, gnasha, scrapa
	Diarrhea	diarrhea, diarrea, diarhea, the runs, squirts
	Weakness	weaka, feeble, puny, scrawny
	Dizziness	dizza, faint, wobbly, shaky, lightheaded, light-headed, woozy, dazed
	Headache	headache, migraine, migrain, migrane
	Sweating	sweata, perspira, dripa
	Nausea/vomiting	nausea, vomita, throw up, stomach pain, stomach ache, upset stomach, puke, barf, heave
	Loss of appetite	hungry, food, eata, ate, weight, appetite, meal, thin, skinny, starva, slima, slender
	Obsessive compulsive behavior	can't stop, cleana, brusha teeth, washa hands, nails, nail-biting

means, medians, and standard deviations were used to describe the Adderall abuse. ArcGIS 10 was used to create maps of rates for GPS Adderall tweeters.

The Institutional Review Board (IRB) at Brigham Young University approved this study.

3.3 Results

Using the Twitter Streaming API with the keyword filter "Adderall," all tweets mentioning Adderall for the dates of November 29, 2011, to May 31, 2012, were collected. There were 14,282 tweets from users whose screen-names included "Adderall" or "pharm" that were removed from the sample because they were not representative of typical users, but rather those that were pushing or promoting Adderall or other pharmaceuticals. The resulting sample consisted of 213,633 tweets mentioning the term Adderall, from 132,099 unique user accounts.

The vast majority of tweets discussed Adderall use in a joking, sarcastic, or casual manner. Observed tweets included (original spelling and punctuation preserved): "I need adderall. Can't focus on studying or finishing these reviews", "this whole no adderall for the past 3 days is really getting to me #StillDoingWork #DontKnoHowTho", "Does anyone have adderall? #desperate", "adderall + school = winning", "wish i had adderall to get my room cleaned faster", "Adderall stockpile for finals", "We would all graduate with a 4.0 if adderall was sold over the counter", "Running on coffee and Adderall", "yay for adderall-induced optimism #givemeaprescription", and "Adderall, Coffee, Red Bull. Epic focus. Or a heart attack." Note that words beginning with "#" are hashtags, or user-defined topics that are often used in Twitter as a means of self-classification.

Table 3.2 lists the number of tweets matching each of the categories defined in Table 3.1. It should be noted that the results shown in Table 3.2 capture words that occur in the same tweet as the term Adderall. In this sense, they may be a conservative underestimate of actual events because it is possible that a user may tweet about Adderall but mention a side effect,

Table 3.2: Frequency Distribution of Adderall Tweets for Search Terms.

Topic	Subtopic	n	%
Alternative Motive (study aid)		27,473	12.9
Co-ingestion	Alcohol-related	10,229	4.8
	Stimulants	10,043	4.7
	Cocaine-related	1993	0.9
	Marijuana	1696	0.8
	Anti-anxiety	881	0.4
	Meth-related	788	0.4
Total Unique Co-ingestion Tweets		24,167	11.3
Side Effects	Sleep Deprivation	10,687	5.0
	Anxiety	1204	0.6
	Teeth Grinding	605	0.3
	Diarrhea	11	0.01
	Weakness	140	0.07
	Dizziness	77	0.04
	Headache	223	0.1
	Sweating	381	0.2
	Nausea/vomiting	154	0.07
	Loss of Appetite	5562	2.6
	Obsessive Compulsive Behavior	1937	0.9
Total Unique Side Effect Tweets		19,539	9.1

motive, or another substance in another tweet. Because subtopics are not mutually exclusive, some tweets match multiple subtopics and are counted for each. Thus, the total number of unique tweets for a topic is not a sum of the subtopic values.

3.3.1 Adderall Use by Hour, Day, and Week

Figure 3.1 illustrates the average number of Adderall-related tweets per day of the week, over the course of the study. Tweets tend to peak on Wednesday and reach a low on Saturday. As shown in Figure 3.2, the number of Adderall tweets per day varied significantly throughout the year, with consistently more tweets on the weekdays than the weekends. Large spikes in Twitter conversations were observed during the months of December and Mayduring traditional final exam times. The one-way ANOVA results indicate a significant difference between Adderall mentions between weeks ($P < .001$). Tweets regarding Adderall peaked December 13th at 2813 and April 30th at 2207 and dropped to a low of 292 on December 25th and 440 on May 27th. Over the course of 6 months while data were collected, the mean number of Adderall tweets per day was 930 with a median of 855. The large spike on May

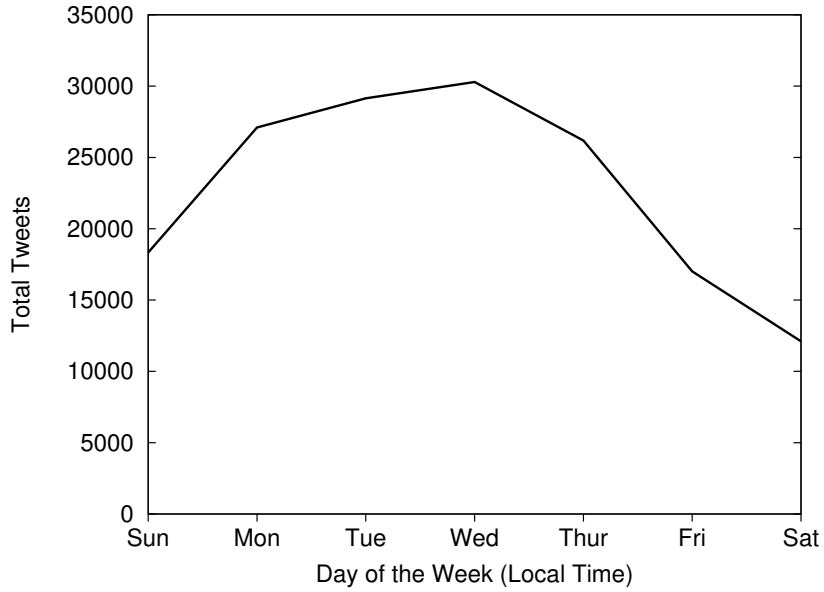

Figure 3.1: Adderall-related tweets by day of the week.

30-31 was attributed to a US Food and Drug Administration news release warning consumers of counterfeit versions of Adderall being sold on the Internet in response to its being on the FDA's drug shortage list [236]. This FDA news release was reported by news agencies, and links to the subsequent stories were tweeted by many users. The 10 days in the middle of April when no tweets were observed is the result of a failure of the investigators' servers.

3.3.2 College and University Clusters

Of the 213,633 tweets referencing Adderall, 27,473 (12.9%) also included reference to an alternative motive for use (eg, finals, studying, project, all-nighter), as shown in Table 3.2. Several of these alternative motives seem to be indicative of misuse among college-age students. To focus the analysis on college-age students, Adderall tweets were analyzed in clusters of colleges and universities that were within 150 miles of each other. A total of 586 colleges and universities in the United States were identified with a student body population of at least 10,000. Colleges and universities within 150 miles of each other resulted in a total of 87 clusters ranging in size from 1 to 48 colleges and universities in each cluster. The mean size of student-body population per cluster was 131,562, and the median was 93,281.

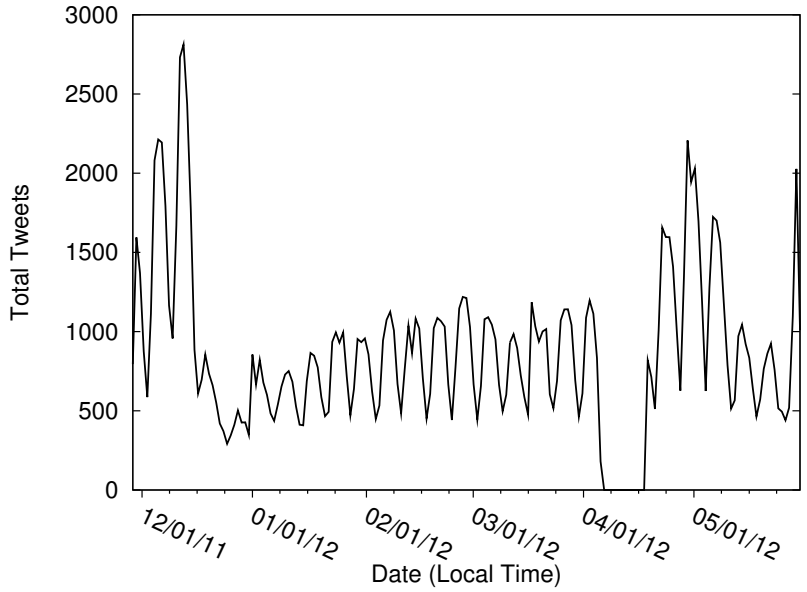

Figure 3.2: Distribution of Adderall-related tweets over 6 months.

Of the 132,099 unique users in the sample, 3698 (2.8%) provided GPS data. In order to restrict this set of GPS-enabled users to include only those users who are likely to be students, we obtained the 3200 most recent tweets (the maximum provided by Twitter) from each user with GPS data and searched these tweets for the following student-related terms: "homework", "teacher", "professor", "class", "final", "test", "exam", and "study." Of the 3698 users with GPS information, 2335 (60.7%) included one of these student-related terms in their tweets and are referred to as GPS Adderall Tweeters.

Figure 3.3 illustrates the 150-mile college clusters in the contiguous 48 states of the United States according to the rate of GPS Adderall Tweeters per 100,000 students, where the center of the circle is the average of the locations of the colleges in the cluster, and the size of the circle corresponds to the rate. Table 3.3 lists the ten clusters with the highest rates, and Table 3.4 lists the ten clusters with the lowest rates. Cluster identifications (ID) represent the state(s) to which the majority of colleges and universities in the cluster belong. As shown in these tables, the amount of GPS Adderall Tweeters per 100,000 students ranges from a high of 66.4 in the Vermont cluster and 54.6 in the Massachusetts cluster to a low of 1.4 in the South-Eastern Texas cluster and 2.1 in the Central Illinois cluster. Rates reveal

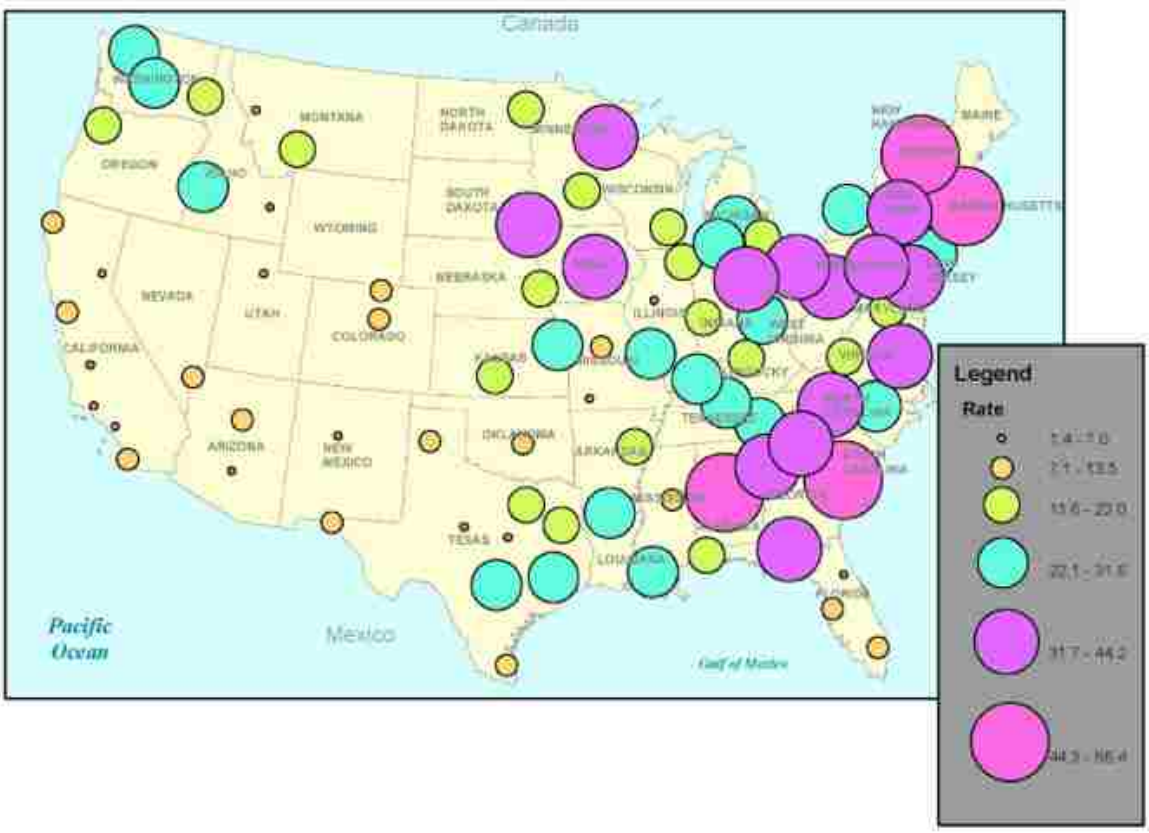

Figure 3.3: Rates of Adderall tweets by 150 mile college clusters in the United States (rate per 100,000 students).

a greater rate of GPS Adderall Tweeter in the northeast and south regions of the United States.

3.3.3 Co-ingestion and Side Effects

A total of 24,167 (11.3%) tweets also mentioned another substance along with Adderall in their tweet (see Table 3.2). Analysis revealed that the most common substance terms were alcohol-related (4.8%, $n = 10,229$) and stimulants, such as coffee or Red Bull (4.7%, $n = 10,043$). Other substances were cocaine-related (0.9%, $n = 1993$), marijuana (0.8%,

Table 3.3: Top 10 Rates of Adderall Tweets for 150-mile College and University Clusters in the United States

Rank	ID	Rate	GPS Adderall Tweeters	Total Cluster Population	Number in the Cluster
1	Vermont	66.4	9	13,554	1
2	Massachusetts	54.6	162	296,704	16
3	Alabama	52.2	38	72,748	3
4	South Carolina, Southern Georgia	48.8	57	116,891	6
5	Central Georgia	44.2	52	117,765	6
6	North Georgia, Southern South Carolina	44.0	36	81,773	4
7	Northern Florida	44.0	18	40,921	3
8	Southern Pennsylvania, Northern West Virginia	42.9	43	100,336	5
9	Ohio	37.3	54	144,659	7
10	Western North Carolina, Eastern Tennessee	37.0	41	110,718	6

Table 3.4: Bottom 10 Rates of Adderall Tweets for 150-mile College and University Clusters in the United States

Rank	ID	Rate	GPS Adderall Tweeters	Total Cluster Population	Number in the Cluster
77	Central Texas	5.8	3	52,076	3
78	Alaska	5.5	1	18,154	1
79	Southern California	5.5	55	1,008,210	48
80	Puerto Rico	5.0	3	60,579	4
81	Northern Nevada	4.5	3	66,242	4
82	New Mexico	3.9	3	77,236	3
83	Northern Utah, Southern Idaho	3.6	1	27,476	1
84	Northern California	3.5	4	115,026	7
85	Central Illinois	2.1	1	46,797	3
86	South-Eastern Texas	1.4	1	69,949	3

$n = 1696$), methamphetamine-related (0.4%, $n = 788$), and depressants, such as Xanax and painkillers (0.3%, $n = 728$).

Sleep deprivation (5.0%, $n = 10,687$) and loss of appetite (2.6%, $n = 5,562$) were the most common side effects associated with Adderall tweets (see Table 3.2). Diarrhea (0.01%) was the least common side effect mentioned followed by weakness (0.01%, $n = 140$) and nausea/vomiting (0.07%, $n = 154$).

3.4 Discussion

This study demonstrated the use of Twitter posts (ie, tweets) as a way to examine Adderall abuse among a sample of college students in the United States. More specifically, the study sought to determine: (1) When do Twitter users typically tweet about Adderall?, (2) To what extent do tweets about Adderall abuse differ among various college clusters in the United States?, (3) What, if any, substances do Twitter users tweet about commonly abusing in combination with Adderall?, and (4) What common side effects are mentioned?

Findings indicate that Twitter posts regarding Adderall vary across day of the week and week of the month. Consistent with traditional college final exams schedules, tweets regarding Adderall peaked during December and May. Similarly, tweets regarding Adderall peaked during the middle of the academic week and declined to fewer mentions over the weekend. These findings are consistent with previous research that has suggested that college students who abuse prescription ADHD stimulants do so primarily during times of high academic stress [62]. In addition, preexisting attention difficulties have been shown to be a predictor of nonmedical use of prescription ADHD medication in order for college students to experience greater academic success [203].

Grouping colleges within 150-mile clusters ultimately provided a mechanism for comparing geographic regions within the United States. Analysis of these college clusters revealed a concentration of GPS Adderall tweeters along the northeastern portion of the United States and in some of the southern states. The rates of GPS Adderall twitters per

100,000 students in the east and south clearly indicated greater Twitter conversations related to the use and abuse of Adderall. These findings are consistent with previous studies that examined the nonmedical use of prescription stimulants. McCabe [154] observed geographical patterns of nonmedical use of prescription stimulants with higher rates of use among college students in the north-eastern region of the country. Additionally, these findings are consistent with the Monitoring the Future study where higher rates of nonmedical methylphenidate use were found among college-age young adults in the northeastern region of the United States [111]. Other studies at select colleges in the east have shown high rates of nonmedical use of prescription stimulants [16, 147].

Additional research is needed to better understand the reasons for geographical variations in use. One possible explanation includes the fact that the U.S. fraternity/sorority system has deep historical roots at northeastern colleges and universities, and prevalence of nonmedical use of prescription stimulants is higher among fraternity/sorority members [154]. Future research might explore the link between nonmedical use of prescription stimulants and the geographical distribution of colleges and universities and their admission standards, student/family income, as well as the distribution of prescription drug monitoring program in the United States. Research has associated nonmedical use of prescription stimulants with competitive admission standards [154] and students coming from families with higher incomes [231].

Geographical findings can provide practitioners with evidence necessary for prioritizing intervention resources for targeting priority populations. This study has demonstrated how grouping can occur; however, and more importantly, it provides a social media solution for segmenting a broader population into more meaningful and manageable groups for intervention purposes. Colleges can be clustered in numerous different ways as needed and defined by researchers.

Because social media is, by its very nature, a social endeavor, the users' postings can have a great impact on the social norms of others. This is particularly relevant in the context

of drug abuse, where drug abuse behavior can be represented. Social norms theory suggests that individual behavior (eg, drug use) is influenced by individual perceptions of what is perceived as "normal" or "typical." This theory is rooted in Social Cognitive Theory [20] as well as the Theory of Planned Behavior [5]. In this light, the data that 8.9% of Adderall tweets mention another substance in the same tweet is significant because it may influence others to think that co-ingestion is normal and not dangerous. This is particularly troubling because it is through poly drug use or co-ingestion that morbidity and mortality risk increases. Poly drug use occurs among college Adderall abusers and combining Adderall with other stimulants like cocaine increases risk of heart attack and stroke [228]. Also in this regard, even tweets that are sarcastic, joking, or simply restating song lyrics, are relevant in their misrepresentations because of their impact on social norms.

Nearly 1 in 10 tweets included in this sample referenced a side effect of Adderall use/abuse. Effects relative to sleep deprivation and loss of appetite were discussed the most. Whereas more tweeters discussed an alternative motive for use (ie, study aid), individual tweeter perception of the benefits of Adderall use (eg, study aid) may outweigh the costs of use (eg, side effect such as irritability). Future research might further explore individual perceptions of Adderall side effects among college students to gain a better understanding of why some college students abuse, while others do not.

3.5 Limitations

These findings should be interpreted based on the following limitations. First, not every Adderall tweet is related to actual use. For example, we observed song lyrics that impact these counts, such as the two often quoted lines "College hoes love alcohol and popping adderall" and "I've been up for 3 days adderall and redbull." In our sample, there were 4,275 tweets that have the words "college hoes love" and 894 that have the words "been up for three/3 days". These numbers likely inflate the number of matches for "college", "alcohol", and "redbull" above the number of people tweeting about actually using these substances.

However, as discussed, even sarcastic mentions, or the quotation of song lyrics, are pertinent because of the impact they may have on social norms. Second, our study did not consider misspellings of the word "Adderall" or other ADHD medications, such as Ritalin. While our sample would have been increased by these inclusions, it is not likely that their absence resulted in any particular sampling bias. Third, our analysis focused exclusively on public tweets. It is unclear, and indeed difficult to assess, what the impact of other tweets (eg, direct messages) may have on our results. Fourth, our analysis focused only on colleges and universities with a student population of 10,000 or more. No attempt was made to designate whether the colleges and universities in this sample were on a quarter or semester system. Finally, the keyword approach to identifying college students may have included other students (eg, high school) or others that simply mentioned academic-related terms. While these additional users could inflate our overall values, we have no reason to believe they would be substantially biased toward different areas of the nation.

3.6 Conclusions

The twitter-based surveillance methodology in this study produced similar findings to traditional survey designs. In response to the noted research questions, Twitter posts regarding Adderall vary across day of the week and week of the month among users. Consistent with college traditional final exams schedules, tweets regarding Adderall peaked during December and May. Similarly, tweets regarding Adderall peaked during the middle of the academic week and declined to fewer mentions over the weekend, which suggests that college students who abuse prescription ADHD stimulants do so primarily during times of high academic stress.

Additionally, tweets about Adderall abuse differ among various college clusters in the United States. Using 150-mile college clusters, regional comparisons identified a concentration of GPS Adderall tweeters along the northeastern portion of the United States and in some of the southern states, and thus indicate greater Twitter conversations related to the use and

abuse of Adderall. Further, co-ingestion of other substances, notably alcohol, stimulants (such as coffee or Red Bull), cocaine-related, marijuana, methamphetamine-related, and depressants (such as Xanax and painkillers), are the substances most commonly mentioned with Adderall. Such poly drug use or co-ingestion is known to increase morbidity and mortality risk. Finally, the most common side effects associated with Adderall tweets include sleep deprivation and loss of appetite. Thus, Adderall abuse is associated with college or university life. Given the risks and trends for Adderall acceptance among college-age students, there is a need to renew interest and priorities to influence college campus norms, promote the safe and legal use of these substances, and promote stronger student wellbeing and study habits to better manage the academic demands and pressures that are typical on college campuses in the United States.

Chapter 4

Leveraging Social Networks for Anytime-Anyplace Health Information

Abstract

The health landscape is shifting to one in which common individuals are no longer merely consumers, but also producers, of health information. We demonstrate that social media platforms provide the means to seek and receive personalized, credible health advice from peers at any place and time, by tracking dental health advice sought and received in Twitter. We show that for genuine dental advice-seeking questions, answers are received 32% of the time, with the first reply coming less than 6 minutes after the question is posed, in the median. We compare our results to studies focusing on generic questions and find stronger relationships between users that answer health questions. Additionally, we find that users with more social capital, in the form of more reciprocal follower/following relationships, are more likely to receive responses and receive them faster, and are thus better able to leverage their social networks in receiving advice.

4.1 Introduction

Historically, individuals have been regarded as patients, or mere consumers of health information and care, provided directly, and controllably, by experts including medical doctors and other health practitioners. The advent of the Internet and the proliferation of online content of all kinds, including health-related content, means that more information is now readily accessible by lay individuals, thus enabling health information consumption on a much larger scale, and independently of the traditional channels of distribution (e.g., doctor's

office, hospitals). Furthermore, individuals are not limited to simply looking up information; they are also able to use such technology as blogs and online forums to discuss, comment on, and share experiences about various health issues they themselves, or their loved ones, may be facing. In doing so, individuals begin to change slowly from being only consumers of health information to becoming producers of the same. What they share with others about symptoms, side-effects, remedies and other relevant experiential knowledge becomes information for others to consume. We regard this as the first phase of the transformation of the health care landscape.

The second phase, and the one really responsible for a major shift in attitude and behavior, is the emergence of rich, interactive social media applications, such as Facebook, MySpace, YouTube or Twitter. These applications allow individuals to connect, collaborate, and exchange their current thoughts, feelings and activities with one another without concern for geographical boundaries. What that means is that whereas the role of providing support and advice regarding health issues has generally been limited to health care providers one-on-one visits, and to close associations in one's family or small network of friends, it may now be extended to all participants in one's social network. Hence, we are witnessing a dramatic paradigm shift in health care. The one-way flow of health information and solutions from health care professionals who produce them to lay individuals who consume them is gradually being replaced by a more fluid and distributed flow where any and all individuals, professionals or otherwise, may act as both consumers and producers of health information, advice and solutions. This new state of affairs is of course not without its own challenges, including privacy, quality and trust issues. We do not address these here, however, but focus instead on showing how social media are indeed being used to seek and receive health advice.

A 2008 survey indicates that a large number of people turn to the Internet (59%) and social media (34%) for health information [66]. The fact that turning to search engines is often the first action people take for health questions is precisely what enables identifying

early-stage outbreaks through search query tracking [30, 84, 195, 248]. When it comes to personal health, there are however, some limitations to the Internet:

1. One has to search for the needed information and possibly wade through many results.
2. The information tends to be of a generic nature and hence responses are not personalized.
3. The information available is limited to what authors have already posted, which may or may not include what one is looking for.
4. The credibility of the information is a concern as inaccuracies can arise from under-informed people sharing opinions, as well as businesses and other invested parties promoting their own agendas or manipulating the content to their own ends [59, 60, 80, 100, 166].

Some level of context and credibility (or trust) can be established through focused social media groups [88] and stand-alone e-communities such as *Patients Like Me*[1], which provide opportunities for people with common conditions to connect. However, despite sharing common conditions, users of these forums often have little history outside these interactions. The blogosphere is another rich source of health data [161, 176, 243], where users are sometimes familiar with blog authors, either personally or through consistent online interaction and following. So while unknown bloggers carry the same risks to validity as other Internet sites, trusted authors can provide a sense of integrity. However, even respected authors and sites can only be probed for existing information, and not questioned in real-time for advice.

Asking questions in social networks provides a natural mechanism to overcome all of the above limitations, since it exhibits the following characteristics.

1. *No Search.* One needs not search for answers but simply ask questions to his or her network and wait for answers to come.
2. *Personalization.* One is more likely to receive personalized answers because such answers are to a specific question and come from people one knows.

[1] www.patientslikeme.com

3. *Versatility.* One may obtain information about almost any topic, including some that are not easily obtained through search engines [164].

4. *Credibility.* Answers and advice received from one's network are more credible, or carry a higher level of trust, since social network connections are based on established relationships (either in the real or virtual world).

Furthermore, social media provides the additional advantage that questions may be asked at any time, which is desirable since health needs and questions arise in many different settings, often outside of the doctor's office or hospital. And responses may also be received at any time, often shortly after the question has been asked (*Timeliness*), as we shall see.

In this paper, we demonstrate the value of social media for health advice seeking, as discussed above, by showing how people use Twitter to ask and receive advice about dental health issues. Twitter is particularly attractive for this type of study for a number of reasons. First, it is a very rich source of timely, spontaneous, and uncensored excerpts of users' emotions and activities. It is estimated that over 200 million tweets are generated each day. Second, Twitter implements a one-to-many broadcast communication mechanism in which a user may pose a question to all of his or her followers at once. And finally, Twitter possesses a rich application programming interface (API) that allows information to be filtered and/or searched programmatically.

The contribution of this work is two-fold. First, we establish that despite possible concerns of anonymity or privacy, social media users are seeking advice, and receiving responses, in at least some areas of health (in our case, demonstrated by dental advice in Twitter). Second, we highlight social factors that contribute to speed and quantity of responses.

The remainder of the paper is organized as follows. First, we discuss related work and then outline our methodology for finding dental advice-seeking questions and their responses on Twitter. Next, we present our results followed by a discussion of the implications of our

findings and the differences of our results compared to general question and answer research. Finally, we offer conclusions and suggest areas for future work.

4.2 Related Work

Social network analysis is becoming increasingly important in health and bioinformatics as relationships are modeled between people, cancer cells [27], or even between related diseases and genes [250]. The impact of online social networks on public and personal health is increasingly being recognized [47, 69, 116, 149, 213, 239]. Several recent studies have specifically identified health topics in Twitter data. Scanfeld et al. [214] mined Twitter content and demonstrated that social media provides a means for sharing health information, especially as it relates to antibiotics misuse and understanding. Paul and Dredze [192] employed topic modeling to 1.5 million tweets and were able to discover that numerous health related conditions (e.g., allergies, obesity and insomnia) were mentioned in the tweets. Prier et al. [201] were able to identify tobacco related conversations through Twitter. Chew and others conducted content analyses of tweets on H1N1 and swine flu mentions and demonstrated the value of using the tool for monitoring pandemics [13, 46, 131]. Finally, in the area of dental health, Heaivilin et al. [96] characterized tweets relating to dental pain. We build upon their work by considering dental advice being sought, as opposed to merely statements of pain, and also identify the answers received to the dental questions.

Ma et al. [149] have shown that online social interactions may carry enough positive peer pressure to encourage healthy behavior. It has also been found that, while in some cases anonymity may promote increased antagonism [137], adolescents generally feel more comfortable discussing potentially embarrassing topics with some degree of anonymity, as afforded by chat rooms and bulletin boards [87, 229]. While these studies have shown the effectiveness of several Internet tools, such as bulletin boards and chat rooms, to our knowledge, health advice seeking has not been studied in social media platforms, such as

Twitter, which introduce a different dynamic of at least partially-surrendered anonymity because of explicit connections to either real- or virtual-world friends.

Advice seeking presupposes the formulation of question to be asked of one's social network. Identifying questions is a non-trivial process, especially in micro-text posts where space limitations discourage proper grammar and promote abbreviations and slang, which produce challenges for traditional natural language processing techniques such as part-of-speech tagging [61]. Because of the difficulty of directly applying NLP techniques, to find questions for study and analysis, other approaches have been taken. Morris et al. [164, 165] were the first ones to study the use of social media for asking questions. In their work, however, they do not analyze media content directly, but rely instead on survey techniques where a number of individuals were asked about their experience in asking questions, receiving responses and providing responses themselves on Twitter or Facebook. Efron and Winget [64] look at tweets directly and employ a keyword approach. We adopt a keyword-based approach focused specifically on finding advice-seeking questions, as opposed to the more general topic of all interrogative statements. In that sense, we are influenced by the work of Paul et al. [193, 194], who analyzed questions and answers found in Twitter based on the presence of a question mark, and then used Amazon's Mechanical Turk to restrict the candidate set to valid questions, as judged by the turkers. We follow a similar approach, where we first identify likely advice-seeking questions related to dental health, and post-process them through human readers to increase precision.

4.3 Methods

To illustrate the value of social media in seeking and receiving health advice, we focus on dental health issues in Twitter. The topic of dental work is of general interest, because all people must manage their dental health to some degree. It also provides an area that people are generally comfortable discussing and where the vocabulary is accessible to common individuals. The common vocabulary of dental health, and the fact that complex medical

terminology is typically not used, helps in identifying dental advice, and better enables a keyword-based approach.

Because the Twitter platform limits tweets to 140 characters, it may inherently promote questions and responses that are less complex or elaborate. On the other hand, the simplistic nature of tweets may also cause more directly-asked questions and more succinct responses. The direct nature of tweets is helpful to our study, wherein we are seeking to determine if health advice is being sought and obtained.

We received an exemption from the university Internal Review Board to study these public-facing tweets.

4.3.1 Observing Dental Tweets

The first step to identify dental advice, is obtaining a sample of tweets on the dental topic. Twitter provides a streaming API that returns a portion of the complete stream of tweets filtered by a search query. To identify potential dental tweets, we filtered the Twitter stream by the keywords: "tooth," "teeth," "dental," "dentist," "gums," "molar," "moler," "floss," and "toothache." This keyword-based filter does not guarantee that all resulting tweets are related to dental health. For example, tweets containing the words "sweet tooth" or "molar mass" will pass through our filter even though they clearly have nothing to do with dental issues. However, this simple mechanism provides a good starting point.

Using our filter, we observed all tweets for two separate weeks, from October 26 to November 1, 2011, and from November 9 to November 15, 2011, and received a total of 1,032,754 tweets over the 14-day period, for an average of approximately 74,000 tweets per day.

4.3.2 Identifying Advice-seeking Questions

Twitter essentially implements a broadcast, one-to-many, form of communication in which a user posts messages (status messages, reactions to current events, questions, etc.), generally

intended to be read by all of that user's followers. This is an ideal mechanism for soliciting advice, because the question can be posed once to multiple potential respondents, as opposed to, for example, making individual phone calls to friends. Probably due in part to this broadcast-style of communication, we have observed that many advice-seeking tweets tend to contain words such as "anyone" or "anybody" with a question mark at the end of the sentence (e.g., "anyone know of a good dentist in Lancaster?", "Cold and sore throat has developed into painful tooth/mouth ache. This one's totally new to me. Can anyone enlighten me?").

Interestingly, Morris et al. [165] who characterized questions on Twitter, found that, in their set, 81.5% of questions contained question marks, and 20.9% contained the word "anyone." In their work on characterizing questions on Twitter, Paul et al. [194] simply used the question mark to identify questions, which allowed them to find more questions, many of which, however, were rhetorical. Because we are not concerned with categorizing all questions, but rather, are focused on genuine, health advice-seeking questions, we have found the use of the additional anyone/anybody criterion to help in removing some of the rhetorical, sarcastic, and advertising questions.

From the roughly one million potential dental tweets, looking for the words "anybody," "anyone," or 'any1," together with a question mark, we identified 2,035 candidate dental advice-seeking questions. To further improve the precision of our set of questions, we followed an approach similar to Paul et al. [193], except that, since we had several available, we used willing volunteers rather than Amazon's Mechanical Turk. In all, we had 18 independent individuals read the candidate tweets and manually classify them. Each person classified approximately 200 tweets, according to the following criterion:

> *Mark the tweet as a health advice-seeking question if it seems clear that the individual posting the tweet is asking for advice about a dental health issue regarding themselves or their family, with the expectation of receiving a response.*

The condition about "themselves or their family" allowed us to eliminate generic questions and questions about pets, while the condition about "expectation of receiving

a response" helped us focus on questions most likely to seek timely advice. Each tweet was independently classified by two different people. The separate classifications were in agreement in 87% of cases, and the remaining tweets were arbitrated by the authors. Of the 2,035 candidate questions, 432 (21%) were labeled as dental advice-seeking questions, such as: "does anyone know how long it takes for swelling on your mouth to go down after getting teeth out?" and "Can anyone suggest some home remedies for a #toothache?". Many of the tweets not matching the above criterion were in fact valid questions but did not seek dental advice or did not seem to be expecting an actual response (e.g., "Going to the dentist this morning. Anyone want to trade? I'll even throw in my best marble!", "anyone know how do to the putty for vampire teeth?" (sic.))[2].

4.3.3 Identifying Responses

One of the shortcomings of the Twitter API is that it does not allow direct querying of responses to a particular tweet. To overcome this limitation, we used the Search API to identify any tweets after the question was issued that were directed to the author of the question, using the *@username* syntax. Then, using the detailed REST API, we examined each of these possible replies individually to determine if it was listed as being "in-reply-to" the original question, as specified by a meta-data field of the tweet. While it is possible that users might respond by simply creating a new tweet addressed to the author, we assume that most users actually make use of the "reply" feature of the Twitter website (also available in most popular 3rd-party applications), which ensures that the reply-to meta-data field is correctly populated. Because of this assumption we may overlook some replies, causing some of our results to be underestimates, but we can have high confidence that the responses we identify are truly replies to the original question.

Because Twitter is asynchronous, there are potentially many different strands of conversation occurring simultaneously. This means that a user may pose dental questions to

[2]Note that one of our weeks of study included the Halloween holiday, which resulted in several questions about costume elements such as vampire teeth

Table 4.1: Tweets at each stage of the experiment

Set	Tweets	Percent of Previous Set
Matching dental keywords	1,032,754	
Candidate advice-seeking questions	2,035	0.2 %
Quality advice-seeking questions (human verified)	432	21.2 %
Questions that received answers	140	32.4 %

their followers but then continue to converse with others about different topics and receive a response later. However, the longer it has been since a question was asked, the less likely it is to receive a response. Paul et al. [194], for example, observed that 67% of their responses came within 30 minutes and 95% came within 10 hours. Because it is possible that health-advice replies may take longer than replies to other questions, but still recognizing that they become less likely over time, and less relevant, we searched for those occurring within 48 hours of the original question tweet.

We applied this process of determining dental replies to our 432 dental advice-seeking questions, and found that 140 (32%) received at least one response. In the median case, the first response was received 5.5 minutes after the question was asked. As noted, because our approach focused on minimizing the number of false positives, we cannot deduce that the other questions, for which we did not identify a response, were truly left unanswered. The number of tweets at each stage of the experiment are shown in Table 4.1.

4.4 Results

Because *Timeliness* is one of the desirable properties of social media, we feel it is useful to determine the time of day and week when questions were asked. To do so, we converted the question tweet's time to the user's local time wherever a time zone was listed on the user account. Week days are defined as Monday-Friday from 8 a.m. to 5 p.m. Week nights are defined as Monday-Thursday from 5 p.m. until the next day at 8 a.m. Weekends are defined as Friday at 5 p.m. until Monday at 8 a.m. Finally, "any after hours" is a combination of week nights and weekends. The results are summarized in Table 4.2 and Figure 4.1.

Table 4.2: Distribution of questions and responses by time of day and week

	Total	Receiving Replies	Without Replies	Reply Percent
Week Day	123	48	75	39.0
Week Night	105	44	61	41.9
Week End	72	21	51	28.4
Any After Hours	177	65	112	36.3
No Time Zone	132	27	105	20.5

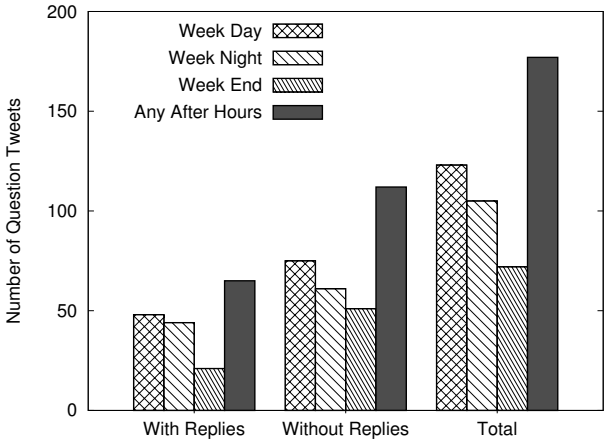

Figure 4.1: Number of questions occurring during week days, week nights, and week ends. "Any after hours" includes both week nights and weekends, and represents the majority of the questions.

As shown in Table 4.2, the majority of the dental advice-seeking questions were posted during the evening and weekend hours, which is not necessarily surprising given that we observed that Twitter activity is highest in the evening in general. However, this may be particularly relevant in the context of dental advice because it represents advice sought when traditional channels, namely dentist offices, are not available. It is interesting to note that over 36% of the questions asked after hours received answers, with slightly more of the week nights questions being answered (42%) than the weekend questions (28%). The latter could be explained by the fact that users may be less apt to consume others' content on the weekends, possibly catching up on their feeds on Monday morning, thus responses, even if they were to be given, would likely appear beyond our 48 hour limit, at which stage they would also have become much less useful as more traditional channels would have re-opened.

Table 4.3: Advice-seeking questions receiving replies and relationship to ego network

	Total	Percent	Median Followers	Median Following
Receiving Replies	140	32.4	331.5	256.5
Without Replies	292	67.6	136.0	157.0

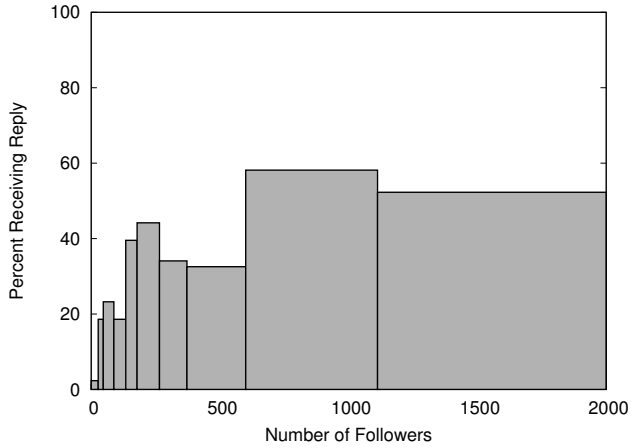

Figure 4.2: The percent of questions receiving replies based by number of followers, grouped into 10 bins of equal-question frequency.

While we cannot measure *Credibility* directly, we do, as others have (e.g., see [194]), look at the influence that an individual's ego network (i.e., its followers and its followings) may have on the responses they receive. As shown in Table 4.3, those users who received replies had significantly (unpaired, two-tailed t-test, $p = 0.005$) more followers (median of 331.5) than those that did not receive replies (median of 136). This is further demonstrated in Figure 4.2 which shows the percent of questions receiving answers based on the number of followers, grouped into 10 bins of equal frequency with regard to the number of questions. While on average 32% of questions received replies, users that had more than 200 followers had their questions answered 45% of the time, and users will less than 100 followers received answers in only 14% of cases.

Additionally, as shown in Figures 4.3 and 4.4, those users with more followers received their first replies faster on average. For questions that received replies, Figure 4.3 shows the delay between the question and the first response based on the number of followers, and Figure 4.4 shows the same data, grouping the questions into 10 bins of equal frequency with

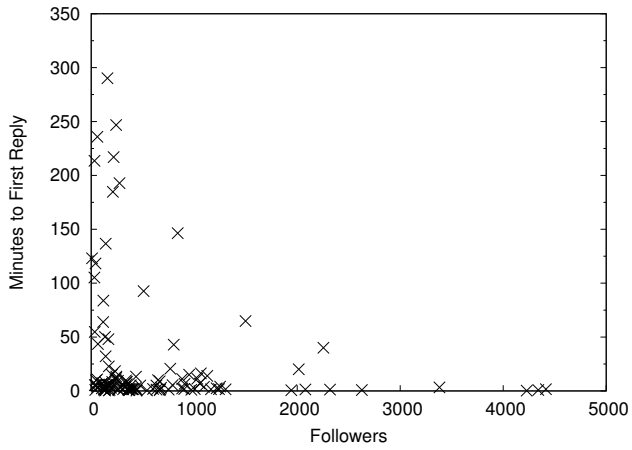

Figure 4.3: Time taken to receive the first reply versus number of followers.

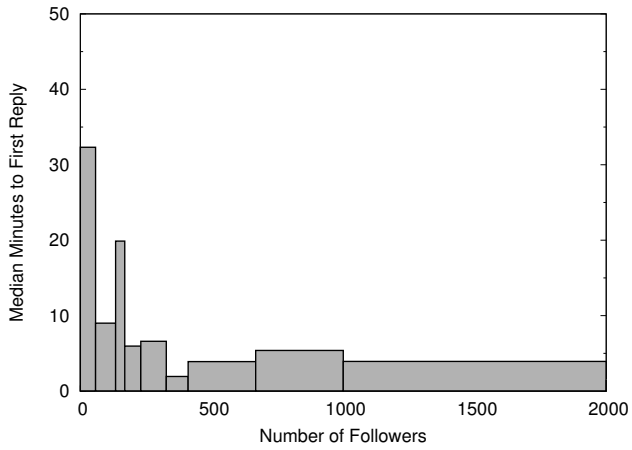

Figure 4.4: Median number of minutes to the first reply by number of followers, grouped into 10 bins of equal-question frequency.

respect to the number of questions. For questions that received answers, the *number* of replies also correlates positively with the number of followers (Pearson's $r = 0.48$), suggesting that users with more followers are not only more likely to receive responses, but are more likely to receive more of them.

In addition to considering followers and following independently, it is interesting to look at the reciprocity of the relationship between the asker and responder, as this provides a better indication of the strength of the relationship between the two. As shown in Table 4.4, 93% of responses came from users following the person asking the question and 69.5% came from users with a mutual following/follower relationship.

Table 4.4: Reciprocity of relationships between askers and responders

Relationship	Amount	Percent
No relation	16	6.6
Responder following asker	226	93.0
Asker following responder	170	70.0
Mutual following and follower	169	69.5

4.5 Discussion

Our finding that, overall, 32% of advice-seeking questions received answers is significantly higher than the rate of 9% that Paul et al. [194] observed for personal and health-related questions. This may be due to a number of factors, such as dental topics being less sensitive or personal than other health topics, but most likely it is the result of our focus on questions where a response was actually expected, rather than including rhetorical ones. The fact that users that are genuinely seeking advice receive it 32% of the time suggests that Twitter is a valid resource to turn to for personalized answers. And, seeing that users with more than 200 followers received answers to 45% of their questions and those with less than 100 followers only received answers to 14% of questions demonstrates that users with more social capital are better able to leverage their network to receive value—in this case, health advice.

The implicit social capital graph among Twitter users implies a weighting of different connections, where a person values their relationship with others at very different levels, ranging from very little weight with unknown users to a strong connection with close personal friends or family members. While the actual weighting of the graph would be very fine-grained, the following/follower structure of Twitter could provide a coarse approximation where low value exists between users that are not following or followed by one another, and high value exists in mutual follower/following relationships. Possible coarse approximations for social capital values from the asker's point of view are summarized in Table 4.5.

The fact that having more followers results in a greater likelihood of receiving a response as well as more timely responses is not necessarily surprising. Statistically speaking, the simple fact that more people are likely to view the message means that it is more likely to

Table 4.5: Coarse approximations of asker's view of social capital value

Relationship	Value	Reason
No relationship	Low	Users may not know each other
Followed by responder	Low–Med	Asker may not know responder
Following responder	Med–High	Asker trusts responder but responder may not know asker well
Mutual relationship	High	Asker trusts responder who can give personalized advice

be seen, and seen sooner. However, this result also says something about the *Personalization* available in social media. Indeed, people generally invest a lot of time and energy into building their social networks. In the case of Twitter, this means following other users, as well as responding to questions and posting relevant status updates regularly in an attempt to gain followers. In doing so, users create social capital and maintain a list of followers who come to know them. The more social capital an user has, the better his or her chances of getting timely responses, and of obtaining responses that are more personalized.

Furthermore, since 69.5% of the answers came from responders who had a mutual relationship with the asker (both following and followed by), we may be able to argue that the responses are not only more personalized but also more credible. This number (69.5%) of reciprocal relationships is significantly higher than the 36% found by Paul et al. [194]. The difference may be due to the fact that giving health advice is more personal than answering other questions. Thus, the mere act of answering a health question may indicate a strong relationship between the two users. In any case, these results suggest that success in obtaining advice on social media may be directly related to an individual's social capital. Others have similarly suggested that answering the questions of others could be used as a means to increase one's social capital thus resulting in higher chances of having one's own questions answered [165].

The short delay to answer is rather remarkable. In the median case, the first response was received within 5.5 minutes of the question being asked. Given that many responses were given outside of normal office hours, this suggests that Twitter may be effective at handling non life-threatening health emergencies.

4.6 Conclusions and Future Work

Social media offer unique opportunities for people seeking health advice in that information may be obtained in a more timely manner, on a potentially broader set of issues than present in other media (e.g., Internet), with increased credibility and better personalization. We have used Twitter and dental health issues as an example to demonstrate that 1) people do ask dental health related questions on Twitter; 2) a large number of questions are answered; 3) users receive timely advice after business hours thus making social media a valuable addition to traditional channels; and 4) the pattern of connections between askers and responders suggest that social capital is a determinant factor in the process.

The fact that advice can be obtained from established relationships, in particular mutual follower/following connections, provides an increased level of personalization and trust over anonymous Internet forum posts. And the fact that users with higher social capital are better able to leverage their networks for health advice demonstrates the value in building and maintaining on-line social relationships.

There are several interesting areas of future work. First, we have done nothing here to test the validity of the responses received, but have assumed that since they came from "trusted" sources, they too could be trusted. It would be interesting to test this hypothesis formally, perhaps involving subject matter experts to evaluate the actual quality (and safety) of the health advice offered. Second, while we obtain promising results with dental issues, we would need to repeat our study with other health topics to see whether the results generalize or whether there are any differences across health topics, possibly due to the sensitivity of the topic. Finally, we have discussed social capital and argued that there was evidence that social

capital had a direct impact on one's ability to obtain answers to advice seeking questions. Again, this result deserves more analysis. Furthermore, it would be interesting to expand the study of the role of social capital on Twitter by checking whether people are indeed more likely to turn to Twitter (or some other social media) than to a less personal medium, such as the Internet, to get answers to their question. Also, recognizing that social networks are dynamic, it would be valuable to study (and potentially predict) how the network might change as a result of asking or answering questions, especially recognizing that links could be both added and dropped as a result of this interaction [7].

Part III

Capitalizing on the *Social* of Social Media: Integrating the "Who"

S. Burton, R. Morris, J. Hansen, M. Dimond, C. Giraud-Carrier, J. West, C. Hanson, and M. Barnes. Public Health Community Mining in YouTube. In *Proceedings of the Second ACM International Health Informatics Symposium (IHI 2012)*, pages 81-90, 2012.

S.H. Burton and C.G. Giraud-Carrier. Discovering Social Circles in Directed Graphs. *In Submission*, 2013.

S.H. Burton, C.V. Tew, S.S. Cueva, C.G. Giraud-Carrier and R. Thackeray. Social Moms and Health: A Multi-platform Analysis of Mommy Communities. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (in press), 2013.

C.L. Hanson, B. Cannon, S.H. Burton, C. Giraud-Carrier. An Exploration of Social Circles and Prescription Drug Abuse through Twitter. *In Submission*, 2013.

Chapter 5

Public Health Community Mining in YouTube

Abstract

YouTube has become a vast repository of not only video content, but also of rich information about the reactions of viewers and relationships among users. This meta-data offers novel ways for public health researchers to increase their understanding of, and ultimately to more effectively shape, people's attitudes and behaviors as both consumers and producers of health. We illustrate some of the possibilities here by showing how communities of videos, authors, subscribers and commenters can be extracted and analyzed. Tobacco use serves as a case study throughout.

5.1 Introduction

One of the difficulties of research in public health is determining —and ultimately influencing— the perception and reaction of the public with regard to health-related issues. Typical approaches, such as questionnaires (e.g., NHANES, HINTS), can be difficult and costly to administer. Furthermore, processing results and preparing them for analysis are tedious activities that cause studies based on questionnaires to be delayed and thus to lag behind important, relevant, and detectable, social media health communications, which arise spontaneously and much faster.

Social networking websites, such as Facebook, MySpace, Twitter and YouTube, contain vast amounts of content that is constantly being updated by the public as a whole. One of the great advantages this offers is the ability to observe, in a timely manner, the attitudes and

behaviors of people in their natural interactions with others. There has thus naturally been a growing focus on the importance of online social media in public health research [47, 239]. Recent studies have, for example, shown that online social interactions may carry enough positive peer pressure to encourage healthy behavior [149]. It has also been found that, while in some cases anonymity may promote increased antagonism [137], adolescents generally feel more comfortable discussing potentially embarrassing topics with some degree of anonymity, as afforded by social media [87, 229]. For public health practitioners, social media offer yet another significant advantage in that the interactive nature of Web 2.0 applications facilitates not only observations, but more importantly intervention, such as through tweets or chats [56].

In this paper, we focus our attention on YouTube. With the ability to easily post, view, and comment on videos, YouTube allows the ideas of a single user to be seen by millions in a matter of days. In addition to the obvious video content, YouTube is also a repository of rich meta-data giving relationships among related videos, users, and comments. Indeed, while it was designed primarily as a video-sharing platform, each entry or submission to YouTube goes far beyond the video content and author's name alone, to include such things as a list of author-defined tags or keywords to describe the video, a list of subscribers (i.e., people who "follow" the video's author), comments left by viewers, ratings left by users (now simplified to *like* or *dislike*), statistics collected by YouTube, such as number of views and viewing history, and a ranked list of links to 20 related videos, as determined by YouTube's proprietary algorithm based on viewers' clickstream data, recency, etc.

Unlike some who have argued that medical research based on statistics from YouTube may lend false credence to a "conduit of popular culture" [95], we believe that YouTube remains a valuable medium both for observing and for interacting with the public. Additionally, it has been noted that health topics are already being discussed in social networks, and in many cases the associated communications are dominated by businesses that have vested commercial interests [239]. It would seem not only reasonable, but in fact desirable, for the

public health community to take advantage of YouTube, and other social media, to ensure that accurate, constructive and health-promoting viewpoints are widely represented and adequately expressed.

While much of what is presented here extends in principle to other social media platforms, there are several features of YouTube that make it particularly well suited to our applications. In particular, 1) YouTube is an open forum, 2) YouTube's data is rich in natural relationships among its elements (e.g., friends, comments, related videos), and 3) YouTube possesses a rich application programming interface (API) that makes almost all of its data available for easy consumption by data mining tools.

We take advantage of these features here and show how social network analysis tools can be used to build and analyze communities of videos, authors and comments from YouTube's rich data. We give examples of the value of these communities with regard to public health, with specific emphasis on tobacco usage as a relevant case study.

5.2 Related Work

There is certainly no way for us to be exhaustive here about work in social network analysis or the use of social media in public health. However, we highlight several pieces of work most relevant to our own in the context of YouTube and community mining.

The increase of YouTube's popularity and the accessibility of its audio/visual material, textual comments, and friendships, is leading public health researchers to leverage this oracle to public perception and interaction. Most of the studies so far have focused exclusively on the content or message of videos returned by certain keywords. For example, videos have been examined for their potential role in implicitly influencing normative beliefs formation or for their explicit attempts at eliciting positive or negative sentiment in areas as varied as vaccinations/immunizations [118], recreational partial asphyxiation (i.e., the choking game) [144], and tanning beds [100], with a significant body of studies specifically targeted at smoking behavior [17, 76, 80, 122, 186]. Our research goes beyond content. It is interested in

how videos and authors are connected to each other, and how such networks can inform our understanding of health issues.

While others have studied structural properties of YouTube as a general social network (e.g., see [189, 212]), our work focuses on the unique notion of community. Indeed, social networks differ from other types of networks, such as technological or computer networks, in many ways that can be traced to the fact that they are inherently composed of communities [180]. Understanding these communities with regard to health concerns can lead to valuable research insights, yet discovering these communities within the context of YouTube is a non-trivial computational problem. At least part of the difficulty arises from the fact that many community mining algorithms depend on a complete enumeration of the network [43, 121, 178, 179, 246, 251]. Yet, YouTube does not make available a complete list of its videos, and even if it did that list would be much too large for its enumeration to be computationally feasible. Recently, new algorithms have begun to emerge that perform community discovery through a controlled iterative process [42]. We follow and extend this latter approach to community building here.

5.3 YouTube Communities

One approach to defining a social network from YouTube data is to 1) consider videos as nodes, and 2) use the related video list provided by YouTube to define the edges of the graph [44]. An alternative method still views the videos as nodes, but defines the edges based on "video responses" posted by users in response to an original video [23]. Once a social network of videos is defined, a social network of the users that authored those videos can also be derived rather straightforwardly [23].

We capitalize on the richness of implicit relationships embedded in YouTube's data to build on these ideas. Indeed, while YouTube is essentially an extensive network of videos, the additional information available in tags, friends' lists, subscribers' lists, and comment trails can be used to build various focused communities of videos, authors and commenters

relevant to public health research. In what follows, we present a basic analysis of several of these communities and illustrate their value in the context of tobacco usage.

5.3.1 Video Communities

As stated above, YouTube does not make available a complete list of its videos, and even if it did, that list would be unmanageable due to its sheer size. Hence, it is not possible to use most community building algorithms, which require a knowledge of the complete social network. Instead, an iterative approach to building communities must be employed, beginning with one or more (seed) videos and expanding from that point. A mechanism for this iterative expansion consists in exploiting the set of related videos provided by YouTube alongside each video, and also available through the public API. While the details of the algorithm used by YouTube to produce the set of related videos are proprietary, the related videos are a valuable resource for understanding behavior in that they represent what users see and click on when navigating the site.

We describe two complementary ways of building communities of videos. The first tries to capture the general behavior of viewers. The second is more directed and focuses the community on a specific topic.

Breadth-first Search

Perhaps the simplest way of iteratively producing a community of videos is to begin with a specific seed video and to proceed with a breadth-first search of related videos. Breadth-first search consists of going from the seed video to its related videos, followed by their related videos, and so on, until all videos have been visited or a certain number of iterations has been reached, as detailed in Algorithm 1 [127].

Because there are 20 related videos provided by YouTube for each video, the size of communities discovered using this technique would grow very quickly, on the order of $O(20^d)$, where d is the depth or number of iterations performed. An alternative in such contexts is to

ALGORITHM 1: Breadth-first Search

Require: An initial video v_0, and for each video v, a set of related videos defined by $v.relatedVideos()$

Ensure: The set C contains the community

```
Initialize set C and queue Q to be empty
Q.enqueue(v_0)
repeat
    v ← Q.dequeue()
    C.add(v)
    for all related videos r in v.relatedVideos() do
        Q.enqueue(r)
    end for
until Q is empty or terminating condition reached
return C
```

constrain the breadth-first search to a beam search [259], where, for each video considered, only the first b most related videos, as per YouTube's rankings, are added to the queue. The size of the community now only grows on the order of $O(b^d)$. If $b = 20$, the result of the beam search is identical to the result of the traditional breadth-first search.

As an example, Figure 5.1 shows the community of videos discovered around an anti-smoking seed video using a beam search with beam size $b = 5$, run to a depth of $d = 3$ from the initial video. Table 5.1 shows a subset of the titles of these videos. For the sake of space, only the first four titles are included for depths 2 and 3.

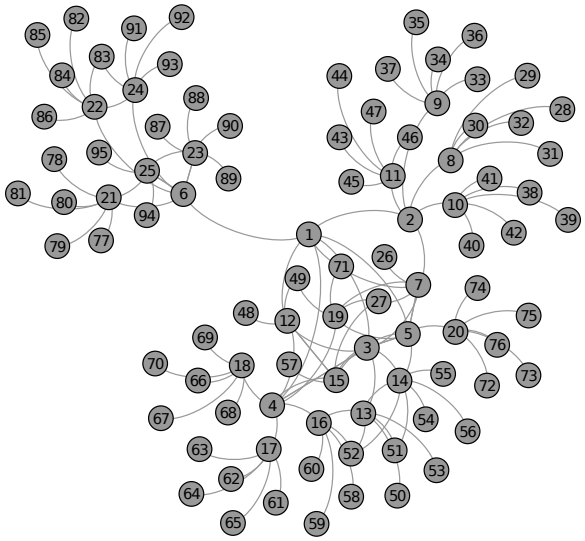

Figure 5.1: Beam Search-generated Community of Anti-smoking Videos ($b = 5$, $d = 3$)

Table 5.1: Subset of Titles of the Beam Search-generated Community of Anti-smoking Videos

Video	Depth	Title
1	0	Tobacco Free Florida: Kid Tossing Ball
2	1	Tobacco Free Florida: Kid Tossing Ball (CC)
3	1	Tobacco Free Florida: Mirror
4	1	Tobacco Free Florida: 31 Flavors
5	1	Tobacco Free Florida: Buckle Up (en Espanol)
6	1	The Sexiest Commercial Ever.
7	2	Tobacco Free Florida: Buckle Up
8	2	Grey Poupon Original Commercial
9	2	Bounty Paper Towel Ads with Captions
10	2	Gray Bright, Jack In The Box Taco Adventure (from Sydney Australia to Los Angeles USA for Taco's)
...	...	...
26	3	Tobacco Free Florida: Video Game (en Espanol)
27	3	Tobacco Free Florida: Light It Up
28	3	Wayne's World - Grey Poupon (Parody)
29	3	Grey Poupon "Son Of Rolls" 30 Sec Commercial
...	...	...

As may be expected, many of the videos in this community, even a small number of links away from the starting video, are about very different topics. Specifically, many of the videos at a distance of three and four steps from the first video are not about smoking behavior at all, but rather focus on humorous or sexual content (top left-hand side of Figure 5.1). A reasonable way to assess the likely subject matter of these videos is to observe keywords in the titles. After surveying the titles, we noted that a large subset of the videos could be identified as very likely to focus on smoking behavior or sexual appeal based on a few specific keywords. We recognize that there are clearly some tobacco and many sexual related videos that do not contain these specific keywords, but using them gives an objective way of summarizing the list of titles and illustrating the point. As shown in Table 5.2, this community of 363 unique videos has only 73 whose titles contain the smoking-related words "tobacco," "smoke," or "smoking," whereas 65 contain the sex-related words "hot," "sex," "ass," or "Megan Fox."

It is interesting to note that such findings would be difficult, if not impossible, to bring out without building communities. One valuable insight gained from this finding is that if a user is navigating through content using the related videos links, even if they start watching

Table 5.2: Statistics on the Titles of the Beam Search-generated Community of Anti-smoking Videos

Depth	Unique Videos	Smoking-related	Sex-related
0	1	1	0
1	5	4	1
2	19	9	5
3	70	18	17
4	268	41	42
Total	363	73	65

an anti-smoking video, it is very likely that they will end up viewing content with sexual or humorous appeal, rather than continuing to view multiple anti-smoking productions. From a public health standpoint, there are several follow-up questions one may consider:

1. Is the current observation representative of a more general human behavior? In other words, is it true that whatever the first video is (i.e., whatever the reason a user was drawn to a specific video on YouTube), users quickly (i.e., 2 or 3 hops) drift away to gravitate around videos with sexual content?
2. As far as conveying health-promoting messages is concerned, should content be packed into the first videos users are most likely to watch?
3. If viewers do indeed tend to be distracted by other content, is it possible to design health-related videos that are more likely to cause viewers to stick with the topic? How?

Answers to these questions, and other related ones, would help the preventive and intervention efforts of public health practitioners within social media.

Multiple Sub-community Expansion

As shown, a breadth-first search or even a beam search using the related videos provided by YouTube quickly diverges to many different topics. While this discloses possibly interesting aspects of human behavior, alternative methods must be employed to discover communities of videos that are more interrelated and therefore closer to the same topic.

Chen et al. recently introduced Iterative Local Expansion (ILE), a community discovery process designed for iterative expansion in large networks [42]. The first part of this process is a local community identification algorithm, which attempts to identify communities with a "sharp" boundary to the rest of the network. A community is considered in two parts: 1) nodes in the *core*, which only link to other nodes in the community; and 2) nodes on the *boundary* which link to other nodes in the community but also to those outside the community. The local modularity factor R is used to evaluate the quality of a community [50]. It is specified in terms of the boundary nodes and is defined as $R = \frac{B_{in}}{B_{total}}$, where B_{in} represents the number of links from the boundary nodes that stay inside the community and B_{total} is the total number of links from the boundary nodes. The local community identification algorithm begins with a single node and adds nodes to the community in a greedy fashion in order of most improvement to R, until R can no longer be increased.

ILE can, of course, be applied to videos on YouTube by using the set of related videos to define the nodes to which a particular video links. One of the limitations of this approach however is that, while a small set of videos (typically between 10 and 30) is discovered that are related around a certain topic, the topic may not be the exact one desired. For example, beginning with an anti-smoking commercial featuring a superhero, a community may be discovered that is focused on tobacco, or alternatively a superhero related community may be brought out. As an illustration, we have run ILE starting with ten different anti-smoking videos, and observed that in many instances the communities are, in fact, closely centered on tobacco, but in many instances the communities tend to focus closely on other topics, as summarized in Table 5.3. As above, tobacco-related videos are designated by titles containing the keywords "tobacco," "smoke," or "smoking."[1]

Chen et al. do suggest that their algorithm could be applied iteratively to eventually build communities covering the whole graph, by selecting random starting nodes from those

[1]We have found that even running ILE with the same starting anti-smoking commercial in successive weeks can result in rather different behaviors because the greedy algorithm is highly influenced by the selection of the first few nodes.

Table 5.3: Tobacco Relatedness of ILE-generated Communities of Videos

Community	Videos	Smoking-related	Percent
1	18	4	22.2
2	16	2	12.5
3	33	4	12.1
4	9	0	0.0
5	17	1	5.9
6	12	1	8.3
7	11	9	81.8
8	13	12	92.3
9	29	1	3.4
10	30	27	90.0
Total	188	61	32.4

not in a community [42]. These potential starting nodes consist of those linked to by a boundary node, but outside a community, and are referred to as the *shell* of the community. This random selection approach would result in assigning additional videos to communities, but as with a beam search, it would quickly diverge to more diverse topics. If we consider each of these communities as a sub-community of a larger set of videos related to the desired topic, this iterative process could be used to identify additional starting nodes and subsequently additional sub-communities. However, to discover additional sub-communities about the same overall topic, the selection process must be guided.

We propose an extension to ILE, called Multiple Sub-community Expansion (MSCE), that implements an alternative selection process to identify starting videos that are more closely related to the original topic, composed of two components. First, each node in the shell set S is given a community link score L of the number of unique sub-communities that link to the node. On the first iteration, this will result in a score of $L = 1$ for each node in S because they are each linked to by the single existing sub-community. On subsequent iterations, when more sub-communities have been identified, videos that are linked to by more than one sub-community will receive higher L scores.

The second component consists of a keyword score K. These keywords are related to the overall topic and are supplied by the user at the beginning of the process. The K score of a video is determined by the number of keywords contained in that video's title. Using these

two components, an overall expansion selection score E can be determined as the weighted sum of these, i.e., $E = L + \alpha K$, where α denotes a constant that can be defined to indicate the importance of keyword score. The node with the highest E score is then selected as the starting node for the next community. Details of MSCE are shown in Algorithm 2.

ALGORITHM 2: Multiple Sub-community Expansion

Require: An initial video v_0, a set of *keywords*, and a weight parameter α
Ensure: The set C contains a set of sub-communities

```
Initialize Set C to be empty
Let s = v_0
repeat
  Run ILE starting with s to produce sub-community C_i and shell set S
  C.add(C_i)
  for all videos v in S do
    Let L = 0 and K = 0
    for all sub-communities c in C do
      if v is connected to any nodes in c then
        L = L + 1
      end if
    end for
    for all keywords key in keywords set do
      if v.title contains key then
        K = K + 1
      end if
    end for
    Let E = L + αK
  end for
  s = arg max_v E
until S is empty or terminating conditions reached
return C
```

One of the benefits of MSCE is that even when the first sub-community is not as related to the central topic, subsequent sub-communities are likely to return back to the desired topic. For example, Figure 5.2 shows the composite community that is the set of ten sub-communities discovered by MSCE, using the single keyword "smoking," and beginning with the same single seed video used for our earlier beam search ("Tobacco Free Florida: Kid Tossing Ball"). In this case, the first sub-community (highlighted by the rectangular region on the top right part of the figure) contains some anti-smoking videos featuring

superheroes, which results in also including several videos that are solely about superheroes. Despite the fact that this sub-community is not completely focused on smoking, subsequent sub-communities are much more focused on the topic, as shown in Table 5.4. Note that because sub-communities can overlap, the total values are computed with regard to the total number of unique videos, not as sums of the corresponding columns. While the first sub-community consists of only 22.2% (4/18) videos containing the words "tobacco," "smoke," and "smoking," when considering all ten sub-communities, 84.4% (157/186) of unique videos contain these words.

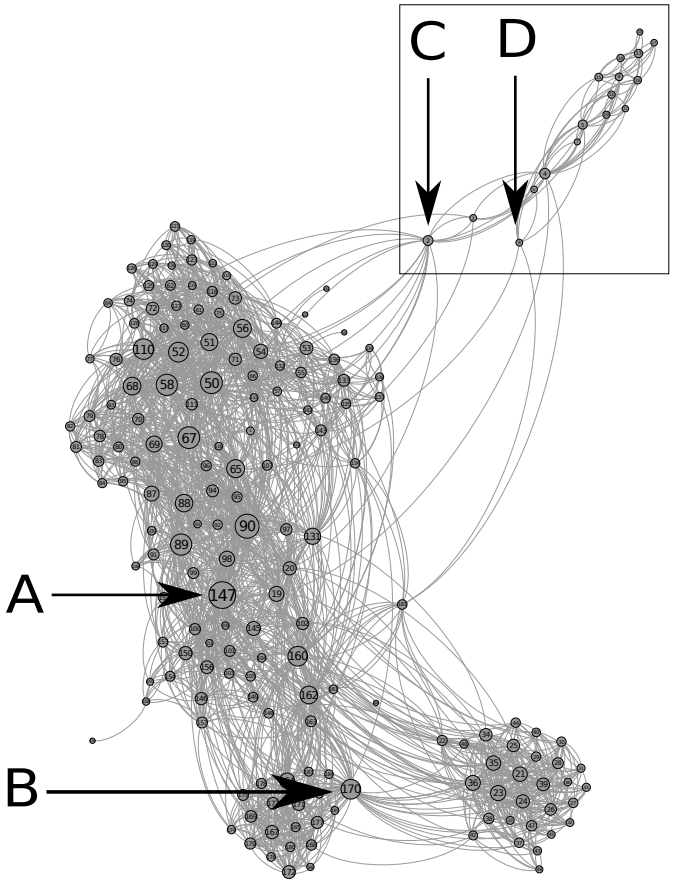

Figure 5.2: MSCE-generated Community of Anti-smoking Videos in 10 Sub-communities

To further demonstrate the robustness of the MSCE algorithm in finding subsequent sub-communities that return to the desired topic, we have run the algorithm for ten iterations (i.e., building a community composed of ten sub-communities) beginning with each of the

Table 5.4: Tobacco Relatedness of MSCE-generated Sub-communities for a Single Anti-smoking Video Community

Sub-community	Videos	Smoking-related	Percent
1	18	4	22.2
2	31	29	93.5
3	15	14	93.3
4	33	32	97.0
5	24	23	95.8
6	35	29	83.0
7	16	15	94.8
8	15	12	80.0
9	22	21	95.5
10	15	13	86.7
Total (unique)	186	157	84.4

anti-smoking commercials used above. Thus, where before we built a single whole community for each of these videos (as shown in Table 5.3), we now build a composite community (made up of ten sub-communities) for *each* of the ten anti-smoking videos. Table 5.5 shows statistics regarding these ten communities.

Table 5.5: Tobacco Relatedness of MSCE-generated Communities for Ten Different Anti-smoking Videos

Community	Videos	Smoking-related	Percent
1	186	157	84.4
2	163	144	88.3
3	165	87	52.7
4	145	100	69.0
5	161	79	49.1
6	145	85	58.6
7	111	100	90.1
8	139	99	71.2
9	161	104	64.6
10	149	104	69.8
Total	1525	1059	69.4

Even when the first sub-community (as shown in Table 5.3) was not on topic, the subsequent nine sub-communities included in the final composite community bring the overall community back on topic (as shown in Table 5.5). The percentage of smoking-related videos in the first sub-community compared to the percentage for the entire community, for each of the ten videos, are depicted in Figure 5.3. These results show that on average the percentage

of smoking-related videos increases significantly from 32.4% to 69.4%, when including the additional nine sub-communities. This suggests that MSCE can be successful at returning to the desired topic even when the initial community was further away.

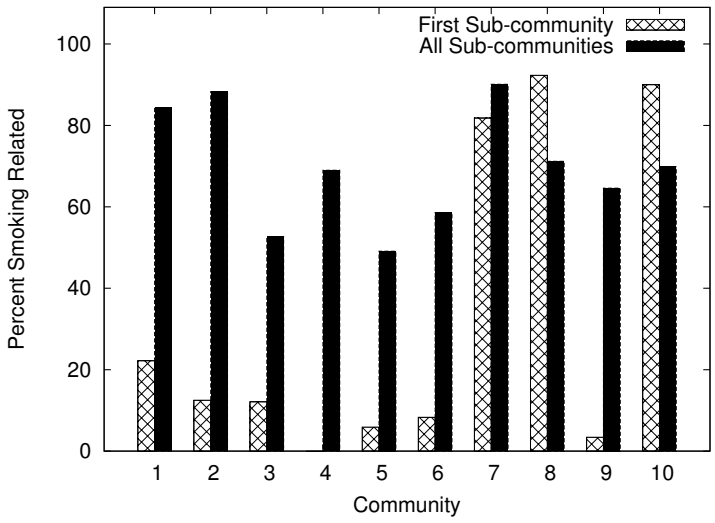

Figure 5.3: Percentage of Smoking-related Videos in the First Sub-community vs. in the Complete Composite Community for Ten MSCE Communities

Additionally, we observe that videos that are more archetypical of the topic and focus solely on the message rather than including other themes or personalities from popular culture are more likely to remain centered on the original topic. However, even in these cases the keyword component of MSCE is able to guide some of the subsequent sub-communities back toward the topic. This is demonstrated with regard to videos 4, 5, and 6, where, as shown in Table 5.3, the first sub-communities for each had as little as 0%, 5.9%, and 8.3% of the videos containing the smoking-related words in the title, but after iterating to ten sub-communities, the percentage of smoking-related videos in the corresponding final communities had risen to 69.0%, 49.1%, and 58.6%, respectively. Interestingly, the communities built from videos 8 and 10 experienced a reduction in the percentage of videos on the topic (from 92.3% to 71.2% and 90.0% to 69.8%, respectively), where in each case the expansion included sub-communities that focused more on humorous commercials rather than strictly tobacco centered ones.

From the point of view of public heath practitioners, discovering a community of related videos using MSCE may prove useful in at least a couple of important ways.

1. As discussed later, obtaining a community of videos is the first step in many other types of analysis, such as considering the relatedness of the authors or commenters of videos. In addition, insight can be gained by examining the community of videos directly. For example, nodes with a very high degree are likely to be very central to the topic and could represent those with higher social capital among the set. Nodes that are bridges between different sub-communities are interesting because they represent a clickstream that a user may follow to transition between topics. For example, a video that bridges a sub-community of anti-smoking commercials and unrelated humorous commercials could represent the point a which a user stops consuming the health related content. in Figure 5.2, node A has a very high degree which is the video "Graphic Australian Anti-Smoking Ad" that has been viewed 2.5 million times and is very central to the topic of anti-smoking commercials. Also, node B, entitled "How to quit smoking," has high degree and is the bridge between three sub-communities focusing specifically on "the effects of smoking," "do you still want to smoke," and "how to quit smoking." Nodes C ("Star Wars Anti Smoking Ad") and D ("Anti-Smoking : Superman Versus Nick O'Teen (1981)") are examples of bridge videos that are about tobacco, but could also represent a clickstream taking a user to more superhero or movie related videos than health ones.

2. MCSE could be used as an alternative sampling method for other studies. Almost all previous public health work involving YouTube has the researchers choose a set of keywords to search through YouTube's website and using the resultant videos as the sample for their work [17, 80, 100, 118]. While this approach has a higher chance of returning videos that are well on topic, it also presents a number of drawbacks.

In particular, finding adequate keywords is notoriously difficult,[2] and in the case of YouTube (as many other online search systems) the number of results returned per query is limited to 1,000. Alternatively, the MSCE approach can retrieve any number of videos. Another interesting aspect of building a sample based on MSCE is that it more closely matches actual user activity. Research has shown that users rarely look beyond the first few pages of results: 41% are reported as continuing their search by changing keywords when the desired content is not found on the first page of results and 88% as changing their keywords when they do not find it on the first three pages [103]. Thus, performing analysis on over 900 videos retrieved from a search does not match user behavior. Our own intuition and experience suggests that users often hop from one video to the next by way of the related video links.

5.3.2 User Communities

While YouTube is well-known for its video content, users are also at the heart of YouTube. Users are part of a larger community of friends, subscribers, subscriptions, videos, and authors. Models of these communities offer researchers ways to identify important authors and their characteristics, influential videos, and interesting users. YouTube also acts, in some fashion, as a social networking service, allowing users to identify other users as friends, subscribe to authors, personalize a page with user info and videos posted by the user, and exchange messages with other users.

Author-Friend Community

An author-friend community is an example of the communities that can be built based on the YouTube users. This community is built starting with a set of videos (such as those obtained using the community mining algorithm mentioned above) and identifying the author of each video in the set. Then each of the friends of these authors is identified, and a graph is

[2] Ambiguous words, mismatch between practitioners' vocabulary (e.g., smoking cessation) and layman's terms (e.g., quit smoking), etc.

built with each of these authors and their friends as nodes, and edges denoting the friendship relation between users. Anomalous users can be identified from this graph, such as those with an unusual number of friends or those who are friends with an unusual amount of other authors in the community.

As an illustration, Figure 5.4 shows the author-friend community built from the authors of the same MSCE anti-smoking community discussed earlier, showing only those users who are friends of at least four authors in the set. Nodes corresponding to authors are shown in black, while nodes corresponding to friends are shown in white. The size of each node is proportional to its degree.

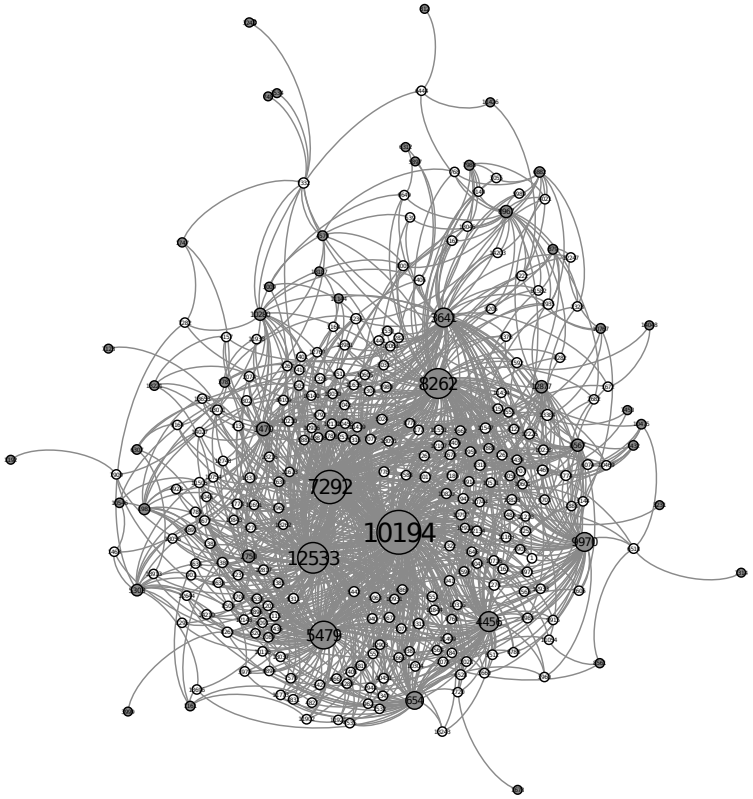

Figure 5.4: Community of Authors and their Friends

Figure 5.5 shows the distribution of number of authors per number of friends. Not surprisingly, the distribution follows a kind of power law with most authors having a small number of friends and few authors having a very large number of friends. The maximum

number of friends for an author in this community is 6,249. Note that we did not distinguish between authors with 0 friends and authors who choose to keep their list of friends private, which may bias our results.

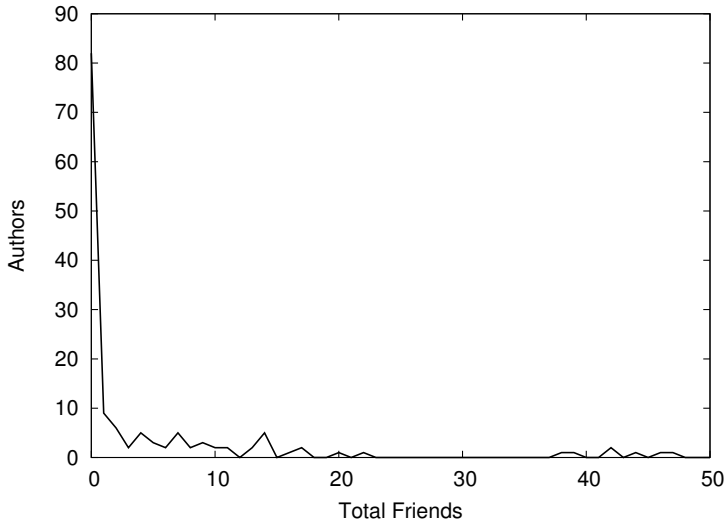

Figure 5.5: Number of Authors vs. Number of Friends

The users of this author-friend community represent those who are likely to have some affinity toward the topic, because they have either authored a video on the topic themselves or are friends with at least four authors of videos on the topic. Nodes with a high degree could represent users of higher social capital who potentially have influence in this community. One of the reasons that some of the users in this graph have a high degree is that they try to become friends with many others in an attempt to increase their own exposure as advertising means. For example, the three users with the highest degree (appearing in the center of Figure 5.4) are a health and beauty company, an online fitness company, and a documentary film maker. This may have interesting ramifications for governmental or non-profit public health producers, in that it may not be sufficient to simply produce content and upload it to YouTube. Authors likely need to become involved in the community so as to gain social capital, subsequently getting exposure to content. Such involvement can be built from the ground up, or it could take advantage of the novel understanding of the target community provided by the foregoing community mining approach. Indeed, rather than waiting to

acquire the needed social capital, authors of public health videos may benefit from tying into established users with high social capital, getting them to upload and/or promote their content.

Additionally, users who are friends but not authors in such communities and who have high degree, may be highlighted as the prime consumers of the community's video content. These consumers, in turn, could be observed in terms of their susceptibility to or targeted with specific messages.

Commenter Community

Another rich source of metadata in YouTube is found in the comments made by viewers on the videos. A commenter community can be discovered by, for example, identifying those users that leave comments on the same videos. Specifically, this commenter community is built by beginning with a set of videos and identifying all users that have made comments on each one.[3] Then, a link is made between users that commented on the same videos, where the strength of the link (or the weight of the edge) between two users is the number of videos in the sample on which both users commented. Additionally, thresholds can be used to indicate a link only if the users have commented on at least some number of common videos. This graph can also be restricted by considering users whose comments occur within a certain distance of each other in the list of comments.

Due to the number of comments per video, the commenter community can quickly become difficult to visualize if the set of videos is large and the threshold parameters are set low. Figure 5.6 shows the community of commenters for the anti-smoking videos in Figure 5.2, restricted to users who commented on at least four common videos. The thickness of an edge is proportional to the number of common videos on which the adjacent users commented.

[3]In the current API, YouTube returns a maximum of 1,000 comments per video. Even with this limitation valuable insights can be found, but this limitation should be considered when attempting to generalize from this data.

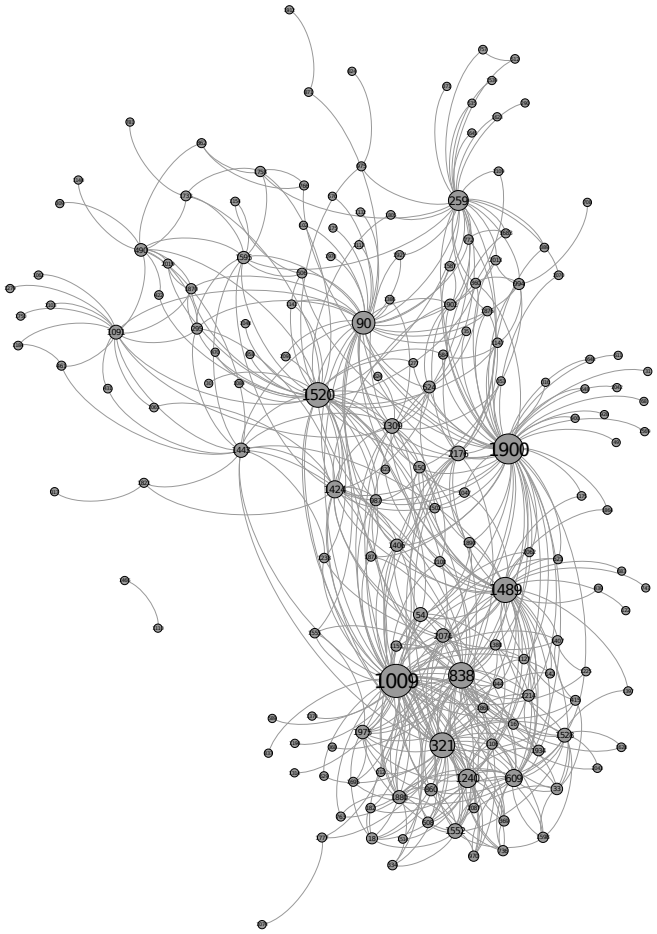

Figure 5.6: Community of Users Who Commented on at Least Four Common Videos in the Set of Anti-smoking Videos

Figure 5.7 gives the distribution of the number of comments made by users in this set, as well as the number of unique videos on which these users commented. Again, unsurprisingly so, the distribution follows a power law, with most commenters leaving only very few comments behind.

It should be noted that commenter communities do not imply that users feel the same way about an issue, but rather that they are both interested in the issue, and may in reality have opposite views on the topic. The two nodes from Figure 5.6 with the highest degree are both users promoting their own stop smoking programs, leaving almost identical comments on many videos in the set, encouraging others to follow a profile link. Because these users

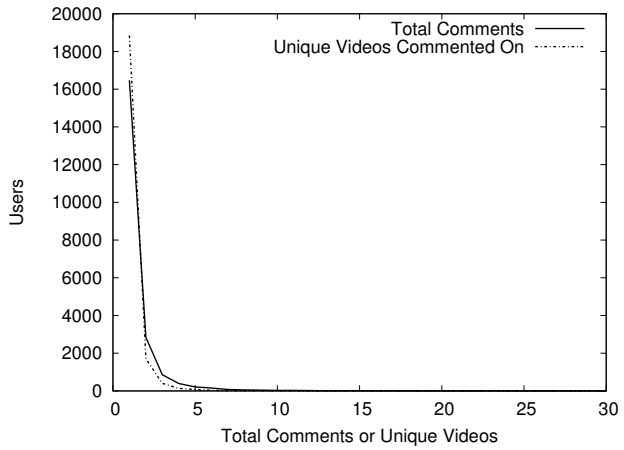

Figure 5.7: Distribution of the Number of Unique Videos Commented on by Users

left comments on so many videos in the set, they have an implicit relationship with a large amount of other commenters.

Another, more explicit commenter community can also be built by considering directed edges between users that direct comments at one another using the conventional "@*username*" syntax. Figure 5.8 shows the resulting community of commenters over the same set of videos as above. Links between users appear only when at least two directed comments have been made. The thickness of the links is proportional to the number of times users referenced each other.

The central node in this figure with a disproportionately high degree is a spammer similar to those found in Figure 5.6. However, in this case, the user left multiple directed comments to others promoting a political and ideological agenda, which in many cases elicited antagonistic responses. Alternatively, the pairs of users with disproportionately high weight on the edges between them represent users that maintained long-lasting conversations with one another.

Additionally, because the explicit commenter network is directed, users can be identified that have a high in-degree, representing those at whom many others direct comments. These users may have higher social capital in that they have attracted the attention of many others. In the case of this network of anti-smoking videos, the user with the highest in-degree made a

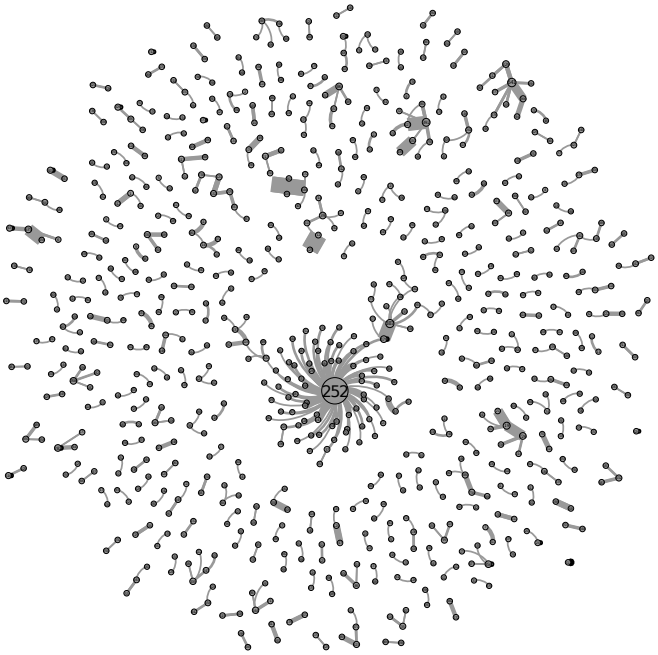

Figure 5.8: Community of Users Defined by the "@*username*" Syntax

single comment asking the question: "if smoking is so bad why isn't it illegal?" This elicited the responses of over 30 other users.

Comment Trails

Finally, it is also possible to utilize video comments to track the comments of a specific user through time to observe a type of path followed by the user. Because the current version of the YouTube API does not provide all the comments of a user, this information must be acquired by first identifying a set of videos and then considering all comments left on those videos. The set of comments can then be sorted by user and comment time, to show the trail of users through the set of videos.[4]

This type of user trail can be valuable in two ways. First it can help to further identify characteristics of a single user of interest. Second, and perhaps more importantly, it can help to show trends of what videos users are seeing and how they move through these. An obvious

[4]This analysis is also subject to the limitation of only being able to consider the first 1,000 comments on a video.

limitation in identifying the trails of users is that not all users leave comments, and those that do do not leave them on each video they watch. Despite this limitation, these trails may still be valuable in discovering overall trends and relationships among videos. Also, we submit that those that do leave comments are in many cases the ones with the most extreme views on either side of an issue. Depending on the topic being studied, this may actually be more valuable as a way of identifying those that are more interested or passionate about the issue.

Using the same set of comments (for the community of anti-smoking videos) discussed above, we identify users' trails. In the sample of 186 videos, the maximum number of unique videos commented on by a single commenter is 27. Table 5.6 shows the trail of a prototypical commenter through our community of anti-smoking videos as defined by the date/time the comments were authored. The author engaged in a conversation with other users on videos B and D resulting in returning to leave additional comments on that video. The fact that a user returns to the same video to continue a conversation may result in additional exposure to its content. Thus, there may be a correlation between high-impact videos and increased conversation in their respective comments, either because the video itself drew increased discussion, or because the increased conversation led to more exposure of the message.

Table 5.6: A User Comment Trail Sorted by Time

Date/Time	Video
06/02/10 10:39 AM	A
06/19/10 06:47 AM	B
06/19/10 06:54 AM	B
06/19/10 07:03 AM	C
06/19/10 07:07 AM	C
06/19/10 07:19 AM	C
06/19/10 07:28 AM	D
07/17/10 05:48 PM	E
07/17/10 05:49 PM	E
08/03/10 08:32 AM	B
09/08/10 12:32 PM	D

5.3.3 Comment Communities

In addition to considering the users that made comments on videos, the text of the comments themselves can be valuable in discovering the views of content consumers. We turn to Latent Dirichlet Allocation (LDA) to exploit this text data and gauge the feeling, perception, and reaction of the public to the messages that are presented. LDA is a probabilistic model that, when applied to documents, hypothesizes that each document in a collection has been generated as a mixture of unobserved (latent) topics, where a topic is defined as a categorical distribution over words [26]. While not strictly the case, we can usefully regard the set of topics as a community of comments over the video community.

As an illustration, we consider the text from video titles, descriptions, and comments in the set of $4,407$ videos gathered by breadth-first search ($d = 3$), starting from the video titled "Quit Smoking." We assemble the title, description, and comment data as a corpus of documents consisting of one document for each video containing both its title and its description (if any), plus one document for each video comment. We use the popular MALLET implementation of LDA [155] to automatically discover topics in this corpus. The number of topics K is set to 10 to give a high-level sense of the themes dealt with in the corpus, and to simplify analysis. The topics discovered, represented by their most prominent words, are given in Table 5.7.

A clear "quit smoking" topic emerges—Topic 6. However, its weight is relatively small (0.08751) indicating that the discussion has diverged substantially from the topic of the starting video. A near-universal of topic modeling on YouTube comments is the presence of an expletives topic. In this case, Topic 7 combines expletives with other colloquial forms such as "lol", "ur", and "wtf".

5.4 Conclusions and Future Work

We have illustrated ways in which community mining techniques may be applied to YouTube to inform public health practice. We recognize that we have only scratched the surface

Table 5.7: Ten Topics Inferred on the "Quit Smoking" Videos and Comments

#	Weight	Top Words
0	0.08614	scary videos ur die life dont read fake post press video ghost lol works comment love
1	0.13785	video lol youtube watch thumbs videos amir check xd love channel remember justin
2	0.17116	don people game real time lol make good car video fake guy batman man dont thing
3	0.02474	de la el es en se si lo por una los video mi le con xd di che tu
4	0.25734	people video love baby good life don time god sad feel im man make wow girl kid dont
5	0.10693	people god don world jesus life make religion human time truth things country good
6	0.08751	smoking smoke people cancer don weed quit dont good cigarettes years stop day bad
7	0.45418	lol funny f⋆ s⋆ f⋆ xd people guy a⋆ haha stupid im dont ur man dude gay
8	0.01	allah bu ha bir ve ne fap mart de wal ya da bean ama mr bi sen ben var
9	0.10818	movie love song great good film watch movies trailer awesome amazing watched music

and that, while tobacco usage provides an intuitive case study, we have not here produced any significantly new knowledge in this area. However, we have showed the potential and highlighted a number of relevant follow up questions that the approach presented here brings to light naturally, and can help answer in more thorough and focused analyses.

Chapter 6

Discovering Social Circles in Directed Graphs

Abstract

We examine the problem of identifying social circles, or sets of cohesive and mutually-aware nodes surrounding an initial query set, in directed graphs where the complete graph is not known beforehand. This problem differs from local community mining, in that the query set defines the circle of interest. We explicitly handle edge direction, as in many cases relationships are not symmetric, and focus on the local context because many real-world graphs cannot be feasibly known. We outline several issues that are unique to this context, introduce a quality function to measure the value of including a particular node in an emerging social circle, and describe a greedy social circle discovery algorithm. We demonstrate the effectiveness of this approach on artificial benchmarks, large networks with ground-truth community labels, and several real-world case studies.

6.1 Introduction

Humans are inherently social beings that tend to associate with one another through homophily [158]. It follows that human society, both online and offline, is characterized by a complex network of interconnections within which somewhat homogeneous groups, or communities, emerge naturally. In turn, these communities tend to have a powerful influence on the attitudes and behaviors of their members [58], so that an individual's social environment can often be leveraged to infer important information about that individual's attitudes, behaviors and decisions [6, 85, 163, 208, 225, 242, 256]. For example, a health professional may be

able to improve the efficacy of his/her intervention by considering the social circle of an at-risk individual, or a workshop organizer may better target potential participants by issuing invitations around a core group of known experts.

The problem of discovering such communities around one or more individuals has recently been referred to as the *community search* problem [221], to differentiate it from the well-known *community detection* problem (e.g., see [51, 77, 177, 181, 206]). Unlike community detection, which is concerned with finding arbitrary highly-interconnected subgraphs within larger networks, the goal of community search is to identify a single subgraph that includes an initial set of query individuals. The role of the query set is to provide some context to the search. Indeed, most individuals belong to multiple overlapping communities, such as work organizations, clubs, and neighborhood associations. While the issue of overlapping communities has received some attention in the context of community detection [85, 183, 222, 240, 253], it is clearly intrinsic to the local community search problem where a single node cannot uniquely identify the community of interest. Instead, by adding other nodes to the query set it is possible to extract different overlapping communities for the same individual depending on the content of the query set. Overlapping communities are thus handled naturally, because a node's community membership is established for each query set separately. The specification of the additional seed nodes is what determines the desired community, and ideally, these nodes are selected such that their only element of commonality is the characteristic that defines the desired community, e.g., co-workers, teammates, fellow hobbyists. For example, the social circle of an individual, that begins with two of his/her sisters, is likely to center around family relationships, while the social circle of that same individual, that begins with two of his/her professional colleagues, is likely to include mostly business relationships. In that sense, such local communities resemble what sociologists refer to as social circles [114, 115], and we refer to them as such in the following.

Here, we focus our attention on the *local* social circle discovery problem. Analogous to the difference between community detection and local community detection [42, 50, 139, 148,

190], the local variant of the social circle discovery problem operates under the constraint that the entire graph is not known *a priori*, and that new edges and nodes are discovered only through their adjacencies to the currently-known portion of the network. The local constraint is intrinsic to many contexts wherein knowing the entire graph is either impossible or infeasible (e.g., Web pages, Twitter users, YouTube videos).

We further focus on *directed* graphs, since many relations are naturally directed and opposite-directional links are not synonymous (e.g., publication citations, links on Web pages, followers on Twitter). Most extant community mining algorithms are designed for undirected graphs, with the assumption that they can be applied to directed graphs simply by ignoring direction and treating the graph as if it were undirected. However, if edge direction is ignored, valuable information is lost [140]. Furthermore, incoming links to a node may not be known without an exhaustive search of the graph rendering this approach clearly inadequate.

In this paper, we propose an effective local social circle discovery algorithm for directed graphs. Ideally, seed nodes are selected such that the only element of commonality among them is the underlying characteristic, or shared interest, that defines the desired social circle. We adopt a greedy expansion approach where nodes adjacent to the social circle are iteratively added, or those in the social circle are periodically removed, by maximizing a particular heuristic function, until a specified size is reached. We demonstrate the effectiveness of the proposed algorithm using standard benchmarks as well as case studies in large real-world social networks.

6.2 Related Work

While there is no consensus on the exact definition of the term community [181], a community is usually defined as some variant of a subgraph of nodes that are more densely related to each other than to the rest of the graph. Most of the research regarding communities has focused on detection, where a graph is partitioned into distinct communities, based on random walks [123, 199, 209], label propagation [89, 204], spectral methods [37, 219],

modularity [25, 51, 78, 177, 206], and generative models of affiliation [255]. A recent, and excellent, survey of the field is in [77].

In the past decade, several researchers have begun to consider a natural variation on the community detection problem, that rather than partitioning a graph into a number of communities, builds a single community, or social circle, from one or a small number of nodes. For example, Palla et al. propose finding k-cliques around a start node [187], while Mislove et al. build a community using normalized conductance, an idea derived from circuit analysis [163]. Sozio and Gionis, who coined the phrase *community search problem*, offer a solution that starts with a set of nodes and expands a community from them using a variant of density based on the minimum node degree rather than the average node degree [221]. Unfortunately, in all of these cases, knowledge of edges outside the community and its boundary, or even the complete graph, is necessary, which is often impossible or infeasible.

Hence, other researchers have focused on local methods. Clauset, for example, introduces a local extension of modularity, based on the steepness of the boundary [50], while others have proposed related approaches based on such criteria as internal and external links [148], bridges to other communities [190], triangles to outside nodes [81], and the rate of adding new links [19]. While they do not require knowledge of the graph, these approaches strongly depend on the notion of a boundary, which, as we show here, may exclude relevant nodes. Local methods that focus less on the boundary, and more on density-related measures, have also been developed. They include Iterative Scan, which alternates through phases of adding new nodes and removing community members to maximize a density metric [21], Greedy Clique Expansion, which builds upon earlier work from [135] and adds/removes nodes in a greedy fashion to maximize a ratio of internal to total edges [139], Max-flow [73], internal density maximization [182], and spectral clustering [11, 254]. Interestingly, all of these local community mining approaches assume undirected edges, with the stated (and sometimes only implicit) assumption that the algorithm can be applied to directed graphs by ignoring edge direction. However, because edge direction may limit the knowledge of links *into* an

emerging community, applying undirected local approaches to directed graphs is not trivial and may embed assumptions that adversely affect the algorithm or metric. By contrast, we propose a local social circle discovery algorithm for directed graphs.

Recently, McAuley and Leskovec have presented a generative, unsupervised approach to discover an individual's social circles among their friends, which combines link and profile information [152], while Qin et al. do something similar as they cluster blogs around a given vertex of the blogosphere [202]. To the best of our knowledge, these authors are also the first, within the Computer Science research community, to use the term *social circle*. Their definition is similar to ours since they "expect that circles are formed by densely-connected sets of alters...[and] each circle is not only densely connected but its members also share common properties or traits" [152], but their motivation is different. They focus exclusively on ego networks, and essentially cluster ego's alters, building a number of circles around ego. By contrast, we take two (or more) individuals (think of ego and a small set of its alters only) and build a single social circle around them. One significant distinction is that we may get in our circle someone who is not directly connected to ego (i.e., not one of the current alters) but who is strongly connected with others in the social circle. One may think of this as a case where ego may not have yet established an explicit connection to that individual but probably should. A simple example would be a situation where an individual, say John, is connected to a number of people in his family but has no direct link to aunt Sally, whereas most others in his social circle do. McAuley and Leskovec's algorithm would not be able to put aunt Sally in any of John's circles since she is not one of his alters. Our algorithm, on the other hand, would add aunt Sally to John's family circle, on the strength of her associations with John's other family alters. Hence, while their work focuses on *organizing* the neighbors of a node into different groups, we seek to *discover* nodes that belong with the initial query set, including those that are not directly adjacent. Hence, we extend the concept of social circles to include nodes within the same community, not just those connected to a particular ego.

Finally, we note two problems that bear similarity to community search, or social circle discovery, but also differ in significant ways. First, the team formation problem, whose goal is to identify a compatible team of experts possessing required skills, may involve an initial set of query nodes, yet the problem itself is quite different in that the defining requirements are the skills and personalities of the potential members, not their connections [138]. Second, the graph theory problem of finding a minimum set of nodes connecting an initial query set shares some similarities with the local community search problem, but is also very different in that local community search seeks to build a cohesive set of nodes around the query set, as opposed to simply finding paths among them, and it is also very likely that the initial query nodes are already connected [71, 234].

6.3 Social Circle Discovery

The local social circle discovery problem for directed graphs consists of identifying a set of cohesive and mutually-aware nodes surrounding an initial query set, using only information from known nodes and the directed edges among them. This definition raises a number of important issues that must be addressed in order to formulate a node selection function that captures the underlying intuition of what "good" social circles should be like. We examine these issues in turn, and show how they affect the design of our node selection function.

Before we proceed, however, we first consider one of the fundamental tenets of our work, namely that we work explicitly with directed graphs. Many relations are naturally directed yet not inherently reciprocal (e.g., publication citations, links on Web pages, followers on Twitter), resulting in graphs of directed edges. Interestingly, most existing community mining algorithms are designed for undirected graphs with an assumption that they can be applied as is to directed graphs by simply ignoring direction. Leicht and Newman note that ignoring edge direction works reasonably well in some cases, but not in others, and in all cases it discards potentially valuable information that could enable more accurate community

discovery [140]. Consider the four graphs shown in Figure 6.1, which are all isomorphic to graph 1 if edge direction is simply ignored.

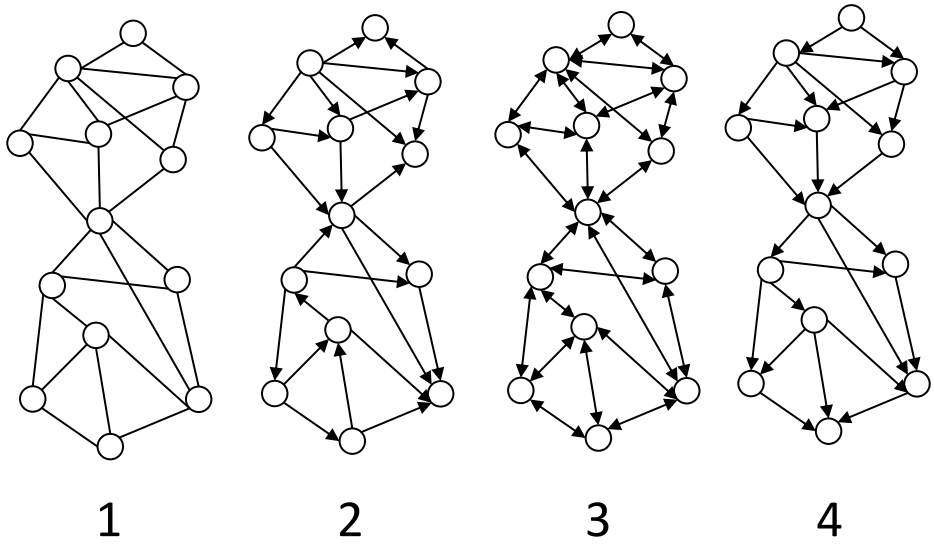

Figure 6.1: Differences among graphs when edge direction is taken into account.

When edge direction is taken into account, obvious differences emerge among these graphs. For example, since Graph 3 is composed exclusively of bidirectional links, it has twice as many edges as Graph 2, which is composed of only unidirectional links. Even more relevant to social circle discovery, since all of the edges in Graph 4 point downward, the nodes at the bottom of the graph may be completely unaware of those above them that link to them. Applying undirected algorithms to directed graphs requires two important assumptions to be made. First, an assumption must be made about how to count edges. For example, if a metric requires counting the number of edges between two nodes, should a bidirectional edge count as 1 or 2? Treating directed graphs as undirected, implicitly causes bidirectional edges to be counted as 1, like any other edges in the graph. Second, an assumption is needed about whether both incoming and outgoing edges should be considered. While the natural answer may be that all edges should be used, in many instances incoming links cannot be directly discovered (e.g., links to a website) [162]. This is exactly the situation in the local discovery context, where nodes can only be found through their links from the known portion

of the graph. Because some inward links may be known through exploration of other nodes, and yet, there may exist any number of additional unknown inward links, using any of these links in calculations could lead to unexpected behavior. Furthermore, even if all edges were known, there seems to be a significant semantic difference in terms of social circle membership between a node with high in-degree (e.g., a news site that many readers link to) and a node with high out-degree (e.g., a directory-like service providing pointers to a large number of resources). Yet, treating edges as undirected would view both cases as identical.

For all of these reasons, we contend that it is important to design algorithms that handle directedness explicitly. We now return to the specific issues raised by the local discovery of social circles in that context.

6.3.1 The Lab Advisor Problem

Since a social circle is defined as a cohesive group of nodes around an initial query set, one would expect that the decision to include a new node in a given social circle should be independent from the existence of other collateral social circles to which that node may also belong [81].

As an example, consider the task of discovering the social circle around a few students who work in the same research lab. One would expect that social circle to encompass all students in the lab, as well as the lab advisor. Now, for the most part, the students are likely to have limited professional contacts outside the lab. The advisor, on the other hand, is likely to be well connected within the broader research community to many individuals outside the lab. This scenario is depicted in Figure 6.2, where there is a link between two nodes if the corresponding individuals have a professional relationship, Node A is the advisor, and the shaded nodes represent the students that make up the current lab social circle.

While it is true that A is part of a select group of people with whom she interacts in her research community, it is equally true that A is part of her research lab. Provided that the focus is originally on a few of A's students (query set), the lab should here be the discovered

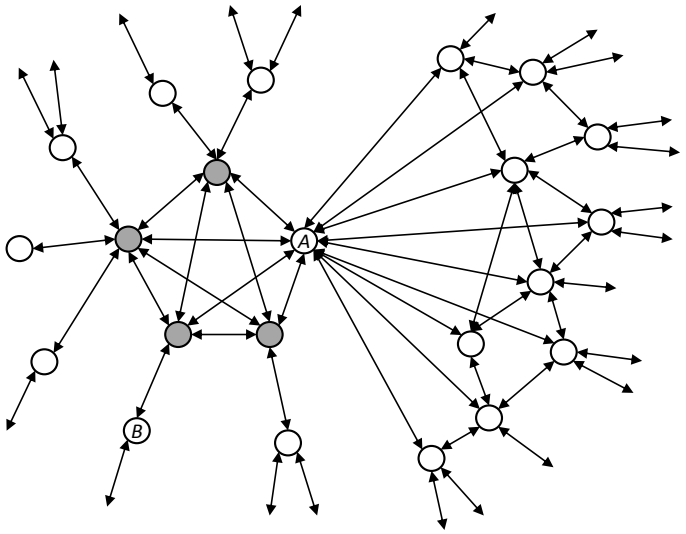

Figure 6.2: The problem of selecting an advisor (A) as a member of her research lab (shaded nodes).

social circle. Note that many community mining algorithms, that view a community simply as a subgraph of nodes that are more densely interrelated among themselves than with the rest of the graph, put emphasis on the community's boundary to outside nodes, and would thus miss Node A and instead prefer an outside colleague of a single lab member, such as Node B, because of its fewer total number of connections.

In the case of naturally overlapping social circles, the fundamental assumption of "more in than out" is just not valid [4]. A node's membership to a specific social circle should depend solely upon the strength of its ties to that circle, and not on the presence of links (or lack thereof) to others outside of it. Hence, as [182], we turn our attention to the idea of maximizing internal density. Given a directed graph with N nodes and E edges, density is defined as [241]:

$$Density_d = \frac{E}{(N)(N-1)} \tag{6.1}$$

When selecting the next node to add to an existing social circle, the denominator is the same for all candidates, since in all cases the social circle's size increases by 1, regardless of the number of links of the candidate node to the social circle. Hence, to maximize $Density_d$,

one only needs to maximize its numerator, E. Now, E counts the number of edges in the social circle so that for all candidate nodes, E starts at the same value, and the differentiating factor among candidate nodes is the number of links that exist between these nodes and the social circle.

Let $e(x, y)$ be an edge indicator function defined by $e(x, y) = 1$ if there is an edge from x to y, and $e(x, y) = 0$ otherwise. Let SC be a social circle and n a node that may be added to SC. Then, the number of links between n and SC, denoted by $d_d(n, SC)$, is the sum of the number of links from nodes in SC to n and the links from n to nodes in SC, namely:

$$\begin{aligned} d_d(n, SC) &= \sum_{c \in SC} [e(c, n) + e(n, c)] \\ &= \sum_{c \in SC} e(c, n) + \sum_{c \in SC} e(n, c) \\ &= InDeg(n, SC) + OutDeg(n, SC) \end{aligned} \tag{6.2}$$

where $InDeg(n, SC)$ is the in-degree of n with respect to SC and $OutDeg(n, SC)$ is the out-degree of n with respect to SC. It follows that maximizing $Density_d$ (Equation 6.1) is the same as maximizing $d_d(n, SC)$ (Equation 6.2) across candidate nodes.

It is clear that, starting with the shaded nodes of Figure 6.2, maximizing $d_d(n, SC)$ would allow A to be added to the lab social circle. Similarly, as expected, if the set of query nodes were to include a few of A's colleagues from her broader research community, rather than a few of her students, the resulting social circle would include A and her colleagues, but none of her students. Hence, maximizing $d_d(n, SC)$ provides a principled solution to the Advisor Problem based on maximizing internal density in the context of directed graphs. Further refinements are needed, however, in response to other important issues.

6.3.2 The Fringe Problem

A social circle is defined as a cohesive group of nodes that surround a set of query nodes. Recall that the role of the query set is to provide context, such that, ideally, its nodes capture the characteristic that defines the desired social circle. As a result, one would expect the query set to remain somewhat prominent in, or central to, the discovered social circle, and not to be pushed out to the fringe by dense but remote groups of nodes.

One such scenario is depicted in Figure 6.3. Assume that the query set consists of the shaded nodes (area labeled 1). As some of the nodes in the area labeled 2 begin to be added to the growing social circle, they will have a tendency to cause the highly-connected nodes in the area labeled 3 to be added, thus leaving the initial query set on the fringe of the social circle.

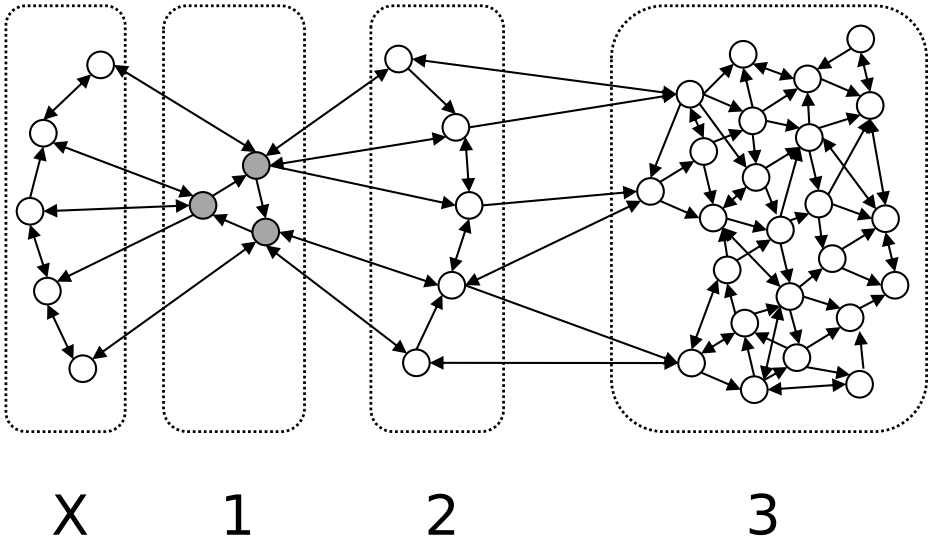

Figure 6.3: The problem of adding nodes away from rather than around the query set, leaving it on the fringe of the final social circle.

Given the position of the query set in the graph, it would seem natural to expect that the nodes in the area labeled X be part of the final social circle, rather than the nodes in the dense area marked 3. Interestingly, Sozio and Gionis noticed the same problem in the context of community search [221]. In their case, in undirected graphs and having a complete

knowledge of the graph, the solution was to use the minimum degree rather than the average degree of the nodes of a community as a measure of density for that community.

We are, of course, operating at the local level only, where nodes are added one at a time, based only on information from nodes in the growing social circle. If newer members of the social circle are treated the same as older ones, then they exert the same influence on new ones and can thus easily cause the social circle to divert from the initial query set, as illustrated above. Hence, our solution is based on the idea of discounted importance through length of membership to the social circle, as follows.

To maintain the relative importance of early members of the social circle, especially the query set, edges between nodes are discounted according to the time when the nodes were included in the social circle. Let $s(n)$ denote the step in the social circle-building process at which node n was added to the social circle. We modify Equation 6.2 to obtain the step-discounted value $\delta_d(n, SC)$ of $d_d(n, SC)$ as:

$$\begin{aligned}\delta_d(n, C) &= \sum_{c \in SC} e(c, n) s(c)^{-\alpha} + \sum_{c \in SC} e(n, c) s(c)^{-\alpha} \\ &= WgtInDeg(n, SC) + WgtOutDeg(n, SC) \end{aligned} \tag{6.3}$$

where $WgtInDeg(n, SC)$ is the weighted in-degree of n with respect to SC and $WgtOutDeg(n, SC)$ is the weighted out-degree of n with respect to SC. The parameter α is the discount factor. A value of $\alpha = 0$ treats all nodes equally, while a value of $\alpha = 1$ treats each node inversely according to the step in which it was added.

Maximizing $\delta_d(n, SC)$ allows us to avoid the Fringe Problem while retaining the advantages of maximizing $d_d(n, SC)$. Yet, one more problem remains, which we alluded to above when introducing directed graphs, and the distinction between in-degree and out-degree and its impact on social circle membership.

6.3.3 The Famous Person Problem

We have already addressed the issues of cohesiveness and overlap, and of query set centrality. There remains as part of the definition of a social circle the fact that it should be composed of nodes that are mutually aware, in line with Shaw's view that a group is "two or more persons who are interacting with one another in such a manner that each person influences and is influenced by each other person" [215]. While we do not require a social circle to be a k-clique, it is reasonable to expect that each member of the social circle influences and is influenced by at least some other members of the circle. It is clearly not sufficient for a potential node to have links from every member of the social circle if there are no links back, and vice versa.

As an example, consider two cases. In the first, the graph is made up of research scientists and there is a link from one research to another if the former has cited the work of the latter. There likely exist in such a graph dense groups, or social circles, of respected research scientists who have cited each other's work extensively. For a new researcher to cite the work of these scientists (i.e., link to them) does not make her part of their social circle in any meaningful way. In the second, and somewhat reciprocal, case, the graph is made up of individuals with varying levels of popularity and there is a link from one individual to another if the former is interested in the latter's activities and life events. While such a graph will contain a number of what may be viewed as genuine friendship networks, it will also contain celebrities whose social status makes them more visible to the graph at large. Then, one may likely find a celebrity who garners the interest of (i.e., is linked from) members of the same social circle. Surely again, this does not make the celebrity a part of the social circle in any meaningful way (it is unlikely that any celebrity is keenly interested in the life of any of her fans). Both of these scenarios are captured abstractly in Figure 6.4 where the shaded nodes mark the current social circle, Node A represents the new researcher in the first instance and Node B represents the celebrity in the other instance.

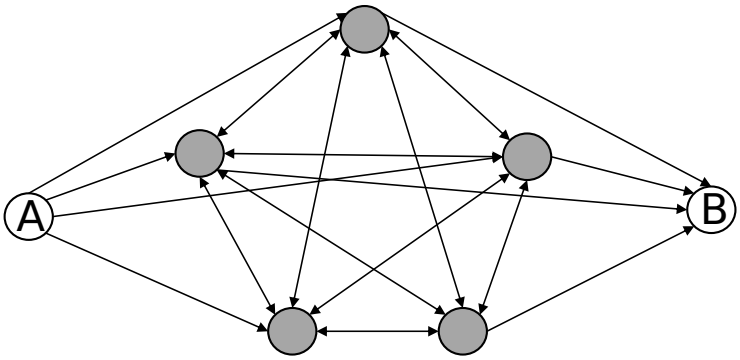

Figure 6.4: Two types of nodes that should not be included in the community because they do not have mutual influence.

Note that what makes Node A and Node B unusual is that they possess only one type of directed edges. Node A links to several nodes in the social circle, but there are no links from members of the circle back to it. Conversely, Node B has links from several members of the social circle, but does not link back to any of them. Neither one of these nodes should be part of the social circle. To help exclude such nodes and enforce some level of mutual influence as per the definition of social circles, we make one final change to the node selection function, wherein we modify Equation 6.3, so that rather than summing over the step-discounted in-degrees and out-degrees, we select their minimum, as follows:

$$\phi(n, SC) = \min\left(WgtInDeg(n, SC), WgtOutDeg(n, SC)\right) \tag{6.4}$$

By maximizing $\phi(n, SC)$ over all candidate nodes, we ensure that the social circle is dense, centered around the initial query set, and its members have a significant level of mutual awareness. Furthermore, the use of min in $\phi(n, SC)$ naturally handles the problem caused by nodes that link to the social circle, but of which the algorithm is currently unaware (due to its local nature). In this case, the min function will result in a 0, because such nodes have no links from the social circle, thus they can be consistently excluded. Only those nodes that have links from the social circle can have a score greater than 0 (the min term would

result in 0 otherwise), so the set of candidate nodes can be safely reduced to only those that are known.

6.3.4 Social Circle Discovery Algorithm

Sozio and Gionis have proven that a greedy algorithm is guaranteed to solve the community search problem for any node-monotone function to be optimized, where node-monotonicity is defined by [221]:

Definition 1. *Let V be an underlying set of nodes, and let G_V be the collection of all possible graphs defined over subsets of V. Let f be a function that assigns a score value to any graph in G_V and node $n \in G_V$, that is, $f : V \times G_V \rightarrow \mathbb{R}$. A function f is monotone non-increasing if for every graph G, for every induced subgraph H of G, and every node v in H, $f(H, v) \leq f(G, v)$. Node-monotone non-decreasing functions are defined similarly.*

Theorem 1. *$\phi(.)$ is node-monotonic non-increasing.*

Proof. Let G be a graph and H be any induced subgraph of G. Let n be a node in H. Then, it is clear that

$$\begin{aligned} WgtInDeg(n, H) &= \sum_{c \in H} e(c, n) s(c)^{-\alpha} \\ &\leq \sum_{c \in G} e(c, n) s(c)^{-\alpha} \\ &= WgtInDeg(n, G) \end{aligned}$$

Similarly,

$$\begin{aligned} WgtOutDeg(n, H) &= \sum_{c \in H} e(n, c) s(c)^{-\alpha} \\ &\leq \sum_{c \in G} e(n, c) s(c)^{-\alpha} \\ &= WgtOutDeg(n, G) \end{aligned}$$

It follows immediately that $\phi(n, H) \leq \phi(n, G)$, which establishes the result. $\square$

Other than the function to optimize, which captures the specific group properties one is interested in, the formal definition of the community search problem and that of the social circle discovery problem are identical. Hence, it follows from Theorem 1 that Sozio and Gionis' greedy algorithm, equipped with the function ϕ, is guaranteed to solve the social circle discovery problem. However, as one may expect, that algorithm, and the subsequent guarantee of optimality, require complete knowledge of the graph. Here, we are concerned specifically with the local version of the problem, where only those nodes that members of the growing social circle link to are available. While we cannot guarantee global optimality in this context, if the algorithm adopts an alternative greedy approach where at each step it selects the node that maximizes ϕ among all candidate nodes, then we retain at least some local optimality. Details are shown as Algorithm 3.

Algorithm 3 takes as input the query nodes, the maximum size of the desired social circle and the frequency of removal, and produces as output a social circle of at most the specified size. Lines 1-5 set the add-step counter to 1, assign that value to all of the query nodes, and initialize the social circle to the query set. The number of iterations is initialized to 1 on line 6. Its purpose is to assist in the node removal process. As we wish to consider node removal with frequency f, i.e., after every f iterations through the main loop, we can use the number of iterations so far and check for node removal every time it is divisible by f, as shown on line 16, where $\ \mathrm{mod}\ $ is the modulo operator. Lines 7-25 contain the main loop, which runs until the social circle reaches the user-specified size. On line 8, the add-step counter is incremented by 1. On line 9, the set of all neighbors of the current social circle is computed as the set of all nodes that any member of the social circle links to. Lines 10-12 handle the possibility that the algorithm runs out of candidate nodes to add to the social circle before reaching the maximum size limit set by the user. If there are no neighbors to consider, the algorithm simply breaks out of the loop. Otherwise, on line 13, the neighbor node that maximizes ϕ is selected. In the event of a tie, the tie is broken by the δ function

ALGORITHM 3: Social Circle Discovery Algorithm

Input: Set Q of initial query nodes, maximum size max of the social circle, and frequency of node removal f

Output: A social circle SC of size at most max

$AddStep \leftarrow 1$
for all q in Q **do**
  $s(q) \leftarrow AddStep$
end for
$SC \leftarrow Q$

$NumIter \leftarrow 1$
while $|SC| < max$ **do**
  $AddStep \leftarrow AddStep + 1$
  $N \leftarrow \{n \mid \exists c \in SC \wedge e(c, n) = 1\}$
  if $N = \emptyset$ **then**
    break
  end if
  $w \leftarrow \text{argmax}_{n \in N}(\phi(n, SC))$
  $s(w) \leftarrow AddStep$
  $SC \leftarrow SC \cup \{w\}$
  if $NumIter \mod f = 0$ **then**
    $c \leftarrow \text{argmin}_{c \in \{SC \setminus Q\}}(\phi(c, SC))$
    $SC \leftarrow SC \setminus c$
    for all $x \in SC : s(x) > s(c)$ **do**
      $s(x) \leftarrow s(x) - 1$
    end for
    $AddStep \leftarrow AddStep - 1$
  end if
  $NumIter \leftarrow NumIter + 1$
end while

Return SC

from Equation 6.3 (i.e., using the sum of the terms rather than the min). If a tie still remains, it is broken arbitrarily. On lines 14-15, the winning node's add-step is set and the node is added to the social circle. Upon successfully passing the test of line 16, every f iterations, lines 17-22 effect node removal. On lines 17-18, the node in the current social circle with the smallest ϕ value is selected and removed from the social circle. Note that we explicitly exclude the nodes of the query set from this selection as it makes little sense to remove them. In order to avoid skipping add-step values, lines 19-22 decrement by 1 the add-step values of all of the nodes that were added to the social circle after the node being removed, and then decrement by 1 the add-step counter. Finally, the number of iterations is incremented by 1 on line 24. Once a social circle of size at most max has been found, it is returned (line 26).

There are two main contributors to the computational complexity of Algorithm 3: the discovery of neighbor nodes (line 9) and the evaluation of the ϕ value with regard to these nodes (lines 13 at each iteration, and line 17 every f iterations). For simplicity, let $c = |SC|$ and let d be the average out-degree of any node in the overall graph. For each node in SC, the algorithm checks all of its out-links and adds the corresponding nodes to the set of neighbors. Hence, the complexity of computing the set N of neighbors (line 9) is $O(cd)$. Now, let n be one of the neighbors in N. In order to compute $\phi(n, SC)$, the algorithm needs the in-degree and out-degree of n with respect to SC. The in-degree, $InDeg(n, SC)$, can be obtained by iterating over the elements of SC and checking whether n is one of the nodes they link to. Hence, the complexity of computing $InDeg(n, SC)$ is $O(cd)$, if we assume a linear search through the out-nodes. The out-degree, $OutDeg(n, SC)$, requires finding all of the nodes that n links to, and for each, check whether it belongs to SC. Hence, the complexity of computing $OutDeg(n, SC)$ is also $O(cd)$, again assuming linear search through SC. Computing the weighted versions of these quantities and finding the minimum is O(1), so that the complexity of computing $\phi(n, SC)$ is $O(cd)$. Since the size of N is $O(cd)$, the complexity of finding the node that maximizes phi (line 13) is $O(c^2d^2)$. All other steps of the algorithm are trivially $O(1)$. Now, the main loop (lines 7-25) is executed a finite number of

times bounded by max, hence the algorithm's overall computational complexity is $O(c^2d^2)$. Furthermore, note that $c \leq max$ ad max is a finite value selected by the user. Hence, the complexity of Algorithm 3 is $O(d^2)$.

Notice that if incoming links can be observed directly, so that any node may have access to all of the nodes it links to as well as all of the nodes that link to it (e.g., Twitter users that follow an account), then with the use of hash tables to store these lists, it is possible to reduce the complexity of computing both $InDeg(n, SC)$ and $OutDeg(n, SC)$ to $O(c)$. And in this case, the complexity of Algorithm 3 is only $O(d)$. This savings can be dramatic in some situations, such as those shown in Section 6.6.1 where the degree of the nodes is large (e.g., $d > 10^6$).

We now turn to an empirical analysis of Algorithm 3 through synthetic benchmark datasets, networks for which communities have been identified a priori (and thus serve as ground-truth for testing purposes), and several real case studies that exercise the unique features of our approach.

6.4 Benchmark Results

Using the established LFR benchmark [133, 134] for directed graphs we can objectively evaluate the quality of our algorithm. It is important to note that these benchmarks were designed for the more traditional community mining problem, in which the community boundaries are clearly defined. Yet, our method is not hurt by this added property. We first consider the case of disjoint communities, wherein every node belongs to exactly one community. Next, we use benchmarks that include nodes with overlapping community memberships, wherein a certain number of nodes belong to multiple communities. For simplicity, we restrict our attention to unweighted graphs, where edges have a value of 1 when a connection exists and 0 otherwise.

For comparison, we consider three common local community mining algorithms. Even though these algorithms were designed for community mining, as opposed to finding social

circles around a query set of nodes, the comparison provides a quantifiable way to evaluate our approach. We consider 1) Clauset's local modularity, which seeks to find a steep boundary [50]; 2) Greedy Clique Expansion, which maximizes the number of internal to total links in a density-like fashion [135, 139]; and 3) Iterative Scan, which alternates between phases of adding and removing nodes to maximize a density metric [21]. For parameters, for the Greedy Clique Expansion, a value of $\alpha = 1$ in the recommend range is used, and for our algorithm we use default values of $\alpha = 1$ and $f = 3$.

6.4.1 Disjoint Communities

First, we consider the case of graphs where every node belongs to a distinct community with no overlapping memberships. We generate a set of directed LFR benchmark graphs, each with 1,000 nodes, varying the community size range to be 20-50 nodes and also 40-100 nodes. In addition, we use two different values, 0.2 and 0.4, for the mixing parameter μ, which defines the amount of linking between nodes in different communities. The other parameters were held constant at standard default values, as follows: average in-degree $k = 15$, maximum in-degree $maxk = 50$, minus exponent for degree sequence $t_1 = 2$, minus exponent for community size distribution $t_2 = 1$, and total number of nodes $N = 1,000$.

For each configuration setting, a separate social circle is discovered around each of the 1,000 nodes as the initial query node. It should be noted that the Greedy Clique Expansion and Iterative Scan methods are designed to find all communities in a network and in so doing, they prescribe processes for determining pockets of nodes from which to begin, and then expand around them. However, in this case we are interested in finding a separate social circle around every node in the graph. Thus, we compare only the expansion phases of these algorithms, not their seeding strategies. Similarly, it should be noted, that our algorithm (and likely the others as well) would perform better if the initial query set included additional nodes from the desired community, but for comparison, only the single starting node is used.

Because local modularity and our approach do not contain a hard stopping criterion, all of the algorithms are stopped when the size of the social circle matches the size of the correct community defined in the benchmark (e.g., if the correct community has 25 members, the algorithms run until the social circle contains at most 25 nodes). In the case of the Greedy Clique Expansion and Iterative Scan methods, if their terminating conditions are reached prior to this point, then the discovery is halted at that point. These benchmarks are directed and are treated as if only outgoing connections can be determined, as is the case with many real-world networks (e.g. blog links, citations, etc.). Thus, if an algorithm seeks to discover the neighbors of a node, only the outgoing neighbors are returned. Once a social circle is discovered, it is compared against the correct community by evaluating the F-Measure. A separate F-Measure value is determined for the social circle around each start node, and then averaged across the 1,000 circles. The results are shown in Figure 6.5, where the error bars represent one standard deviation.

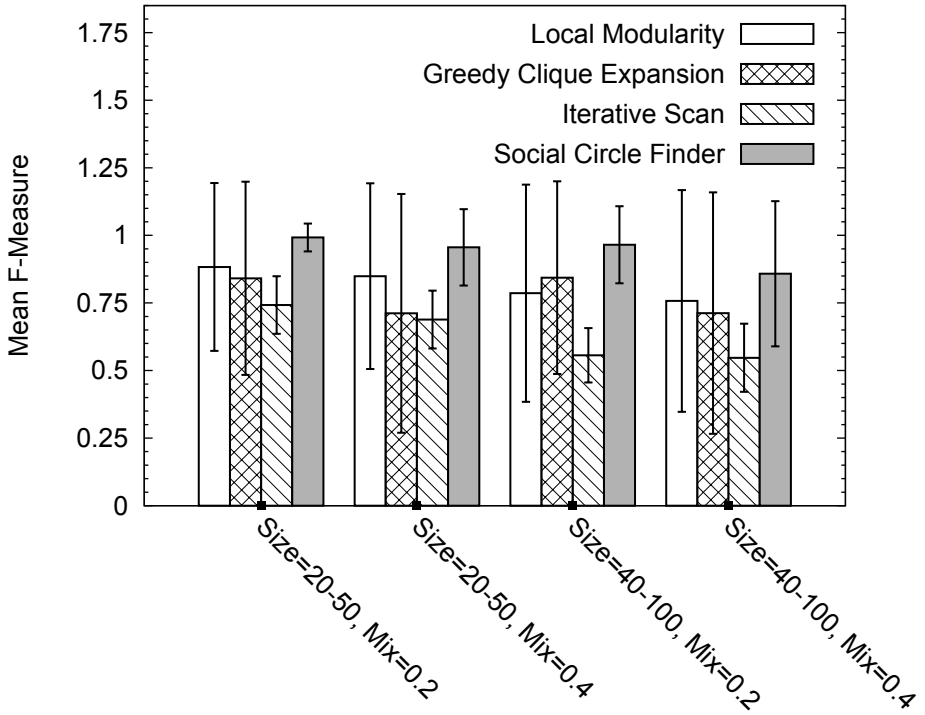

Figure 6.5: Comparison on non-overlapping communities.

As shown, in each case our method performed significantly better than the other three methods on these benchmarks (t-test, $p < 0.01$). The large variance in the values is a result of the fact that if algorithms added nodes outside the community early in the process, they would likely continue adding nodes outside the community. This further demonstrates the value of starting with a set of query nodes, as opposed to just a single one. The relatively smaller variance of our method is the result of enforcing that the start node remains prominent which helps avoid discovering a completely different dense set with the start node on the fringe. The Iterative Scan method had excellent precision (> 0.98 on average for each network), but often terminated before identifying the complete set.

6.4.2 Overlapping Communities

Next, we evaluate the ability of each algorithm to discover social circles in graphs where nodes may belong to multiple overlapping communities. For this comparison, we generate a set of directed LFR benchmark graphs with overlapping communities. We use the same parameters as before, this time holding constant the community size range at 20-50, and the mixing parameter $\mu = 0.2$. We vary the number of nodes that have overlapping memberships to be either 100 or 300 (10% or 30% of the nodes), and also vary the number of memberships for those overlapping nodes to be either 2 or 4 communities.

As above, separate social circles are discovered around each of the 1,000 nodes in the graph. However, in the case of the nodes that belong to multiple communities, it is ambiguous which of the overlapping communities is desired, so in this case we attempt to discover each of the overlapping communities, by starting the algorithm separately with the node and an arbitrary neighbor in each desired community. The results were again averaged across all social circles discovered in the graph with the mean and standard deviations of the F-Measure shown in Figure 6.6.

In this case, our approach outperformed each of the others significantly on the first two benchmarks (t-test, $p < 0.01$). It was also the highest in the third case, but the results

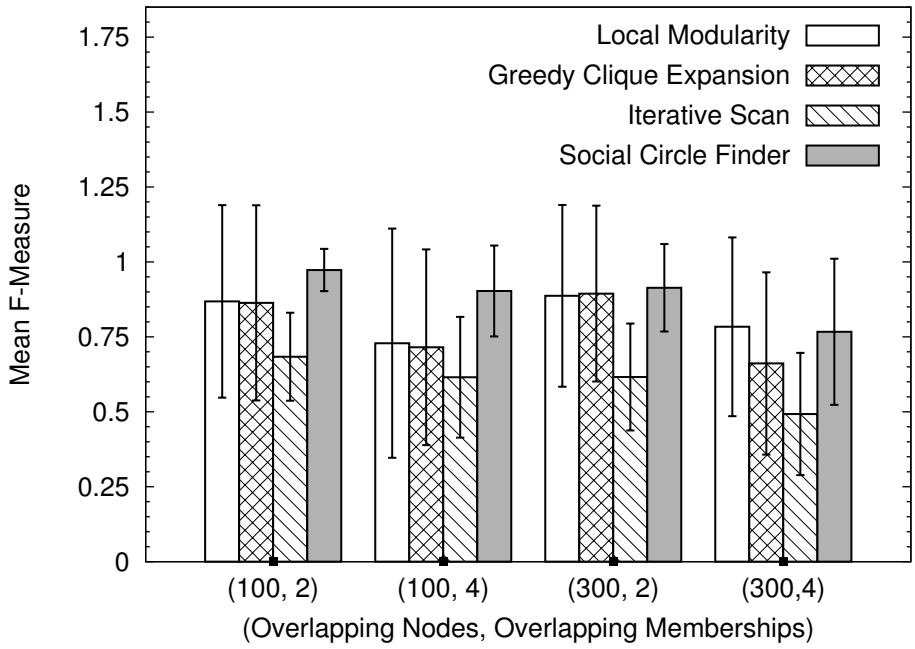

Figure 6.6: Comparison on overlapping communities.

were not statistically significant, and in the fourth case, Local Modularly was slightly higher, but again the results were not significant. As before, the large amounts of variation arise from the fact that when algorithms "missed" the correct community, they tended to completely miss it. On the other hand, our algorithm was less susceptible to adding dense sets away from the start node, resulting in lower variance and suggesting increased robustness.

6.5 Ground-truth Communities

In addition to the artificial benchmarks described above, we also evaluate our approach on real-world networks with user-specified labels. For our evaluation, we use two large directed datasets, one using data from Flickr and the other from YouTube [162]. The Flickr dataset consists of 1.8M users crawled from the site, containing 22M directed links, and 104K user groups. The YouTube dataset consists of 1.2M users crawled from the site, along with 4.9M directed links, and 30K user-groups. The user-groups of these datasets express common interests of the users, and in that way can be seen as a type of ground-truth community

assignment. While these datasets and their user-groups may be more in line with the traditional community mining problem, they provide an avenue for quantitative evaluation of our approach in comparison to local community mining algorithms. Because these user groups are based on common interests, in many cases, the members of the community are actually not well-connected (if connected at all), making it difficult to discover them through structure-based algorithms. Despite these limitations, these user-defined associations still provide a valuable opportunity for quantitative real-world analysis.

6.5.1 Query Node Selection

As discussed, an important element of the local social circle discovery problem is the selection of query nodes. While in many cases the query set is determined beforehand, and is the reason for discovering the social circle, in other cases, it may be that an initial member and a characteristic of interest are known, but neighbors of the individual need to be added to the query set to discover the desired social circle. For example, consider the case of identifying a social circle of business contacts around an individual. The question arises, which co-workers should be included in the query set to best define the social circle? Using the ground-truth datasets, we evaluate different selection mechanisms for selecting a second member of the query set, given a start node and a community of interest. We consider the following possible selection criteria:

1. Arbitrary. Select an arbitrary member of the community.
2. Least Other Groups. Select the node that belongs to the least number of other communities.
3. Least Overlapping Groups. Select the node that has the fewest number of communities in common with the first.
4. Least Outside Friends. Select the node that has the fewest connections outside the desired community.

5. Highest In-group Ratio. Select the node that has the highest ratio of friends inside the community to those outside.

6. Most Inside Friends. Select the node that has the greatest number of friends in the desired community.

To evaluate these different criteria, we select arbitrary nodes from the network, and discover the various overlapping communities it belongs to. For each of these communities, we select a second node based on each of the different criteria and use the two nodes as a query set to discover a social circle. We treat the network as if only outgoing connections are known. For consistent comparison, in each case, we add 20 nodes to the query set and count the number of them that are part of the desired ground-truth set. We exclude communities where none of the neighbors are in the desired set and those that had fewer than 20 additional connected members in the ground-truth community. Figure 6.7 shows the average number of correct nodes added to the set for 8,158 evaluations each on the YouTube dataset and 1,949 evaluations each on the Flickr dataset. As shown, for each data set, selecting the node with the highest ratio of internal friends to external friends is the most effective way of selecting a second query node.

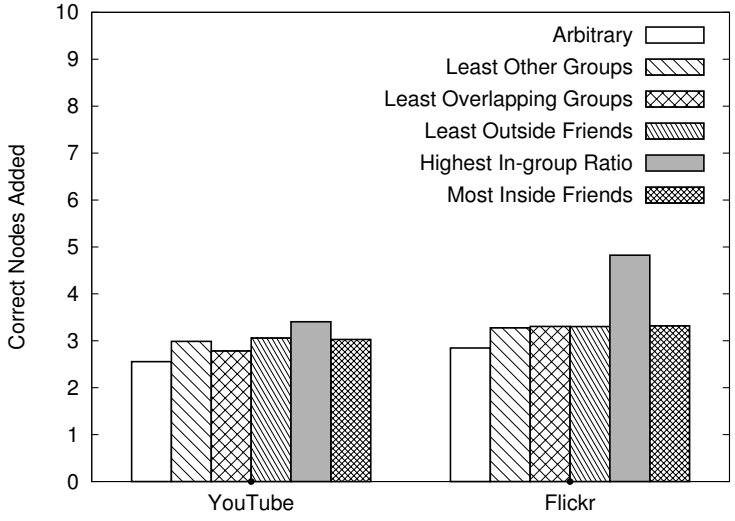

Figure 6.7: Number of correct nodes found in the first 20 for different query selection mechanisms.

6.5.2 Importance of Additional Query Nodes

In addition to the manner of selecting additional query nodes, the *number* of these initial nodes can also potentially impact the effectiveness of discovering other nodes in the desired local social circle. Using the best method of query node selection above (highest in-group ratio) we discover social circles around query sets that range in size from 2 to 10 members. Again, for consistent comparison, we evaluate the number of correctly identified nodes in the first 20 added after the initial query set. Similar to the previous experiment, we select arbitrary nodes from the network and for each of their ground-truth community memberships, we select a query set of different sizes. As the largest query set requires starting with 10 nodes and discovering 20 more, we exclude communities where the initial node has fewer than 9 direct neighbors in the community, and where the connected component of the ground-truth community contains fewer than 30 members. Figure 6.8 shows the number of additional nodes correctly identified in the first 20 added after the query set for 2,132 evaluations each on the YouTube data and 1,331 evaluations each on the Flickr data.

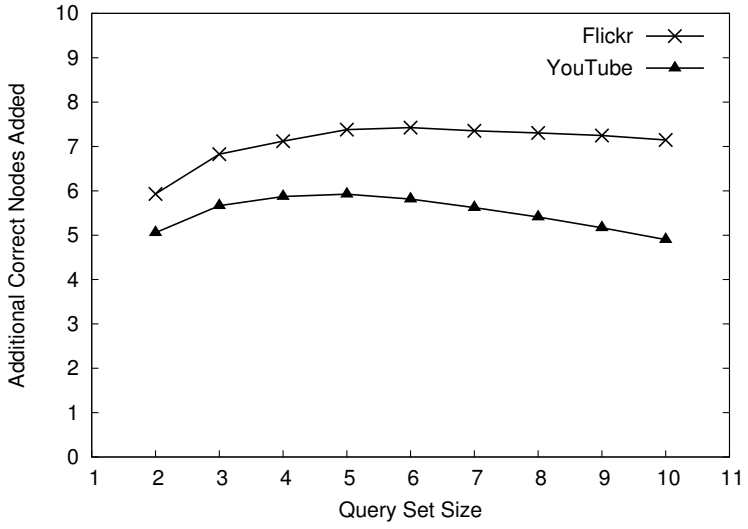

Figure 6.8: Number of correct nodes found in the first 20 (after the query set) for query sets of different sizes.

As shown, using more than two nodes helps to better identify the social circle of interest with the largest increase in value coming from adding a third and fourth member to

the query. Interestingly, for these datasets, having more than 5 or 6 query nodes does not increase effectiveness. The fact that the number of correct additional nodes declines may be because the additional query members were some of the "easier" nodes to identify, so by starting with them already in the social circle, the task is to find other, potentially more "difficult," members.

6.5.3 Algorithm Comparison

Using these same data sets, we can also compare the performance of the different algorithms in discovering the ground-truth communities. As before, we compare the number of correct nodes of the first 20 added after the query set. Because the Iterative Scan algorithm alternates through phases of addition and deletion, it cannot be cleanly stopped at a specific number of members, and therefore is not included in this comparison. As with the previous experiments, we arbitrarily select nodes from the networks and for each community to which they belong, we select a second node for the query set and discover a social circle around them. For the selection of the second node, we use the best approach from before (highest in-group ratio), and exclude communities where the initial node has no direct neighbors in the community and where the number of additional connected members of the ground-truth community is less than 20.

As the algorithms run at different levels of efficiency, some were able to complete more evaluations than others. The relative efficiency of our approach is noteworthy. Thus, we show the rolling average over the number of iterations completed. Figure 6.9 shows the rolling average per iteration for the YouTube dataset and Figure 6.10 shows the averages for the Flickr dataset. As can be seen, our method clearly outperforms and is more efficient than the other methods.

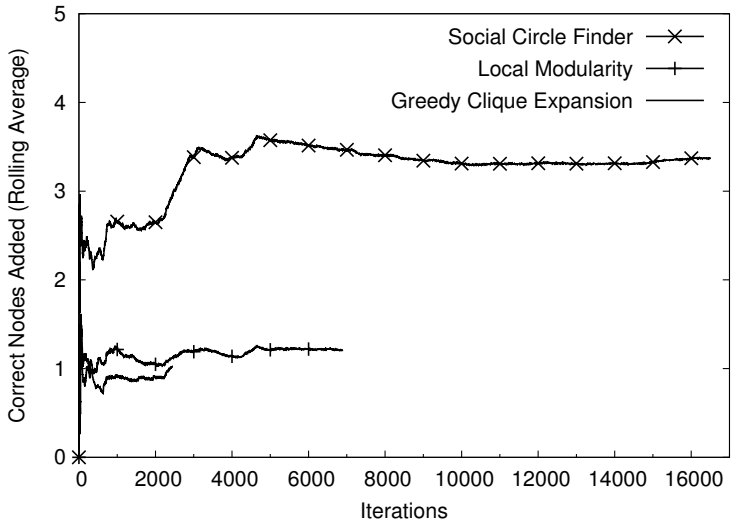

Figure 6.9: Number of correct nodes added out of the first 20 on the YouTube dataset.

6.6 Case Studies

Finally, in addition to analyzing results on benchmark graphs and on anonymous networks with ground-truth community assignments, we also qualitatively validate our approach in two different real-world social networks: Twitter and the blogosphere. These networks complement each other as case studies because they have significantly different graph properties. On the other hand, each of these graphs is directed, incredibly large and complex, and requires a local solution.

6.6.1 Twitter User Social Circles

We first apply our approach to building social circles of users on the social network platform Twitter. For demonstration purposes, we have chosen to consider social circles around well-known individuals (at least in the United States). For all query individuals, we show the number of individuals who follow them (followers) and the number of individuals they follow (following) as of 25 January 2013. Clearly, an individual's lists of followers and following vary over time. We include the numbers here only to give a sense of the relative sizes of these lists and the "social status" of the corresponding individuals. Also, the teams of professional

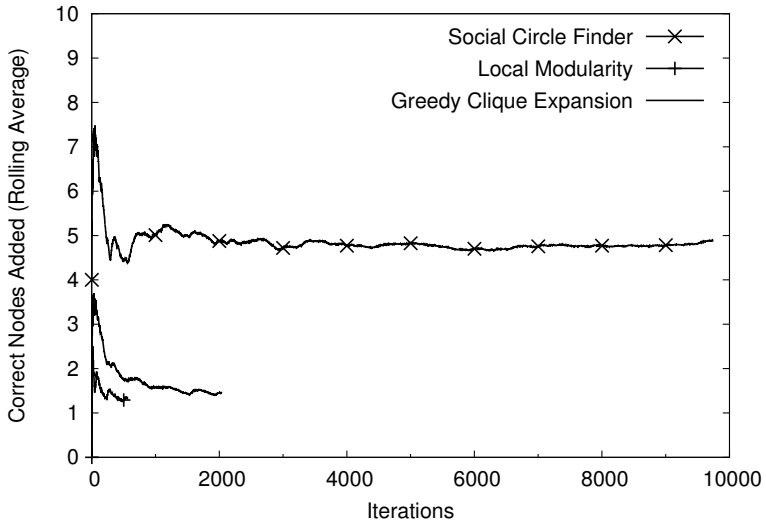

Figure 6.10: Number of correct nodes added out of the first 20 on the Flickr dataset.

athletes and the positions of politicians are not constant over time, and we report them as they were at the time of the discovery in January 2013.

We first turn to professional basketball players, and discover a social circle around three prominent players: LeBron James, NBA player for the Miami Heat (Twitter account: @KingJames; followers: $\sim$ 7M; following: 286), Derek Fisher, NBA player for the Dallas Mavericks and president of the NBA Players Association (Twitter account: @DerekFisher; followers: $\sim$ 930K; following: 189), and Rajon Rondo, NBA Player for the Boston Celtics (Twitter account: @RajonRondo; followers: $\sim$ 885K; following: 62). By choosing players from different cities, we avoid discovering a social circle focused on a certain market such as radio or TV personalities from that city. As discussed earlier, the goal is to choose an initial set such that the only common characteristic is the desired trait (in this case, NBA players). Using these three accounts as the query set, we apply our Social Circle Discovery algorithm to build a social circle of $max = 75$ members, with $\alpha = 1$ and $f = 3$. The first 20 members of the resulting social circle are shown in Table 6.1. Of the 75 members of the discovered social circle, 62 were NBA players or groups, 4 were affiliated with the NBA (such as former players, trainers, and agents), 2 were other professional athletes, 5 were other popular figures (such as musicians and actors), and 2 were athletic news organizations.

Table 6.1: NBA Social Circle Members

Step	Twitter Account	Name	NBA Team
1.	KingJames	LeBron James	Miami
1.	derekfisher	Derek Fisher	Dallas
1.	RajonRondo	Rajon Rondo	Boston
2.	KDTrey5	Kevin Durant	Oklahoma City
3.	rudygay22	Rudy Gay	Memphis
4.	John_Wall	John Wall	Washington
5.	russwest44	Russell Westbrook	Oklahoma City
6.	DWRIGHTWAY1	Dorell Wright	Philadelphia
7.	Baron_Davis	Baron Davis	New York
8.	JCrossover	Jamal Crawford	Portland
9.	CP3	Chris Paul	LA Clippers
10.	NBA		NBA Account
11.	nate_robinson	Nate Robinson	Golden State
12.	MikeVick	Mike Vick	
13.	KyrieIrving	Kyrie Irving	Cleveland
14.	BooBysWorld1	Daniel Gibson	Cleveland
15.	RealTristan13	Tristan Thompson	Cleveland
16.	SteveNash	Steve Nash	LA Lakers
17.	Avery_Bradley	Avery Bradley	Boston
18.	unclejeffgreen	Jeff Green	Boston

Using LeBron James as a starting point, it is actually possible to be interested in, and discover, other social circles or overlapping communities. Indeed, in addition to being a professional basketball player, LeBron James is also a figure of popular culture, so that another social circle may be obtained if we include popular figures rather than professional basketball players in the query set with him. To verify this hypothesis and further validate our Social Circle Discovery algorithm, we re-run the algorithm with a query set comprising Lebron James and two pop culture individuals: Ciara, a musician (Twitter account: @Ciara; followers: $\sim$ 3M; following: 67), and Charlie Sheen, and actor (Twitter account: @CharlieSheen; followers: $\sim$ 9M; following: 106). As before, $\alpha = 1$ and $f = 3$. However, we set $max = 20$ as most of these individuals have very large lists of followers, which greatly affects computation time due to the request rate restrictions enforced by Twitter. The members of the resulting social circle are listed in Table 6.2.

All of the members of this social circle are entertainers of some kind, and each of their Twitter accounts has been "verified" by Twitter as the correct account of a popular figure.

Table 6.2: Popular Culture Social Circle Members

Step	Twitter Account	Name / Stage Name	Status
1.	KingJames	LeBron James	Athlete
1.	ciara	Ciara	Musician
1.	charliesheen	Charlie Sheen	Actor
2.	Ludacris	Ludacris	Musician
3.	SnoopDogg	Snoop Dogg	Musician
4.	lala	La La	Entertainer
5.	iamdiddy	P. Diddy	Musician
6.	NeYoCompound	Ne-Yo	Musician/Actor
7.	chrisbrown	Chris Brown	Musician
8.	KevinHart4real	Kevin Hart	Actor
9.	carmeloanthony	Carmelo Anthony	Athlete
10.	myfabolouslife	Fabolous	Musician
11.	Wale	Wale Folarin	Musician
12.	Tyrese	Tyrese Gibson	Musician/Actor
13.	djkhaled	DJ Khaled	Music Producer
14.	MeekMill	Meek Mill	Musician
15.	CP3	Chris Paul	Athlete
16.	Nas	Nasir Jones (Nas)	Musician
17.	DwyaneWade	Dwyane Wade	Athlete
18.	DJCLUE	DJ Clue?	Musician

Of the 20 members of this set, 11 are musicians, 5 are entertainers (actors, musicians/actors, etc.), and 4 are professional athletes. This group clearly represents a rather different social circle to which Lebron James also belongs. Incidentally, his friendship with the famous rapper, and part-owner of an NBA team, Jay-Z, was made newsworthy over whether the friendship could help lure him to that team.

We note that it would be difficult for boundary-focused community detection algorithms to discover a community of popular figures because of their numerous links with outsiders (the Lab Advisor Problem). Properly handling the links from outsiders also requires a directed approach, and illustrates the importance of accounting for mutual connection to the growing set (the Famous Person Problem). In addition, algorithms that require iteratively trying each outside member as a member of the community, such as those in the benchmark comparison, cannot be effectively run on Twitter with these highly-popular users because it would require millions of calls to the Twitter API (which limits request rates). For this reason, we have not included comparison with the other algorithms used on the benchmark graphs. By contrast,

our approach can be run, albeit still slowly in some cases due to the rate limitations, because we are required only to know the follower/following lists of the members of the growing social circle.

In addition to professional athletes, members of the United States Congress have become prominent users of Twitter, and have strong ties to one another, particularly other members of the same political party. Using our approach and a query set of members of each party, we can discover other representatives from that party. Choosing five Democrats and five Republicans, we build two separate social circles of 100 members (i.e., $max = 100$), with $\alpha = 1$ and $f = 5$. For the initial query set, we selected the party leaders in the House of Representatives, as well as two additional members of the House, and two members of the Senate. The initial query sets for the two social circles are as follows.

- Democratic Congress Query Set
  - Nancy Pelosi, House Minority Leader (Twitter account: @NancyPelosi; followers: $\sim$ 300K; following: 248)
  - Steve Israel, House of Representatives (Twitter account: @RepSteveIsrael; followers: $\sim$ 10K; following: 226)
  - John Conyers, House of Representatives (Twitter account: @RepJohnConyers; followers: $\sim$ 6K; following: 438)
  - John Kerry, Senate (Twitter account: @JohnKerry; followers: $\sim$ 60K; following: 223)
  - Charles Schumer, Senate (Twitter account: @ChuckSchumer; followers: $\sim$ 47K; following: $\sim$ 28K)
- Republican Congress Query Set
  - John Boehner, Speaker of the House (Twitter account: SpeakerBoehner; followers: $\sim$ 437K; following: $\sim$ 14K)

- Jason Chaffetz, House of Representatives (Twitter account: @JasonInTheHouse; followers: $\sim$ 35K; following: $\sim$ 22K)
- Darrell Issa, House of Representatives (Twitter account: @DarrellIssa; followers: $\sim$ 72K; following: $\sim$ 23K)
- John Boozman, Senate (Twitter account: @JohnBoozman; followers: $\sim$ 12K; following: 259)
- Roy Blunt, Senate, (Twitter account: @RoyBlunt; followers: $\sim$ 22K; following: $\sim$ 8K)

Each of the members of the discovered social circles were involved in politics, even though not all of them were actually representatives. Table 6.3 shows statistics of the resulting social circles.

Table 6.3: United States Congress Twitter Social Circles

	Democratic	Republican
Total Members	100	100
Congress (same party)	94	71
Congress (other party)	0	0
News and Reporters	6	14
Foundations and Activists	0	15

Of the 100 members of the Democratic set 94 were accounts for Democratic representatives (either individual accounts, or groups such as the official account for a Democratic congressional committee). In the Republican set, 71 of the 100 members were accounts for Republican representatives or their groups. In each of these social circles there were many accounts of other politically involved users (news organizations, foundations, etc.) that were included due to their large number of mutual connections with the representatives. In the case of the Republican set, there were more foundations and political activists than the Democratic set, possibly suggesting that the Republican representatives are more likely to have mutual links to these users. It is also interesting that no representatives of the opposite

political party were discovered in the social circles, suggesting little direct overlap among members.

6.6.2 Blog Social Circles

One of the characteristics that has contributed to the success of the blogosphere is the fact that authors link to each other's posts. Dense connections between common blogs can define social circles within the blogosphere and because the entire set of blogs cannot be feasibly known, a local discovery method is required to discover these sets. To discover a social circle of blogs, we downloaded the latest 50 blog entries for each blog, and crawled the content for links. The links were then examined and if the resulting page contained a FeedURL in its metadata, it was considered a blog. A case study of blog social circles complements that of Twitter social circles nicely, because whereas on Twitter many graphs are densely connected and it is common for many users to follow those that follow them, a blog social circle defined by links to other blogs is much more sparse.

A prominent interest that exists within the blogosphere is that of "mommy-blogs," where mothers post about their experiences raising children and homemaking, and link to one another. To discover a social circle around a set of mommy-blogs, we selected the query set by choosing an arbitrary blog from the "Top Rated Mommy Blogs" at TopMommyBlogs.com [1], and crawled its neighbors to identify four more that had connections between them, and that by manual inspection appeared to be mothers talking about events, as opposed to an automated feed or coupon service.

Using this initial set, and our algorithm with $\alpha = 1$ and $f = 5$, we identified a social circle of $max = 100$ blogs. A visual representation of this set is shown in Figure 6.11, and the first 25 blogs are listed in Table 6.4.

Each of the 100 blogs were considered mommy-blogs to some degree, in that they dealt with issues related to homemaking, children, and thriftiness. In addition, Figure 6.11 shows

[1] http://www.topmommyblogs.com/pages /top_rated_mommy_blogs.html

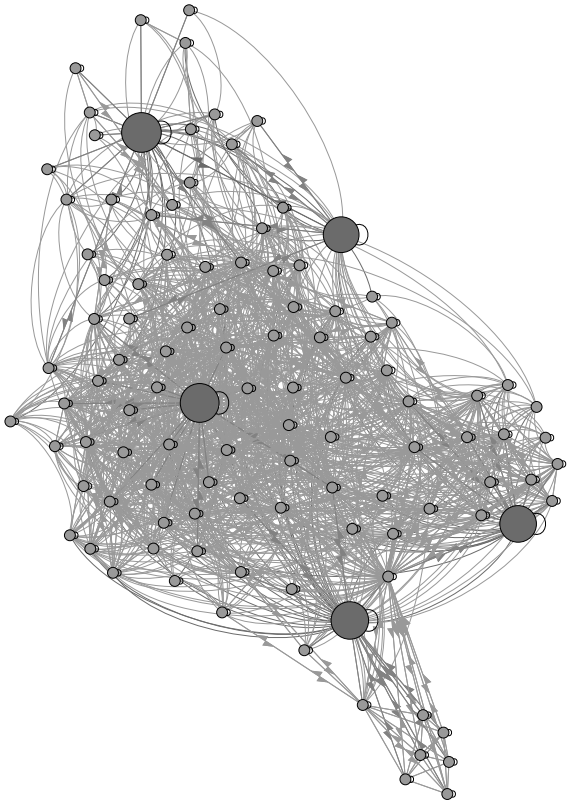

Figure 6.11: The social circle of mommy-blogs. The larger nodes are the initial query set.

that the initial query set (shown as larger nodes) remain highly-connected and prominent in the resulting social circle, as opposed to being left on the fringe while a dense adjacent group is discovered (the Fringe Problem).

6.7 Conclusions and Future Work

In this paper, we have defined the local social circle discovery problem in directed graphs, and proposed a novel algorithm to discover such social circles around an initial query set, based on a degree-inspired quality function that quantifies the value of adding a node to the growing social circle. Our approach does not focus on boundaries and can therefore include appropriate nodes in a social circle regardless of their membership in other circles. In addition it stays focused around the original query set, as opposed to drifting into other parts of the graph, leaving the initial query nodes on the fringe of the final social circle.

Table 6.4: The First 25 Members of the Mommy-blog Social Circle

Step	Blog URL
1.	momtobedby8.com
1.	stuckathomemom.com
1.	guideformoms.blogspot.com
1.	autumnandkids.com
1.	mamaluvsbooks.com
2.	thegiveawaygals.com
3.	blog.stay-a-stay-at-home-mom.com
4.	lifesabargain.net
5.	mewreview.com
6.	confessionsofamessymama.blogspot.com
7.	tinklemonkey.com
8.	swanksavings.com
9.	amedicsworld.com
10.	prmomambassador.com
11.	countingtoten.com
12.	earndollarspinoy.info
13.	to-sew-with-love.com
14.	funnypregnantlady.blogspot.com
15.	nikkicole22654.blogspot.com
16.	budgetearth.com
17.	justjennifer.net
18.	carolscrittercorner.com
19.	mommies-in-orbit.com
20.	alittlesimplicity.com
21.	momat40.com

Further, our approach explicitly accounts for edge direction and avoids including celebrities or unknown nodes that do not have mutual interaction with the social circle. We show that our Social Circle Discovery algorithm performs well on artificial benchmark problems, large networks with ground-truth communities, and through case studies in real-world networks. Our method is able to efficiently discover meaningful social circles even when the degree of the included nodes is extremely high.

There are two interesting extensions to our algorithm that could be pursued. While we have explicitly accounted for directed edges, we are still only handling unweighted graphs. It would be interesting to consider ways to incorporate weighted edges in the algorithm. One simple solution would be to replace the current indicator function $e(x, y)$ by a number-valued function corresponding to the weight of the edge. If the semantic associated with edge

weights is that of a notion of strength of the relationship between the connected nodes, then the interaction between this extension and the existing discounted importance mechanism of our algorithm may result in the expected behavior. If not, further extensions may be needed to properly account for the intended meaning of the weights. Another area of interest, also related to the directed nature of the graphs, has to do with the relative value of incoming and outgoing edges. In the current implementation, both $Wgtin - degree$ and $Wgtout - degree$ are treated equally in $\phi(n, C)$. There may be value, depending on the application, in weighing these quantity differently, perhaps using a parameter β to transform $\phi(n, C)$ into $\min(\beta WgtInDeg(n, C), (1 - \beta)WgtOutDeg(n, C))$. Further experiments are needed.

Chapter 7

Social Moms and Health: A Multi-platform Analysis of Mommy-communities

Abstract

The explosion of online social media has increased people's ability to share content and link with others, thus allowing diverse communities to emerge naturally as a product of interaction among participants. Mothers have certainly not been foreign to this development. Many have embraced the new technology to share experiences, thoughts, current events, reactions, and tips with their peers. Recognizing the role of mothers as decision-makers in their families, especially in the context of health, we focus our attention on "mommy-communities" in Twitter and the blogosphere. We consider what health topics are discussed by mothers in these communities, identify and compare implicit affinities to explicit links, and highlight differences and similarities across the two social media platforms.

7.1 Introduction

Increased user participation online has led to the emergence of a large number of communities arising naturally as people share content with one another and link to each other on social media sites, such as Twitter, YouTube and Facebook, as well as in the blogosphere and on subject-specific Web sites. These online communities supplement more traditional forms of communication (e.g., telephone, email) and greatly enhance the way in which people interact. The science of building, discovering, understanding and leveraging such communities is gaining popularity as the Internet becomes the largest collection of ideas, personalities, and cultures in human history.

One important example of these interlinked communities is that of "Mommy blogs," where mothers share stories about their experiences raising children and nurturing their families. These mommy blogs are read and linked to by other mothers with blogs. Similarly, many mothers use micro-blogging on Twitter to share thoughts, current events, reactions, and tips. Other mothers may then follow or mention them. Such interactions result in the formation of communities that can powerfully influence the social norms of their members, and in turn the decisions made by other mothers in the community [6, 242]. Indeed, among social agents, mothers are a rather interesting group to study because of the central role they play as decision-makers and influencers in the home, especially in the context of health behaviors [57, 91, 159]. Given the influence that social networks have on individuals, it may be valuable to understand how mothers interact online as this is likely to affect their own perceptions, and hence those of their families. The present study is a step in that direction. It aims at discovering, analyzing and comparing mommy communities. To this end, we build a community of mommy Twitter-users to parallel that of mommy-blogs, and highlight differences between the communities. In particular, we address the following research questions and hypotheses.

The present study is a step in that direction. It is aimed at discovering, analyzing and comparing mommy communities in the blogosphere and Twitter. To this end, we build a community of mommy Twitter-users to parallel that of mommy-blogs, and highlight differences between the two communities. In particular, we address the following research questions and hypotheses.

- **Research Question 1 (RQ1): To what extent do mothers discuss health topics, as defined by health scientists? What other topics do mothers discuss?** While certain topics, such as pregnancy, are expected among those discussed by mothers, this may not be the case with other health issues. Discovering the extent, or lack thereof, to which critical health issues are being discussed among this target population can inform health communications and promotion practices. Identifying other topics

of interest may inform about current concerns and emerging interests that health practitioners may capitalize on.

- **Hypothesis 1 (H1): There are differences in the way mothers use the blogosphere and Twitter to discuss health issues.** The blogosphere and Twitter are different communication media. For example, blogs are "pull" technology as others have to follow a link or search for the blog in order to access its content; on the other hand, tweets are "push" technology since all followers are automatically exposed to the content as part of their stream. Similarly, tweets are restricted to 140 characters while blogs can be of arbitrary length. Twitter's immediacy and short messages are likely to favor instant communication about feelings, reactions and events; the blogosphere tends to foster thoughtful, more polished and longer-lasting descriptions of impressions, emotions and opinions. As such, we hypothesize that, at least in the context of health issues, there exist significance differences in the way mothers use each platform.

- **Research Question 2 (RQ2): Are mothers who discuss similar topics connected to each other?** A study of the evolution of discussions (mostly about autism and vaccination) on Cafemom.com suggests that the links in the friendship network are consistent with the topics discussed by mothers (i.e., similar interests are reflected in the links among mothers) [3]. We wish to see whether this carries over to Twitter and the blogosphere in the context of a much larger, less directed number of topics. It is likely that mothers who are connected to each other discuss the same, or similar, topics. The question here is whether the converse is also true, i.e., whether mothers who speak about the same topics belong to the same community. If not, then there is potential for increased interaction, including mutual support.

- **Hypothesis 2 (H2): The patterns of directed links among mothers are consistent across Twitter and the blogosphere.** It would seem reasonable to expect that if a mother's blog links to another mother's blog, then the first mother is also likely to follow, mention and/or retweet the second one on Twitter, and vice-versa (provided

both mothers have accounts on each platform). On the other hand, if this is not the case, one may wonder whether this is an artifact of the medium (e.g., it is easier to follow someone than to link to their blog) or perceived level of commitment (e.g., a retweet seems like a much smaller endorsement than a link from one blog to the other).

We recognize that others have looked at the prevalence of health issues in online social media [192, 201], as well as the flow of information in the blogosphere [92] and in Twitter [82, 207]. However, little work has been done in comparing the behavior of users and communities across these platforms. This is one of the main contributions of this paper, where, using mothers as the target population, we highlight noteworthy differences in terms of content as well as structure between Twitter and the blogosphere.

7.2 Methods

Our first task is to build meaningful communities of mommy bloggers and mommy Twitter users. To do so, we started from a small core group of five typical mommy-blogs, that linked to one another, where the authors also had Twitter accounts that followed one another. This was done by first selecting an arbitrary blog (that also had a Twitter account) from the "Top Rated Mommy Blogs" at TopMommyBlogs.com. We then crawled its Twitter and blog neighbors to identify a set of 5 other mothers that: 1) had accounts in both platforms; 2) were densely connected in both platforms; and 3) were not automated feed or coupon services. These two groups of 5 mothers each, one in Twitter and one in the blogosphere, are used as seeds to guide a social circle discovery algorithm applied to the corresponding platform. Finding such a social circle around a core group of individuals is an instance of the community search problem [221], a query-based version of the traditional community mining problem [77].

We consider all links as directed. For Twitter accounts, we define a directed edge from one user to another by the *following* relationship. For blogs, we define directed edges by searching the 20 most recent blog posts for links to other blogs. We note that *mentions*

in Twitter may be more analogous to Blog links, however, because of rate limitations on the Twitter API, it is not feasible to construct a community using this relation. For our purposes, we considered a website to be a blog if it had an RSS feed. Since neither the complete blogosphere graph nor the complete Twitter graph can be feasibly computed, we require a community search algorithm that can work from local information only. While several such algorithms have been proposed, most assume undirected graphs and emphasize the importance of a boundary (e.g., see [19, 50, 148]). As our graphs are directed and we wish to avoid some undesirable boundary effects, we use a local social circle discovery algorithm designed specifically for directed graphs [33].

Intuitively, the algorithm initializes the social circle to the seed set and adds new members one at a time up to a pre-specified size. At each step, all of the individuals linked to by at least one member of the current social circle are candidates for addition. The score of each candidate is the minimum of the number of individuals in the social circle it links to and the number it is linked from. To avoid drifting away from the query set, scores are discounted at each iteration. Formally, the score of candidate n with respect to a social circle SC is:

$$\phi(n, SC) = \min\left(\sum_{c \in SC} e(c,n)s(c)^{-\alpha}, \sum_{c \in SC} e(n,c)s(c)^{-\alpha}\right)$$

where $e(x, y)$ is an edge indicator function (i.e., $e(x, y) = 1$ if there is an edge from x to y, 0 otherwise), $s(c)$ is the time step at which node c was added to the social circle, and α is the discount factor. The candidate with the largest $\phi(n, SC)$ score is added to the current social circle and the process repeats. Ties are broken using the sum of the components rather than the min operator. To increase cohesiveness, the individual with the lowest score is removed after every five iterations. Note that an individual does not become a member of the social circle simply by linking *to* every other individuals, or alternatively having links *from* them all, but because of some mutual interaction. Hence, upon completion, the algorithm returns a social circle composed of dense connections of mutually aware nodes that surround the query

set. Experiments with artificial benchmarks, large networks with ground-truth communities and real-world case studies have been used to demonstrate the validity and effectiveness of the algorithm [33].

We applied the aforementioned greedy algorithm to each of our seed sets to expand both mommy blogger and mommy Twitter user communities with 750 additional members each. To compare mothers across media platforms, we further processed the two communities to restrict our attention to the set of mothers for whom we could identify both a blog and a Twitter account. To find a blog site for a Twitter account, we checked the Twitter website property, and also searched for URLs in Twitter profile descriptions. If either of these resulted in a website with an RSS feed, we considered this to be the user's blog. Similarly, given a blog, we identified a Twitter account by crawling the website for a link to Twitter from the page. If the page had multiple links to Twitter we evaluated each of them to see if any had links back to the blog, and used the Twitter account if there was exactly one that linked back. Using this procedure, and removing any accounts that were not active and accounts for which we could not get an overlapping set of posts (see below) we identified 889 mothers with matching accounts in both platforms.

In order to make sure that issues of time would not confound the results of our content analysis, we considered the same time-frame across the two media for each mother. Because Twitter limits the number of tweets that can be retrieved for a particular user to 3,200, we crawled the corresponding blog for posts spanning the same duration as the total amount of tweets we could obtain for each particular user. In the event that the blog posts were the limiting factor, we reduced the tweets for that user to match the time span of their blog posts. To obtain the blog entries, we used the unofficial GoogleReader API to download entries from the blogs RSS feed. In most cases, the RSS feed contained the text of the entry itself. However in some instances it only referred to a URL for the page. In the cases where only a URL was provided, we downloaded the webpage directly and, if present, restricted the content to an element with the id of "main" or "content." Then, to prepare for content

analysis, we used the HtmlAgilityPack library to extract only the textual content from each page. There were some blog entries for which this process failed to remove the HTML markup tags. To avoid biasing content analysis with extra elements on the page, we excluded the 31 mothers whose blogs had more than 10 entries failing to have their HTML tags removed. Our analysis is performed on the 858 remaining mothers.

To analyze the topics discussed by mothers, we combined a directed approach, where we selected a number of health issues known to be relevant to mothers, with an unsupervised approach based on Latent Dirichlet Allocation (LDA) [26]. One significant advantage of the directed approach is that it allows us to focus on health issues that experts know are of relevance to mothers, but whose underlying low prevalence would make them difficult to unearth using automatic topic detection techniques that generally rely on frequency and co-occurrence. The list of pre-defined topics, shown in Table 7.1, was prepared in collaboration with a subject matter expert, together with associated search terms for each topic.

Table 7.1: Health Topics and Search Terms

Topic	**Search Terms**
Autism	autism, autistic, asperger*, aspie*, asd, pdd
CMV	cmv, cytomegalovirus
Down Syndrome	down syndrome, down's, trysomy, trisomy, chromosome 21
FAS	fetal alcohol, alcoholic embryopathy, arbd, embryopathia alcoholica, fae, fas
SIDS	sids, sudden infant death, cot death, crib death
Pregnancy	pregnan*, pregnen*, obgyn, ob, maternity
Fitness	fitness, exercis*, cardio*
Illness	flu, sickness, illness, antibiotic*, common cold, influenza, medicine*, allergy, allergi*, virus, fever
Nutrition	nutritio*, vitamin*, minerals, antioxidant*
Breastfeeding	breastfeed*, mastit*, colostrum, breastmilk, breast milk, breast pump
Vaccine	vaccine, immuni*
Hospital	instacare, emergency room, afterhours, hospital
Weightloss	weight loss, scale, obese, obesity, fat, diet, weight*
Mental Health	depres*, anxiety, stress*, breakdown, break down, anx*, prozac, mood*, antidepressant*, postpartum, baby blues

On the other hand, the directed approach is limited to those topics that one is able to envisage, making it impossible to discover other relevant, yet less obvious, or even unexpected, issues. Hence, we complemented our directed approach with LDA. We used the common

technique of combining all the tweets from a mother into a single document. We also combined all the blog posts of that mother into the same document. In this way, we got a single composite document per mother and thus we could interpret the results of LDA as topics being discussed by individual mothers, regardless of the medium used. We used the MALLET implementation of LDA [155].

7.3 Results

Table 7.2 provides a high-level summary of various aspects of the Twitter and blogosphere communities. Data was aggregated across all accounts on each platform and the median value for each statistic is shown. The term "entry" is used to represent either a single blog post or a tweet, respectively.

Table 7.2: Median Summary Statistics

	Blogosphere	**Twitter**
Number of Entries	114.00	2362.50
Total unique health topics	6.50	6.00
Entries with any health topic	21.00	26.00
Words per entry	389.27	15.38
Words per day	279.92	153.94
Entries Per Day	0.72	9.71
Health topics per entry	0.24	0.01
Health topics per day	0.18	0.14
Ratio of entries with health topic	0.18	0.01

It is interesting to note at the onset that a recent study from the Pew Research Center suggests that despite the popularity of social network sites, people only sparingly use them for health information [79]. By contrast, our results show a relatively high rate (i.e., 18%) of blog entries mentioning some health topic, as well as 96% of the mothers mentioning a health topic at least once on their blog, and 94% mentioning a health topic at least once on Twitter. On the other hand, the same study also finds that people caring for loved ones use social media more often than others to gather and share health information and support. It seems reasonable to expect mothers to talk about health topics often since they typically are

the ones taking care of their own family members and loved ones especially in the context of health issues. According to the Pew study, among those that are more likely to gather health information online are women and younger adults. This description also fits the majority of users in our Mommy Community and may be another contributing factor to our findings.

7.3.1 RQ1: Health Topics

As stated above, the predefined list of health topics was compiled by a health promotion professional who selected them due to their relevance to mothers. There was no assumption a priori as to the level of awareness among mothers, although it was anticipated that mothers may demonstrate a lack of awareness of some less well-known, yet important —and therefore selected, topics (e.g., CMV). Table 7.3 shows the list of our predefined health topics, together with the percentage of mothers who mentioned that topic at least once in a blog post (respectively, tweet). The topics are ordered from most prevalent to least prevalent.

Table 7.3: Directed Health Topics

Topic	**Blog (%)**	**Twitter (%)**
Weightloss	87.4	81.2
Mental Health	86.2	79.1
Illness	78.1	73.8
Fitness	69.7	67.1
Nutrition	69.3	62.5
Pregnancy	67.2	68.4
Hospital	57.3	42.4
Breastfeeding	32.1	37.4
Autism	22.8	33.1
Vaccine	18.2	11.5
SIDS	5.7	3.1
Down Syndrome	5.4	5.6
FAS	3.4	3.3
CMV	0.1	0.0

As may be expected, some topics such as Fitness, Weight Loss, Illness and Nutrition, are more prevalent. On the other hand, others have very low, if any, representation in our communities. Part of this may be due to the fact that some health issues are of much broader application. For example, every mother has to worry about a sick child at some point;

thankfully, very few have to face the consequence of SIDS. Similarly, some health issues may be perceived as more "shareable" by nature. For example, pregnancy is typically experienced as good news that one wants to share with others; by contrast, FAS is the result of poor, possibly even shame-ridden, health behavior on the mother's part and is thus more likely to be kept to oneself rather than broadcast to the world.

However, another possible explanation for low representation may be a general lack of awareness of certain issues among women. For example, in the US, 1 in 750 of the approximately 4 million children born each year, or over 5,000 children per year are born with or develop permanent problems due to congenital CMV infection, and congenital CMV infections leads to more deaths than Down Syndrome or FAS[1]. Yet, there is virtually no mention of that condition in our mommy communities in spite of a significant level of discussion about pregnancy. Although it is possible that the mothers found in our communities are not entirely representative of all mothers, these results do suggest that from the perspective of health promotion, efforts should be made to promote CMV as well as other relevant health issues among women.

Table 7.4 shows some of the topics we identified from the terms grouped by LDA (50 topics, 2000 iterations). Note that 6 groups (C1-6) appeared to consist of common, generic terms and 7 others (G1-7) seemed to relate to different forms of "give-away," prize, or review topics, and were thus aggregated. Being unsupervised, LDA is therefore clearly not constrained to finding only health topics. Hence, a number of the topics found, such as Couponing, Recipes, Fashion, and Children, are clearly relevant to mothers yet have nothing do with health. As expected, many of the discovered health-related topics, such as Pregnancy and Fitness, align with prevalent topics in our directed approach. Most notably perhaps in this context is topic 2 about Cloth Diapering, which is clearly a health-related subject, and yet one not included in the directed approach. Interestingly, recent years have seen a resurgence of cloth diapering, fueled either by cost sensitivity or eco-friendliness, and

[1] http://www.cdc.gov/cmv/trends-stats.html

there seems to be rather polarized feelings about the issue. The practice clearly calls for a significant and sustained level of commitment, and it is conceivable that mothers who practice it may be seeking support and encouragement from others with similar interest. The discovery of this topic by LDA not only shows the complementarity of approaches, but may also inform health practitioners of emerging interests, trends or concerns among mothers, where intervention may prove useful. For example, cloth diapering enthusiasts could be given additional information on the health benefits of their practice, which they could in turn pass on to their friends, or an otherwise isolated "practicing" mother could be referred to others for support. We return to this idea of recommendation with RQ2.

Table 7.4: Selected LDA Topics

Topic	**Top LDA-discovered Terms**
1. Pregnancy & Birth	baby birth pregnancy pregnant stories quot home hospital
2. Cloth Diapering	baby diaper cloth diapers giveaway win clothdiapers blog green
3. Recipes	recipe food chicken cup chocolate add cheese recipes make butter
4. Recipes	sundaysupper apple wine love pumpkin familyfoodie good chicken food
5. Fitness	fitfluential mamavation missed run running fitness healthy workout
6. Health & Nutrition	healthy home food children make family organic eat
7. Books & Reading	book life read story books people women world good years reading
8. Social Good	goodwill women photo change world social log blog post children share
9. Children & Family	kids great family day time children child make school home
10. Projects & Crafts	party amp love link submitted make paint project home projects
11. Crafts & Sewing	post diy dress content image tutorial craft background meta title make blouse
12. Fashion	love fashion kids mom day fun style amp great party baby
13. Beauty Products	products skin review love hair product buy great
14. Family Entertainment	fun giveaway disney kids win movie review dvd family
15. Social Media	blog post blogging email twitter content follow posts bloggers
16. Couponing	exp printable oz price product final coupon ss amp free coupons
17. Couponing	free coupon coupons deals save baby deal shipping code online
18. Parties & Events	party twitter rsvp pm join win cravebox blogher congrats
C1-6. Common words	e.g., love day today great people good awesome
G1-7. Giveaways	e.g., giveaway review product enter win prize gift

7.3.2 H1: Platform Usage

Although the number of unique topics covered by mothers is almost the same in the blogosphere (median: 6.5) and on Twitter (median: 6.0), the ratio of entries mentioning a health topic

to the total number of entries has a median of 0.18 in the blogosphere and a mere 0.01 on Twitter (Table 7.2). Hence, mothers blog (and in fact, tweet) about a small number of health topics, found in about 20% of their blogs, but they tweet and/or retweet about a much broader range of issues, with health topics finding their way in only about 1% of their tweets. This may be explained in part by the fact that tweets require less time and thought investment [109].

An interesting aspect of Twitter is that it gives its users the ability to "retweet" messages from others, passing the original message along to their followers. Figure 7.1 shows the percentage of tweets for each directed health topic that come from original tweets (authored by the user themselves) compared to retweets. The topics with the lowest retweet percentage (Pregnancy, Illness, and Hospital) represent personal conditions or experiences that a mother may be sharing with her network. On the other hand, those topics that have the highest retweet percentage (Down Syndrome, Autism, FAS) represent topics that mothers may wish to promote, even if they are not personally experiencing the condition. This suggests that mothers may be willing to retweet for the purpose of increasing awareness of relevant health issues. In fact, while many mothers both author original tweets and retweet about a certain health topic, a large number of the mothers who mention these topics never author an original tweet about them. In other words, while many of the mothers mention our directed health topics (see Table 7.3), a significant proportion of them only do so via retweets, not by authoring their own tweets. For mothers who discuss health topics, Figure 7.2 shows the percentage who do so exclusively via retweet. Again, this percentage is largest for those health topics that are more likely to be awareness-driven than experience-driven (e.g., 64% for FAS, 56% for Down Syndrome, and 40% for Autism). These results begin to suggest that the nature of a health topic impacts mothers' use of Twitter, wherein authoring is more likely to be reserved for health issues experienced first-hand while retweeting may be used to increase awareness.

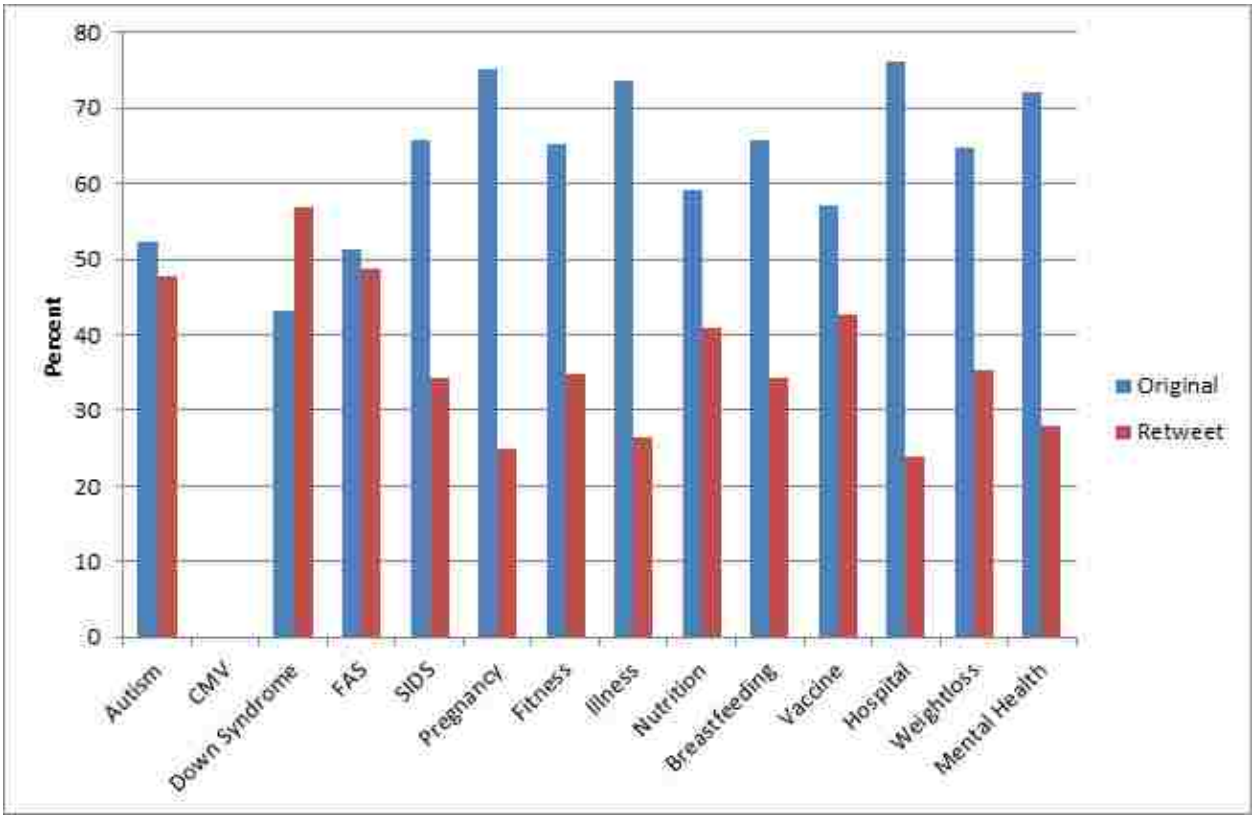

Figure 7.1: Percent of Health Topics from Original vs. Retweet

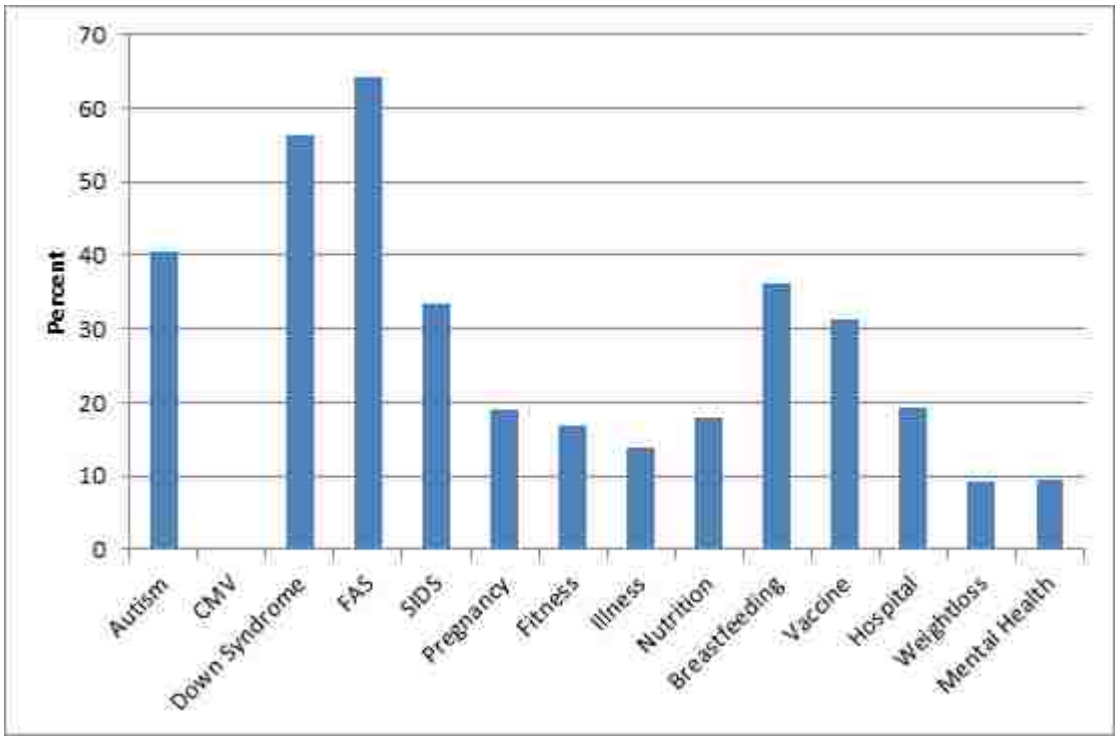

Figure 7.2: Percent of Mothers Mentioning Health Topics Solely via Retweet

We look a little deeper into the idea of authoring about first-hand experiences by comparing across platforms. When considering the relationship between health topics discussed in a mother's blog with those mentioned in her Twitter account, the health topics in her blog relate more strongly with the number of original health tweets as opposed to all tweets (including the retweeted ones). Figure 7.3 shows the likelihood of a mother mentioning a health topic (more often than the median) on her blog given that she mentioned it (more often than the median) on Twitter. It compares the likelihood of a blog post given any tweet on the topic as compared to the likelihood given an original tweet on the topic. As shown, across every topic, the probability of mentioning a specific topic on a blog increased when the mother authored an original tweet on that topic. This seems to add further evidence to the idea of "authenticity" of topics discussed, where mothers may be more likely to write a blog post about a topic if they are actually facing the condition (e.g., being pregnant, raising with an autistic child), as opposed to those that they merely support (e.g., as evidenced by retweeting) but may not personally experience. It is also possible, as noted earlier, that mothers are willing to support a topic via retweet (which takes little effort) whereas they may not treat the subject with a blog post (which requires more resources).

7.3.3 RQ2: Explicit/Implicit Consistency

In addition to topics, LDA also produces a document vector whose entries quantify how much the document, here a mother, relates to each topic. Using these vectors, we calculate the cosine similarity between each pair of mothers to produce an implicit affinity network [218]. To avoid finding relationships among users that had HTML artifacts, the topics composed of these artifacts were excluded from the analysis. The network in Figure 7.4 shows all of the mothers, with edges between them if the cosine similarity of their vectors exceeds 0.8. The clusters of users are labeled by the LDA topic (Table 7.4) that is most prevalent among them, with the largest group being a mixture of the common term topics (labeled "C").

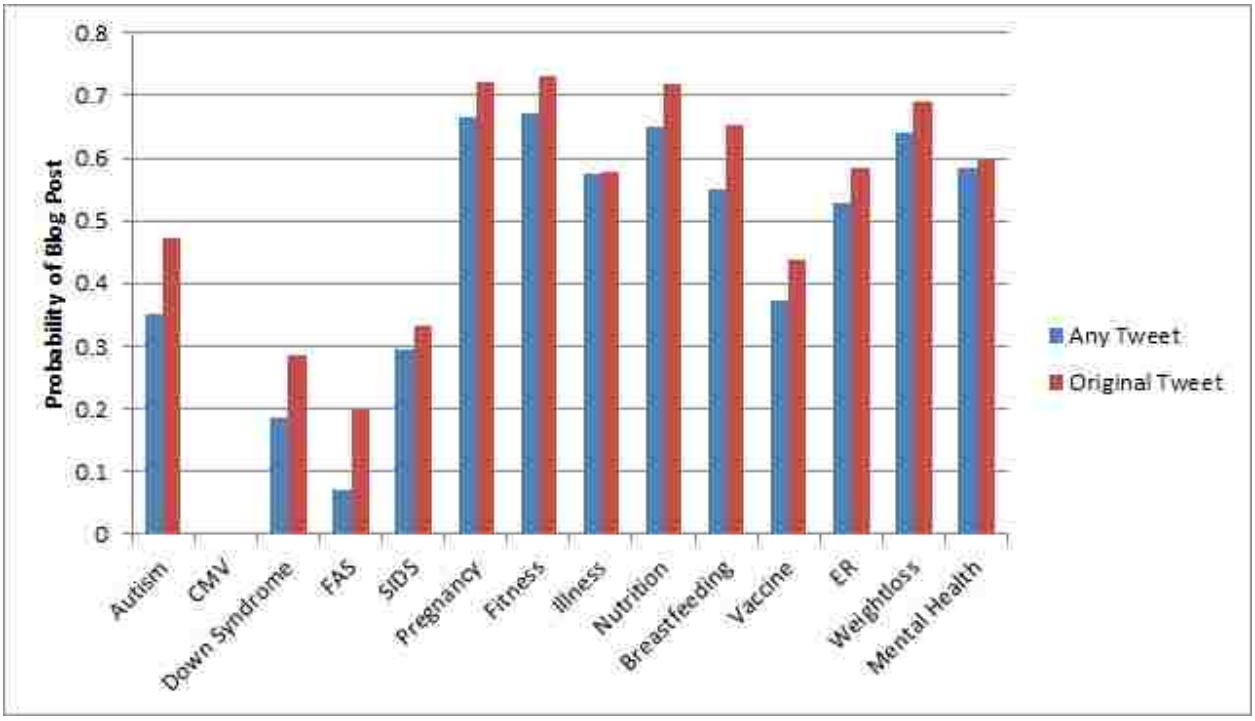

Figure 7.3: Probability of Blog Posts for Health Topics Given Tweets

Evaluating affinities, we found that 10.4% contained blog links, 16.7% mentioned the other on Twitter, and 41.1% followed the other on Twitter. In total, of all the strong implicit affinity relationships, 45.4% had some form of explicit link (i.e., blog link, mention, or follow). This demonstrates that many of these users are in fact connected to others of similar interest, and yet the fact that less than half of these strong affinities have an explicit relationship, suggests that there remain additional opportunities for connection among mothers of very similar interests who are otherwise unaware of each other. Such knowledge could be exploited to expand a mother's support network. For example, consider the case of cloth diapering. Although there is a resurgence of interest in cloth diapering, it is clearly not the norm and a newly practicing mother may face resistance from others around her, which may discourage her. Identifying others with a similar interest would allow that mother to create a network beyond her physical environment that may be sufficient to provide the support needed to maintain her commitment.

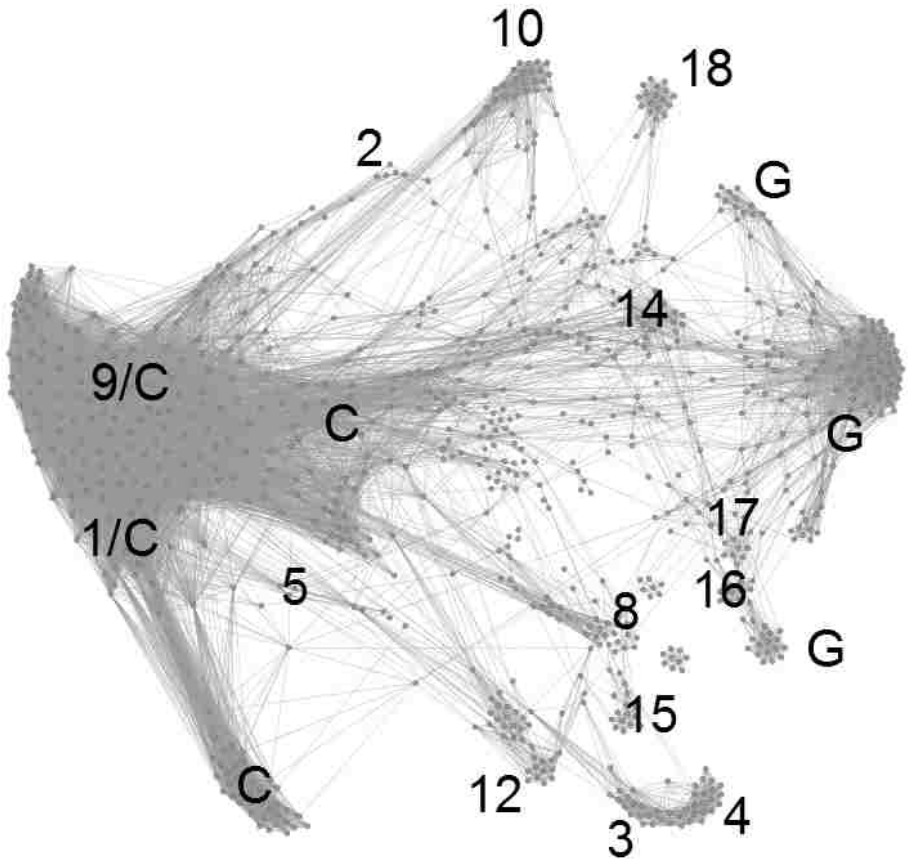

Figure 7.4: Implicit Affinity Network

7.3.4 H2: Linking Pattern Consistency

In order to test H2, we consider the likelihood of a mother linking to or following another in one platform given that they do in the other platform. Let T_{ij} denote the fact that i follows j on Twitter, M_{ij} denote the fact that i mentions j on Twitter, and B_{ij} denote the fact that i's blog links to j's blog. Our data then shows the following.

1. $P(M_{ij}|T_{ij}) = 0.18$ and $P(B_{ij}|T_{ij}) = 0.07$.

   If a mother follows another mother on Twitter, she will mention her only 18% of the time, and link to her blog account only 7% of the time. These relatively small numbers may not be surprising because we would not expect a mother to mention or link to everyone that she follows.

2. $P(T_{ij}|M_{ij}) = 0.87$ and $P(B_{ij}|M_{ij}) = 0.18$.

   If a mother mentions another mother on Twitter, she will also follow her 87% of the time, and link to her blog only 18% of the time. It is not surprising that almost all (87.0%) of the mothers that are mentioned come from the follower list, but the fact that the number of blog links is relatively low may indicate that explicitly linking in one platform does not always translate to explicitly linking in another.

3. $P(T_{ij}|B_{ij}) = 0.62$ and $P(M_{ij}|B_{ij}) = 0.32$.

   On the other hand then if a mother links to another mother's blog, she will follow her on Twitter 63% of the time, and mention her on Twitter 32% of the time. While these numbers are fairly high, given the level of awareness required to link to another mother's blog and alternatively, the relative easiness of following another on Twitter, it may be surprising that these mothers do not follow almost 100% of mothers on Twitter that they link to on their blog. Perhaps these mothers are not aware of the others' Twitter accounts.

These results show that there is no clear consistency of linking pattern across platforms, in spite of the fact that 1) many mothers' blogs push content to Twitter automatically, and 2) many mothers use Twitter to promote their blog. However, our results suggest that there may be something associated with the medium. For example, it is easier to follow someone on Twitter (one click) than to link to their blog (copying link, editing blog), and the endorsement provided by a fleeting retweet seems of much less consequence than one provided via a permanent link from one blog to the other.

7.4 Conclusion

We have identified mommy communities in Twitter and in the blogosphere, and found that health is a frequent topic of discussion within these communities, which is particularly relevant given the importance of mothers in the health behaviors of the family. Some important

health issues, however, are severely underrepresented (e.g., CMV). Given the natural patterns of health discussion among mothers, future work should examine ways of leveraging social networks for health promotion. While a similar number of unique health topics were discussed by mothers on each platform, individual blog posts were much more likely to contain a health topic than individual tweets. In terms of connections, many mothers do explicitly link to others of similar interests, and yet there are many more opportunities for explicit links between mothers of very similar interests. These "missing" links are opportunities for recommendations to mothers who may be seeking information or support. Finally, the methodology used here, including the construction and comparison of social circle structure and content across platforms, can be leveraged in other contexts.

Chapter 8

An Exploration of Social Circles and Prescription Drug Abuse through Twitter

Abstract

- **Background**: Prescription drug abuse (PDA) has become a major public health problem. Relationships and social context are important contributing factors. Social media provides online channels for people to build relationships that may influence attitudes and behaviors.
- **Objective**: To determine whether people who show signs of prescription drug abuse connect online with others who reinforce this behavior, and to observe the conversation and engagement of these networks with regard to prescription abuse.
- **Methods**: Twitter statuses mentioning prescription drugs were collected from November 2011 to November 2012. From this set, 25 Twitter users were selected that discussed topics indicative of prescription drug abuse. Social circles of 100 people were discovered around each of these Twitter users and the tweets of the Twitter users in these networks were collected and analyzed according to PDA discussion and interaction with other users about the topic.
- **Results**: 3,389,771 mentions of prescription drug terms were observed from November 2011 to November 2012. For the 25 social circles, on average 53.96% of the Twitter users used prescription drug terms at least once in their posts, and 37.76% mentioned another Twitter user by name in a post with a prescription drug term. Strong correlation was found between the kinds of drugs mentioned by the index user and his/her network

(mean $r = 0.73$) and between the amount of interaction about prescription drugs and a level of abusiveness shown by the network ($r = 0.85$, $P < 0.001$).

- **Conclusions**: Twitter users who discuss prescription drug abuse online are surrounded by others who also discuss it—potentially reinforcing a negative behavior and social norm.

8.1 Introduction

8.1.1 Prescription Drug Abuse

The 1992 to 2002 decade witnessed a striking surge in the manufacturing and distribution of prescription drugs. In particular, the number of opioid prescriptions increased by 222%, while the number of stimulant prescriptions increased by 368%. These dramatic increases in the prescribing and medical use of drugs have been deemed responsible for the subsequent increase in the misuse and abuse of these same drugs [36], to the point where prescription drug abuse (PDA) has reached epidemic proportions in the United States. Prescription drug overdose is now surpassing the combined number of people who overdosed during the crack cocaine epidemic of the 1980s and the black tar heroin epidemic of the 1970s, and is becoming the fastest-growing drug problem in the United States [257]. There were approximately 27,000 unintentional prescription drug overdose deaths in the United States in 2007 [39]. Results from the National Survey on Drug Use and Health indicate that almost one-third of individuals over the age of 12 who were first-time drug users in 2009 started with abusing a non-medical prescription drug [174]. In addition, it has been estimated that 48 million of Americans (approximately 20% of the population) age 12 and older have used prescription drugs for non-medical reasons at some point in their lifetime [172].

Even though death only occurs in the most severe cases of abuse, the negative health consequences of PDA are many, ranging from simple drowsiness and nausea to lack of coordination, disorientation, paranoia and seizures. A recent study also found that there

may be an emerging trend of (ab)using prescription drugs among adolescents to facilitate unwanted sexual contact [12]. A teen addiction treatment center in Iowa similarly warns against unwanted sexual behavior as one of the consequences of PDA [230]. While there does not seem to be evidence of more at-risk sexual behaviors, such as sex-for-drugs, since most people have easy access to prescription drugs either from friends and relatives or through "doctor shopping," this trend still raises concerns about the limited, yet real, danger of PDA increasing exposure to and spread of HIV.

The Office of National Drug Control Policy (ONDCP) Prescription Drug Abuse Prevention Plan includes four major areas of focus: education, monitoring, proper medication disposal, and enforcement [185]. Current public health intervention strategies are largely aimed at prescribers and distributors. In many states, doctors receive training on how to identify abusers and patients that doctor shop. In some states pharmacies and distributors are required to report the amount of controlled substances dispersed each week. While these measures have proven to reduce rates of overdose and overdose deaths, primary preventative measures among end users of prescription drugs have not been explored or implemented as widely. The inherent difficulty of identifying abusers and redirectors of prescription drugs fosters an easy environment for abuse without real threat of legal repercussion.

8.1.2 Social Networks and Social Media

Relationships embedded in one's social network are an important influencing factor and contributor to health behavior and outcome, even beyond individual attributes such as age, sex, education level, income and occupation [28, 48, 104, 145, 146, 238]. In the context of PDA, a recent study of the co-usage network of a population of 503 prescription drug abusers in rural Appalachian areas shows that daily OxyContin use is significantly associated with higher effective size of ego networks (a measure of social capital), and thus "speak to the importance of peer networks in determining social capital and social norms, which has vast implications for intervention research" [112]. It has been found that people, including

youth, often learn to abuse prescription drugs by observing a family member, or other members of their social network, model the abuse of prescription drugs [52, 141]. Within families, the practice of "friendly sharing" of prescription drugs has become commonplace [36]. Recent research has also identified social groups or informal economic markets where drug transactions can occur. An established market for prescription drug distribution has been identified in junior high and high school classes. Among students in Nova Scotia who had been prescribed stimulants, about 22% reported giving away or selling their medications, while another 7.3% experienced theft or were forced into giving away their prescriptions [200].

Research has revealed that the Internet provides ready access to drugs—including prescription medications [75, 168]. More recently evidence also suggests that participation in social media sites may increase one's risk of substance abuse, especially among adolescents. The National Center on Addiction and Substance Abuse began collecting data to explore the influence of social networking and substance abuse in 2011. Their findings reveal that teens who spend time social networking online are five times more likely to use tobacco, three times more likely to use alcohol, and two times more likely to use marijuana [169].

While studies have demonstrated the influence of social relationships on PDA in the real world as well as ready access to the drugs, little is known about these influences in cyberspace. Social media applications, such as Twitter, provide a way to observe the conversations of individuals and their social circles directly, providing a mechanism to monitor end users of prescription drugs. By monitoring individual conversations, studies have demonstrated the validity of identifying health topics in Twitter [192, 201], including prescription drug misuse [93]. Research has also demonstrated a correlation between online discussion and real-world rates [108]. In addition, social media applications are platforms for networking and as such are rich with relationships. These relationships make up important social circles that have the capacity to influence behavior due to unique norms and values of the group. Indeed, no social media user is an island, and the *social* element of social media has particular relevance in public health research. In this regard, this work extends previous health research

in social media by not only analyzing the content of social media posts, but also relationships among users. Specifically, we discover the online social circles of prescription drug abusers and analyze the discussion and interaction of these networks. Few studies have explored the influence of online relationships on alcohol and other drug use [53, 224]. To the best of our knowledge, this is the first work to focus on the relational component of these networks through social media, with regard to PDA.

The purpose of this study was to explore the social circles of prescription drug abusers on Twitter to observe the discussion and engagement of these users regarding PDA. To fulfill the purpose of this study, the following hypotheses were explored:

- **Hypothesis 1**: People discuss PDA on Twitter.
- **Hypothesis 2**: People who discuss PDA on Twitter belong to social circles that engage with each other about PDA.
- **Hypothesis 3**: Social engagement about PDA varies across social circles of those who discuss it, and higher engagement correlates with higher levels of abuse.

8.2 Methods

A distinction exists between PDA and prescription drug misuse. The former refers to using a drug with the intent of deriving some side-effect, usually of a euphoric nature (i.e., getting high). The latter refers to increasing dosage in an attempt to improve the drug efficacy or to sharing the drug with someone whose symptoms may call for it but to whom the drug has not been prescribed. Either way, one can easily argue that "no matter the intention of the person, ... taking a drug other than the way it is prescribed can lead to dangerous outcomes that the person may not anticipate" [125, p. 1]. Hence, throughout the paper, any improper use and user are referred to simply as abuse and abuser, respectively.

To evaluate the discussion of PDA among social media users, Twitter users mentioning prescription drugs were identified, and their tweets as well as those of their network were analyzed.

8.2.1 Study Setting

Social media applications such as Twitter provide channels for social networking with others who may have similar interests and needs. Twitter provides users with a platform to share short messages ("tweets") among themselves. Twitter users can "follow" others, to subscribe to a feed of tweets from users of interest; they can also broadcast their messages to all of their followers or direct messages at specific users ("mentions"). By default tweets are public; hence, it is generally possible for a user X to see the tweets of a user Y even though X may not be following Y or Y did not mention X explicitly. Because Twitter users tend to post messages as events occur in their lives, tweets are an ideal source for researchers to observe natural, and timely, interactions among people. As such, Twitter was used to observe discussion and engagement with regards to PDA.

This study was approved by the university's internal review board.

8.2.2 Identifying Users and Networks

Twitter provides an application programming interface (API) that enables programmatic consumption of the content and the relationships of its tweets and users. The Twitter Streaming API provides means of obtaining tweets as they occur, filtered by specific criteria, such as a list of keywords. The Twitter API also enables discovering people following and followed by a given user, as well as retrieving up to 3200 of a user's most recent tweets.

To identify a set of tweets mentioning prescription drugs, the Twitter stream was filtered prescription drug terms, producing a set of all tweets mentioning these terms from November 29, 2011 through November 14, 2012. From this set, potential prescription drug abusers were identified for analysis along with their networks. In order to select those Twitter

users who had some discussion of prescription drugs, but that were still regular users (i.e., excluding accounts devoted to online drugs sales, automated feeds, etc.), Twitter users that mentioned prescription drugs in at least 10 tweets but less than 100 were selected at random. These users were then manually evaluated to determine 25 index Twitter users that were most likely prescription drug abusers based on a pattern of prescription drug tweets that matched one or more of the categories of abuse. Users were rejected if their prescription drug tweets did not match any of the categories of abuse. Likely prescription drug abusers were more likely to have tweets that matched the categories of abuse. For example, one of the 25 index users was selected because they had a pattern of tweeting about Adderall and Xanax (45 and 34 tweets respectively) and 26 of those tweets matched several of the abuse categories. Most alarming was that 11 of the abuse tweets were about coingestion. One of these coingestion tweets stated, "Adderall + Benadryl has put me in a weird awake/tired haze. Relatively certain that I'm saying things i wont [sic] remember in the morning."

The social circles of each of the 25 index Twitter users were discovered. Unlike a traditional ego network that consists of all the individuals ego has a direct connection to, a social circle is a densely connected set of mutually-aware individuals that surround ego, where some may be included in the circle by virtue of their many connections to ego's alters. Social circles capture the intuition that someone who influences ego's alters may exert a stronger influence on ego, though indirectly, than some of ego's alters. Finding a social circle around one or a small group of individuals is an instance of the community search problem [221], a query-based version of the traditional community mining problem [77]. In the context of Twitter, however, there are two additional constraints: 1) the Twitter graph cannot be feasibly known, and 2) the "follow" relation in Twitter is directed. As a result, a local social circle discovery algorithm designed specifically for directed graphs must be used [33]. Intuitively, the algorithm initializes the social circle with the index Twitter user and then iteratively adds new members to the social circle until a prespecified size has been reached. At each step, the algorithm considers all Twitter users followed by at least one

member of the current social circle. For each of these, it finds the number of members of the social circle it follows and the number it is followed by, and chooses the minimum as its score. The Twitter user with the largest score is added to the current social circle and the process repeats. To make sure newer members do not make the social circle drift away from the initial Twitter user, scores are discounted at each step of the algorithm. Furthermore, to increase the cohesiveness of the social circle, every 5 iterations, the Twitter user with the lowest score is removed from the social circle. Upon completion, the algorithm returns a social circle composed of dense connections of mutually aware nodes that surround the index Twitter user. Note that in general individuals belong to different social circles that may best be specified by including additional people in the query set (e.g., work colleagues would likely produce a professional social circle, relatives would likely produce a family social circle). Here, however, the index Twitter user is used as the sole query node to avoid biasing the algorithm toward any specific social circle, and instead simply discovering the most natural dense set surrounding that individual.

The size of the social circles was set to 100. After a social circle was identified for each index Twitter user, the most recent tweets of each Twitter user in the social circle (up to 3200 per user, the maximum allowed by the Twitter API) were obtained for content analysis.

8.2.3 Content Categorization

Once a social circle and its corresponding tweets were obtained, tweets were categorized by mention of a particular substance, and further categorized by the manner in which that substance was mentioned. Table 8.1[1] lists the drug categories and the filter terms used to categorize the tweet. For example, a tweet was categorized as mentioning painkillers if it contained terms such as "painkiller," "oxycontin," or "lortab."

[1]The "*" matches 0 or more characters.

Table 8.1: Keywords for Prescription Drugs

Drugs	Keywords
Adderall	adderall
Xanax	xanax
Klonopin	klonopin
Valium	valium; sleeping pills
Painkillers	painkiller*; pain killer*; narcotic painkiller*; oxycontin; vicodin; percodan; percocet; darvon; lortab; lorcet; dilaudid; demerol; lomotil; kadian; avinza; codeine; duragesic; methadone
Depressants	mebaral; nembutal; sodium pentobarbital; halcion; prosam; ativan; librium; depressant*
Stimulants	dexedrine; ritalin; concerta; amphetamines; stimulant*

Tweets matching the drugs in Table 8.1 were further categorized into eight different types of abusive or risk behaviors defined in Table 8.2[2]: taking larger doses (overdose), co-ingestion, taking more frequent doses, alternative motives (dependence or need the drug due to addiction), alternative routes of admission, legitimacy of obtaining, redistributing (trading/selling), and seeking [98].

Tweets that matched the drugs in Table 8.1 were further analyzed to determine if they also contained mentions to other Twitter users (where an author references another user by the @*username* convention). Social network graphs were then constructed to show such connections among Twitter users. The graphs are directed and weighted. The weight of an edge is defined by the number of tweets from one user to another that included prescription drug terms.

8.3 Results

The tweets collected during the study period contained 3,389,771 references to prescription drug terms. Table 8.3 shows the number of co-occurrences of these references with one of the categories defined by the terms in Table 8.2. The large number of references to alternative motives was due primarily to discussion of Valium as a sleep aid.

[2]Coingestion keywords for xanax and adderall did not include the keywords "xanax" and "adderall" respectively. For alternative motives, the keywords "test", "final", "study", and "studying" were exclusively used as keywords for adderall; "Skinny" was exclusive to Stimulants.

Table 8.2: Keywords for Risk/Abusive Behaviors

Risk/Abusive Behaviors	Keywords
Larger Doses/Overdose	too many; two; three; double; too much; overdose; crash; strong enough; max; too many
Coingestion	alcohol; coffee; white; red; wine; vodka; shots; patron; booze; margarita; mimosa; xanax; painkiller; caffeine; alcohol; happy pills ; adderall; concerta; cocaine; rum
More Frequent doses	enough; pop; popping; not enough; another; enough; pop*
Alternative motives/dependance	test; final; study; studying; problems; college; class; breakfast; rely; sleep; sleeping; work; family problems; sleep*; stress*; stressful; stress; skinny
Alternative routes of admission	snort; crush; inject; snort; inhale
Legitimacy of Obtaining	steal*
Trading/selling	buy; sell; trade; share; spend; buy; bring
Seeking	need; want; needing; wanting; wish; need

Table 8.3: Categorization of all Tweets

Category	Adderall	Xanax	Klonopin	Valium	Painkillers	Depressants	Stimulants	Total
Drug Total	412,314	486,670	58,527	917,805	1,215,574	17,364	281,517	3,389,771
Larger Doses/Overdose	11,397	9,508	880	22,263	28,186	218	2,085	74,537
Coingestion	44,179	24,794	5,411	47,657	34,178	1,027	3,181	160,427
More Frequent doses	10,636	18,070	567	15,808	22,764	107	2,566	70,518
Alternative motives/dependance	39,459	18,664	105	617,672	38,135	806	1,868	716,709
Alternative routes of admission	1,316	1,657	73	701	1,641	17	265	5,670
Legitimacy of Obtaining	363	400	16	339	1,032	6	117	2,273
Trading/selling	20,941	63,763	17,000	65,926	95,962	4,913	2,873	271,378
Seeking	46,138	52,852	2,069	165,955	63,165	675	8,808	339,662

The 25 social circles discovered around the 25 index Twitter users gave rise to a total of 2,227 unique Twitter users, 7,290 prescription drug tweets, and 2,788 directed prescription drug tweets. Statistics of these social circles are shown in Table 8.4[3]. As shown, the social circles range from 14% to 87% (mean: 53.96%, median: 61%) of the Twitter users in the social circle tweeting about prescription drugs at least once.

In addition to simply talking about prescription drugs, Twitter users in these social circles also interact with each other about the topic, using the @*username* convention. Examples of mention tweets from the sample include, "@*** Haha! For me it's a nice ritalin/sangria combo :)", "RT @*** I should win a lifetime achievement award...I've been taking Xanax for years without overdosing.", and "@*** lol thanks....but im [sic] pretty emotionally stable. It's called being in a Xanax haze." As shown in Table 8.4, the networks range from 9% to 84% (mean: 37.76%, median: 34%) of the Twitter users in the social circle interacting with another Twitter user about prescription drugs at least once.

Index users and their social circles typically tweeted about similar drugs. For each index Twitter user a topic vector was determined according to the proportion of their prescription drug tweets that matched each of our prescription drug categories, and a topic vector was also created for the aggregated tweets of the rest of the social circle. The topic vectors of index Twitter users were correlated with those of their social circle, and Pearson's correlation coefficients ranged from -0.14 to 0.99. (mean: 0.72, median: 0.72). The mean of these correlation coefficients was computed by first applying Fisher's Z transformation.

Using the abusive behaviors content categories of Table 8.2, each of the tweets of the index Twitter users and their social circles were categorized according to potential abuse. Although not a perfect metric for abuse, the number of abuse categories a Twitter user mentions is used as surrogate for a level of abuse. Thus, a Twitter user who has tweets matching 4 of the abuse categories is considered to be at a higher level than a Twitter user who only has tweets from 1 of them. As shown in Table 8.4, on average 33.24% of the people

[3]The mean of topic correlation coefficients was computed using Fisher's Z transformation.

Table 8.4: Statistics of Prescription Drug Social Circles

Network	Prescription Drug Tweets, No.	Prescription Drug Tweet Mentions, No.	Percentage tweeting prescription drugs, %	Percent interacting about prescription drugs, %	Topic Correlation, Correlation Coefficient	Percent with 1 or more abuse categories, %	Percent with 2 or more abuse categories, %
1	136	55	48	32	0.28	25	9
2	99	22	28	12	0.26	13	1
3	67	26	14	11	0.06	8	2
4	508	290	84	72	0.59	38	18
5	352	97	46	34	0.69	34	22
6	258	37	72	29	0.92	27	12
7	311	40	69	27	0.76	39	18
8	52	14	17	9	0.10	8	6
9	553	142	61	40	0.83	33	18
10	359	156	76	51	0.89	58	21
11	159	73	32	26	0.72	18	11
12	449	300	77	71	-0.14	36	18
13	446	302	87	84	0.74	73	39
14	378	112	79	42	0.65	55	30
15	629	140	61	42	0.99	34	21
16	75	36	31	23	0.82	28	11
17	512	244	84	64	0.93	58	33
18	91	35	25	20	0.89	9	3
19	75	28	30	17	0.37	17	8
20	75	24	20	16	0.77	10	5
21	143	80	46	36	0.30	25	11
22	512	91	79	48	0.86	54	35
23	417	142	69	47	0.60	52	28
24	387	249	83	70	0.97	60	30
25	247	53	31	21	0.61	19	10
Mean	291.60	111.52	53.96	37.76	0.73	33.24	16.80
Median	311	80	61	34	0.73	33	18

in the social circle had tweets matching at least one abuse category, and 16.80% had tweets matching at least two. The level of abuse is strongly correlated with the number of Twitter users interacting with others about prescription drugs. Comparing the percentage of the social circle that interacts about prescription drugs to the percentage that matched at least one abuse category yields a Pearson's correlation coefficient of $r = 0.85$ ($P < 0.001$), and comparing against those who matched two or more abuse categories, $r = 0.81$ ($P < 0.001$).

In addition to the quantitative evaluation of these interactions, interesting patterns can also be observed through visual inspection of the graphs of interactions among Twitter users in each social circle. Figure 8.1 shows three graphs, where the nodes represent users, and the edges indicate that the source user mentioned the destination user along with a prescription drug term. The weight of the edges (as shown by the thickness of the line) denotes the number of mentions. The size of the nodes represents the number of prescription drug tweets.

8.4 Discussion

The purpose of this study was to explore the online social circles of prescription drug abusers to observe the discussion and engagement of these Twitter users regarding PDA. Accordingly, the study was guided by three research hypotheses. As shown in Table 8.3, significant discussion of PDA was observed on Twitter (hypothesis 1). These findings are consistent with previous research exploring PDA through Twitter [93]. While not all of these tweets are necessarily in reference to abuse, those matching the abuse categories defined in Table 8.2, are very likely to be discussion of abuse of prescribed substances. Even if not all of these references denote actual behavior on the part of individuals, the simple act of discussing the behavior within a social circle can impact the social norms of those within that circle.

Those who are not engaged in PDA are still being exposed to others' tweets concerning the matter. They may not be participating in the conversation, but they are observing the sentiment and potentially forming ideas and norms about the abuse of prescription drugs.

Figure 8.1: Prescription Drug Interaction Graphs

While actual drug abuse remains mostly a private affair, it seems to be discussed in a very open manner online for all to observe. It may be that abusers are now, through social media, finding support for their abuse and feel a sense of safety in opening up to others. Uses and gratification theory suggests that individuals make decisions about their media choice based on the extent to which that media gratifies a communication need [117]. Duffy and Thorson [63] expand this idea in their Health Communication Media Choice Model and suggest that connectivity is an important need that can be fulfilled through social media. They define connectivity as the "need to relate, support, engage with and communicate with others face-to-face through media" [63, p. 102]. Social media facilitates the connectivity process by allowing people to engage with and observe others' sentiment on a given subject. Regardless of a person's openness about their behavior, prescription drugs are being discussed on Twitter and many are being exposed to tweets and conversations of an abusive nature through their social circles.

As shown in Table 8.4, there is a significant amount of discussion about prescription drugs in the social circles of the index Twitter users, with an average of 54.0% of the social circles posting about a topic at least once, and an average of 247 tweets per social circle (hypothesis 2). In addition, the high correlation between the substances discussed by the index Twitter user and their social network, shows that these users are engaged in discussions with others of like-minds. These findings confirm our hypothesis and also show consistency with the offline world about the social context of PDA [22, 52, 141, 200].

It is not clear whether index Twitter users developed their behavior from exposure to their online social circle, or whether they sought out the company of others supportive of their viewpoints. But it is clear that each of these Twitter users is in an environment that potentially supports their behavior. This may have interesting ramifications, because these users may not be in close proximity to one another physically, and yet they may find reinforcement for their attitudes from their online connections. Thus, while a prescription drug abusers may not feel comfortable sharing their experiences with their physical neighbors,

who might not approve of abusive behavior, they can develop online associations with those that do. These findings are consistent with recent research exploring the impact of online social circles on young adult alcohol use [53, 224].

In addition to knowing that Twitter users are talking about prescription drugs, it is also relevant to discover if they are also talking to each other about prescription drugs (hypothesis 3). When Twitter users mention one another by their username (using the @*username* convention), these tweets are aggregated into a separate list in the interface, and can also produce other alerts (e.g., email) raising the user's level of awareness of the tweet. In addition, the author may be directly soliciting a response from the user. Thus, the analysis of the number of tweets that discuss prescription drugs and also mention a specific user provides a quantified measure to observe engagement among these users about the topic.

The fact that on average 37.76% of the Twitter users in a social circle interact with another user at least once shows that there is indeed a significant level of engagement in addition to simply talking about the topic. Furthermore, hypothesis 3 is confirmed by the fact that the percentage of social circles interacting about prescription drugs correlates so strongly with the percentage of social circles having tweets that match risk/abusive behavior categories (r=0.85 for one category and r=0.81 for two categories). Social engagement can also be observed through the interaction graphs shown in Figure 8.1. It is interesting to observe how some users that discuss prescription drugs relatively frequently (as denoted by the larger size of the node), in many cases also have a large in-degree, showing that many others mention them in connection with prescription drugs.

With the rise of PDA and its inherent danger, understanding the behaviors of abusers will be vital for public health professionals and prescribers in preventing overdose deaths and the blatant redirecting of the drugs. Many states are implementing prescription drug registries in response to the epidemic of abuse. These registries require prescribers and providers to report the distribution of controlled substances. While these registries can identify patients going to multiple doctors for the same medication they do not address the growing problem

of prescription drug redirection. This drug aftermarket is only facilitated by social media platforms like Twitter. The categorization keywords used in this study were able to identify users seeking, trading, and buying prescription drugs. For example several seeking statements included, "Seriously. Need adderall. Will pay $$$. Help me." and "looking to buy 20-40 mg adderall, email ***". While a drug registry may identify and limit an abuser in one state, that abuser can simply source drugs online from others in states where drug registries are not used and abusers are able to obtain excessive amounts of a drug. Another key risk behavior drug registries cannot address is that of co-ingestion and non-medical use. Co-ingestion is one of the deadliest drug abuse behaviors and a leading cause of overdose death.

Findings from this study have important implications for those professionals involved in prevention and treatment of PDA. Results indicate that Twitter is used as a platform for discussion about PDA within social circles. As such, Twitter provides an additional "access point" to groups of individuals who are abusing prescription drugs. Innovative approaches to reaching these social circles might include online peer health advisors who have been trained to identify PDA and appropriately intervene.

8.5 Limitations

Results from this study should be interpreted in light of the following limitations. First, while a keyword-based approach for identifying and categorizing tweets may exclude misspellings of the term, it does result in a highly precise set for analysis, and at a minimum provides a lower bound for the amount of discussion. Second, through social media it is only possible to observe discussion, not actual behavior. Yet, as these are natural conversations among friends where people post about events that occur in their lives, there is no *a priori* reason to believe that on the whole people are falsifying their posts to portray events or behaviors that do not occur. Third, we may have underestimated the number of PDA tweets. It is possible that there are other PDA-related tweets we missed because they were not covered by our keywords. It is also possible that not all tweets were delivered to us by the Twitter interface,

although that is hard to know for sure. Lastly, tweets containing abuse-related keywords may not always refer to discussion of abuse. Despite these limitations, it is likely that the general trends observed would not be affected.

8.6 Conclusions

Understanding prevalence of a problem or issue through social media is a good place to start; however, prevalence data fails to take advantage of the key aspect of social media: social networks and relationships. This work extends previous work by examining the social context of those discussing an important public health topic. While a major focus of this work has been about the reinforcement of negative behavior, the analysis of the interactions between people can provide insights into the normative aspects of social media. Whereas Twitter is a social media platform used to discuss and reinforce PDA, prevention specialists should be mindful of this communication channel as another setting for understanding and monitoring PDA and potentially intervening online.

Part IV

Conclusion

In this dissertation we have demonstrated computational techniques for mining the *who*, *what*, and *where* of public health surveillance. In doing so, we have made several contributions including:

- developing an experimental framework for use in future research;
- introducing a new algorithm to discover social circles surrounding an initial set of query nodes in directed graphs in which the entire network cannot be known *a priori*;
- demonstrating the use of computational methods to mine health data from social networks at a larger scale than previously available.
- validating location information for research in social media;
- establishing the relevance of social media data for public health research, in that users do, in fact, publicly discuss health topics that might be considered private and that many theoretical results also hold in online settings.

The work in this dissertation is a critical step to understanding, and ultimately influencing, behavior through social media. Building on this work, future studies should further examine the peer-to-peer sharing of health advice, and explore ways to promote positive information flow through this lay health advisor model. For example, one of the elements of suicide prevention successfully employed by the Hope4Utah foundation is to

identify and train a peer-based "Hope Squad," where high school students help their peers that are at risk for suicide. As adolescents continue to use mobile devices and social media at ever increasing rates, these Hope Squads could be empowered with tools to observe and interact with their friends in positive ways. Furthermore, outside of a controlled environment, future work could identify ways to promote other positive horizontal communication in social media, where common users (not just health providers) offer accurate and helpful health advice.

In addition to our work in identifying risk factors in social media, more research is needed to identify pertinent health events in social media (e.g., a loved one that is diagnosed with cancer or commits suicide). This is a difficult challenge because often these events are mentioned very rarely in comparison to the commonly discussed topics of day to day events, making it difficult for them to be discovered by techniques that rely on frequent data to discover patterns. These events often produce significant changes to a user's life, and discussion before them may be different than after. Thus, an important area for future work is to incorporate the timeline of the data, rather than treating everything as if it happened at once. For example, identifying the changes between positive and negative emotions could be helpful in identifying cycles of abuse that would not be found through mining all the text together.

In addition to the content a user produces, understanding and leveraging the networks around them is one the most important areas for future research in public health surveillance. This work could leverage our approach to identifying social circles not only to assist in identifying at-risk individuals, but also to identify influential friends to assist in intervention. Future work is needed to better understand the dynamics of the network support structure (in supporting both positive and negative behaviors), and how to affect it, in order to promote lasting behavior change. For example, how can relationships be identified, established, and maintained between individuals with common goals, and to what extent does a commitment in social media affect offline behaviors?

As users continue to produce increasing amounts of multi-media content (e.g., Instagram, YouTube), future work should explore ways to incorporate this non-textual data into public health research. This may include image processing and computer vision methods, but could also involve taking advantage of more readily available meta-data such as the time or location of an image, which may already be included. To leverage these opportunities, it will also be critical to resolve users across various social media platforms. We have taken a first step in record linkage across social media platforms, and future work could pursue additional ways to match these accounts by incorporating other features such as common friends, similar locations, and usernames.

The role of health in social media also has ramifications for health promotion groups, as mentioned in Chapter 7. We have discussed the implications for health promotion groups attempting to leverage social media to engage their audience [175, 233], but there remains significant work to determine *how* these groups can best produce this engagement. Increasing promotion clearly requires deviating from applying historical methods to new channels to include innovative approaches, to leveraging new technologies. By further identifying latent user attributes, custom tailored information could be developed to provide individual users with what they specifically need.

Finally, it should be remembered that social media is only one facet of an individual's life, and there are many other sources that record, inform, and influence their personal health. Future work in quantified self measurements (e.g., applications that track activity and diet), electronic health records, and citizen sensing (where users help to collect data about their surroundings) can all be brought together in a way to help decode the human exposome[247], including everything that surrounds and influences an individual.

In short, computational health science is an exciting, emerging area of research, with many opportunities for future work, and the work we are doing has even garnered the attention

of the news media (including the Deseret News[45] and the Washington Post[6]). As the domains of health and technology continue to intersect in broadening ways, it will become increasingly important for computer and health scientists to collaborate to exploit the possibilities of their data and solve the computational problems that arise at the intersection of these fields.

[4] http://www.deseretnews.com/article/865571507/Research-finds-Twitter-useful-in-tracking-epidemics.html

[5] http://www.deseretnews.com/article/865579375/BYU-researchers-track-Adderall-abuse-via-Twitter.html

[6] http://www.washingtonpost.com/national/health-science/twitter-becomes-a-tool-for-tracking-flu-epidemics-and-other-public-health-issues/2013/03/04/9d4315c2-6eef-11e2-aa58-243de81040ba_story.html

References

[1] J. G. Adair. The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69(2):334, 1984.

[2] C. D. Advokat, D. Guidry, and L. Martino. Licit and illicit use of medications for attention-deficit hyperactivity disorder in undergraduate college students. *Journal of American College Health*, 56(6):601–606, 2008.

[3] A. Ahluwalia, A. Huang, R. Bandari, and V. Roychowdhury. An automated multiscale map of conversations: Mothers and matters. In *Proceedings of 4th International Conference on Social Informatics*, pages 15–28, 2012.

[4] Y-Y. Ahn, J. P. Bagrow, and S. Lehman. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, 2010.

[5] I. Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, 1991.

[6] I. Ajzen and M. Fishbein. *Understanding Attitudes and Predicting Social Behavior*. Prentice-Hall, Inc., 1980.

[7] W. Almansoori, S. Gao, T. Jarada, A. Elsheikh, A. Murshed, J. Jida, R. Alhajj, and J. Rokne. Link prediction and classification in social networks and its application in healthcare and systems biology. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 1:27–36, 2012.

[8] American Cancer Society. Breast cancer key statistics. http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics, 2013. Accessed: March 2013.

[9] American Foundation for Suicide Prevention. Risk factors for suicide. http://www.afsp.org/index.cfm?page_id=05147440-E24E-E376-BDF4BF8BA6444E76, 2012. Accessed: 2 September 2012.

[10] American Foundation for Suicide Prevention. Warning signs of suicide. http://www.afsp.org/index.cfm?fuseaction=home.viewPage&page_id=0519EC1A-D73A-8D90-7D2E9E2456182D66, 2012. Accessed: 2 September 2012.

[11] R. Andersen and K. J. Lang. Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web*, pages 223–232, 2006.

[12] T. R. Apodaca and N. C. Moser. The use and abuse of prescription medication to facilitate or enhance sexual behavior among adolescents. *Clinical Pharmacology & Therapeutics*, 89(1):22–24, 2011.

[13] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP2011)*, pages 1568–1576, 2011.

[14] M. Arrington. Odeo releases twttr. http://techcrunch.com/2006/07/15/is-twttr-interesting/, 2006. Accessed: 2 September 2012.

[15] S. Asur and B. A. Huberman. Predicting the future with social media. http://arxiv.org/abs/1003.5699, 2010. Accessed: 2 September 2012.

[16] Q. Babcock and T. Byrne. Student perceptions of methylphenidate abuse at a public liberal arts college. *Journal of American College Health*, 49(3):143–145, 2000.

[17] C. L. Backinger, A. M. Pilsner, E. M. Augustson, A. Frydl, T. Phillips, and J. Rowden. YouTube as a source of quitting smoking information. *Tobacco Control*, 20(2):119–122, 2011.

[18] L. Backstrom, E. Sun, and Marlow C. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International World Wide Web Conference*, pages 61–70, 2010.

[19] J. P. Bagrow and E. M. Bollt. A local method for detecting communities. *Physical Review E*, 72(4):046108, 2005.

[20] A. Bandura and D. C. McClelland. *Social learning theory*. Prentice-Hall Englewood Cliffs, NJ, 1977.

[21] J. Baumes, M. Goldberg, and M. Magdon-Ismail. Efficient identification of overlapping communities. In *Proceedings of the International IEEE International Conference on Intelligence and Security Informatics*, pages 27–36, 2005.

[22] N. Bavarian, B. R. Flay, P. L. Ketcham, and E. Smit. Illicit use of prescription stimulants in a college student sample: A theory-guided analysis. *Drug and Alcohol Dependence*, 2013. (in press).

[23] F. Benevenuto, F. Duarte, T. Rodrigues, V. A. F. Almeida, J. M. Almeida, and K. W. Ross. Understanding video interactions in YouTube. In *Proceeding of the 16th ACM International Conference on Multimedia*, pages 761–764, 2008.

[24] S. Bennett. Twitter now seeing 400 million tweets per day, increased mobile ad revenue, says ceo. http://www.mediabistro.com/alltwitter/twitter-400-million-tweets_b23744, 2012. Accessed: 2 September 2012.

[25] J. W. Berry, B. Hendrickson, R. A. LaViolette, and C. A. Phillips. Tolerating the community detection resolution limit with edge weighting. *Physical Review E*, 83: 056119, May 2011.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[27] L. Bortolussi and A. Policriti. Studying cancer-cell populations by programmable models of networks. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, pages 1–17, 2012.

[28] K. Browne-Yung, A. Ziersch, and Baum F. "Faking it til you make it": Social capital accumulation of individuals on low incomes living in contrasting socio-economic neighbourhoods and its implications for health and wellbeing. *Social Science & Medicine*, 85:9–17, 2013.

[29] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff. Digital disease detection – harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21): 2153–2157, 2009.

[30] J. S. Brownstein, C. C. Freifeld, E. H. Chan, M. Keller, A. L. Sonricker, S. R. Mekaru, and D. L. Buckeridge. Information technology and global surveillance of cases of 2009 H1N1 influenza. *New England Journal of Medicine*, 362(18):1731–1735, 2010.

[31] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Medicine*, 5(7):e151, 07 2008.

[32] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, 13(3):e55, 2011.

[33] S. Burton and C. Giraud-Carrier. Discovering social circles in directed graphs. *In Submission*, 2013.

[34] S. Burton, R. Morris, J. Hansen, M. Dimond, C. Giraud-Carrier, J. West, C. Hanson, and M. Barnes. Public health community mining in YouTube. In *Proceedings of the Second ACM International Health Informatics Symposium (IHI 2012)*, pages 81–90, 2012.

[35] S. H. Burton, K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes. "Right time, right place health" health communication on Twitter: Value and accuracy of location information. *Journal of Medical Internet Research*, 14(6):e156, 2012.

[36] J. A. Califano Jr, L. C. Bollinger, C. Bush, J. L. Curtis, J. Dimon, and P. R. Dolan. Under the counter: The diversion and abuse of controlled prescription drugs in the US. New York, NY: National Center on Addiction and Substance Abuse, Columbia University. Available Online: http://www.casacolumbia.org/articlefiles/380-Under%20the%20Counter%20-%20Diversion.pdf, 2005. Accessed: 22 May 2013.

[37] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori. Detecting communities in large networks. *Physica A: Statistical Mechanics and its Applications*, 352(2-4):669 – 676, 2005. ISSN 0378-4371.

[38] H. A. Carneiro and E. Mylonakis. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564, 2009.

[39] Centers for Disease Control and Prevention. CDC grand rounds: Prescription drug overdoses — a U.S. epidemic. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6101a3.htm, 2012. Accessed: 25 July 2012.

[40] Centers for Disease Control and Prevention. Suicide: Risk and protective factors. http://www.cdc.gov/ViolencePrevention/suicide/riskprotectivefactors.html, 2012. Accessed: 2 September 2012.

[41] Centers for Disease Control and Prevention. Suicide facts at a glance. http://www.cdc.gov/violenceprevention/pdf/suicide-datasheet-a.PDF, 2013. Accessed: April 2013.

[42] J. Chen, O. R. Zaïane, and R. Goebel. Detecting communities in large networks by iterative local expansion. In *Proceedings of the International Conference on Computational Aspects of Social Networks*, pages 105–112, 2009.

[43] J. Chen, O. R. Zaïane, and R. Goebel. Detecting communities in social networks using max-min modularity. In *Proceedings of the 9th SIAM International Conference on Data Mining*, pages 978–989, 2009.

[44] X. Cheng, C. Dale, and J. Liu. Statistics and social network of YouTube videos. In *Proceedings of the 16th International Workshop on Quality of Service*, pages 229–238, 2008.

[45] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 59–768, 2010.

[46] C. Chew and G. Eysenbach. Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE*, 5(11):e14118, 11 2010.

[47] D. Chiu, P. Ande, R. A. Coward, and A. Woywodt. The times they are a changin': The internet and how it affects daily practice in nephrology. *NDT Plus*, 2(4):273, 2009.

[48] C. K. Chow, K. Lock, K. Teo, S. V. Subramanian, M. McKee, and S. Yusuf. Environmental and societal influences acting on cardiovascular risk factors and disease at a population level: A review. *International Journal of Epidemiology*, 38(6):1580–1594, 2009.

[49] R. Chunara, J. R. Andrews, and J. S. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medical and Hygiene*, 86(1):39–45, 2012.

[50] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72(2): 026132, 2005.

[51] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[52] W. M. Compton and N. D. Volkow. Abuse of prescription drugs and the risk of addiction. *Drug and Alcohol Dependence*, 83(1):S4–S7, 2006.

[53] S. H. Cook, J. A. Bauermeister, D. Gordon-Messer, and M. A. Zimmerman. Online network influences on emerging adults' alcohol and drug use. *Journal of Youth and Adolescence*, pages 1–13, 2012. doi: 10.1007/s10964-012-9869-1.

[54] C. D. Corley, D. J. Cook, A. R. Mikler, and K. P. Singh. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7(2):596–615, 2010.

[55] A. E. Crosby, B. Han, L. A. G. Ortega, S. E. Parks, and J. Gfroerer. Suicidal thoughts and behaviors among adults aged > 18 years - United States 2009. *Morbidity and Mortality Weekly Report*, 60(SS12):1–22, 2011.

[56] R. Crutzen and J. De Nooijer. Intervening via chat: An opportunity for adolescents' mental health promotion? *Health Promotion International*, 26(2):238–243, 2011.

[57] K. L. Daniel. The power of mom in communicating health. *American Journal of Public Health*, 99(12):2119, 2009.

[58] M. de Klepper, E. Sleebos, G. van de Bunt, and F. Agneessens. Similarity in friendship networks: Selection or influence? the effect of constraining contexts and non-visible individual attributes. *Social Networks*, 32(1):82–90, 2010.

[59] Deloitte Center for Health Solutions. 2010 survey of health care consumers. http://www.deloitte.com/assets/Dcom-UnitedStates/Local Assets/Documents/US_CHS_2010SurveyofHealthCare Consumers_050310.pdf, 2010. Accessed: 20 January 2012.

[60] Deloitte Center for Health Solutions. Social networks in health care: Communication, collaboration and insights. http://www.deloitte.com/assets/Dcom-United States/Local Assets/Documents/US_CHS_2010 SocialNetworks_070710.pdf, 2010.

[61] K. Dent and S. Paul. Through the twitter glass: Detecting questions in micro-text. In *Proceedings of the AAAI 2011 Workshop on Analyzing Microtext*, pages 8–13, 2011.

[62] A. D. DeSantis, E. M. Webb, and S. M. Noar. Illicit use of prescription ADHD medications on a college campus: a multimethodological approach. *Journal of American College Health*, 57(3):315–324, 2008.

[63] M. E. Duffy and E. Thorsen. Emerging trends in the new media landscape. In J. C. Parker and E. Thorson, editors, *Health Communication in the New Media Landscape*, pages 93–116. Springer Publishing Company, New York, 2009.

[64] M. Efron and M. Winget. Questions are content: A taxonomy of questions in a microblogging environment. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[65] P. I. Eke. Using social media for research and public health surveillance. *Journal of Dental Research*, 90(9):1045–1046, 2011.

[66] N. Elkin. How america searches: Health and wellness. http://www.icrossing.com/sites/default/files/how-america-searches-health-and-wellness.pdf, 2008. Accessed: 20 January 2012.

[67] D. Estrin and I. Sim. Health care delivery. open mHealth architecture: An engine for health care innovation. *Science*, 330(6005):759–760, 2010.

[68] G. Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, pages 244–248, 2006.

[69] G. Eysenbach. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, 11(1):e11, 2009.

[70] G. Eysenbach. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of Medical Internet Research*, 11(1):e11, 2009.

[71] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 118–127, 2004.

[72] A. D. Farmer, C. E. M. Bruckner Holt, M. J. Cook, and Hearing S. D. Social networking sites: A novel portal for communication. *Postgraduate Medical Journal*, 85:455–459, 2009.

[73] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of Web communities. *Computer*, 35(3):66–70, 2002.

[74] J. A. Ford and M. C. Arrastia. Pill-poppers and dopers: a comparison of non-medical prescription drug use and illicit/street drug use among college students. *Addictive Behaviors*, 33(7):934–941, 2008.

[75] R. F. Forman. Availability of opioids on the Internet. *JAMA: The Journal of the American Medical Association*, 290(7):889, 2003.

[76] S. R. Forsyth and R. E. Malone. "I'll be your cigarette—light me up and get on with it": Examining smoking imagery on YouTube. *Nicotine & Tobacco Research*, 12(8): 810–816, 2010.

[77] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[78] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[79] S. Fox. The social life of health information, 2011. Washington, DC: Pew Internet & American Life Project (online at http://alexa.pewinternet.com/∼/media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf), 2011.

[80] B. Freeman and S. Chapman. Is "YouTube" telling or selling you something? tobacco content on the YouTube video-sharing website. *Tobacco Control*, 16(3):207, 2007.

[81] A. Friggeri, G. Chelius, and E. Fleury. Egomunities, exploring socially cohesive person-based communities. http://arxiv.org/abs/1102.2623, 2011.

[82] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of 3rd Conference on Online Social Networks*, page 3, 2010.

[83] M. C. Gibbons, L. Fleisher, R. E. Slamon, S. Bass, V. Kandadai, and J. R. Beck. Exploring the potential of web 2.0 to address health disparities. *Journal of Health Communication*, 16(sup1):77–89, 2011.

[84] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232): 1012–1014, 2008.

[85] M. Goldberg, S. Kelley, M. Magdon-Ismail, K. Mertsalov, and A. Wallace. Finding overlapping communities in social networks. In *Proceedings of the 2nd IEEE International Conference on Social Computing*, pages 104–113, 2010.

[86] S. K. Goldsmith, T. C. Pellmar, A. M. Kleinman, and W. E. Bunney. *Reducing suicide: A national imperative*. National Academies Press, 2002.

[87] M. S. Gould, J. L. H. Munfakh, K. Lubell, M. Kleinman, and S. Parker. Seeking help from the internet during adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry*, 41(10):1182–1189, 2002.

[88] J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. *Journal of General Internal Medicine*, 26(3):287–292, 2011.

[89] S. Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.

[90] B. Griggs. Twitter: Number of tweets tripled in past year. http://articles.cnn.com/2011-07-01/tech/twitter.tweets_1_tweets-twitter-kate-middleton?_s=PM:TECH, 2011. Accessed: 2 September 2012.

[91] G. J. Gross and M. Howard. Mothers' decision making processes regarding health care for their children. *Public Health Nursing*, 18(3):157–168, 2001.

[92] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of 13th International Conference on World Wide Web*, pages 491–501, 2004. doi: 10.1145/988672.988739. URL `http://doi.acm.org/10.1145/988672.988739`.

[93] C. L. Hanson, S. H. Burton, C. Giraud-Carrier, J. H. West, M. D. Barnes, and B. Hansen. Tweaking and tweeting: Exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research*, 15(4): e62, 2013.

[94] K. Hawton, D. Zahl, and R. Weaterall. Suicide following deliberate self-harm: Long-term follow-up of patients who presented to a general hospital. *British Journal of Psychiatry*, 182:537–542, 2003.

[95] A. J. Hayanga and H. E. Kaiser. Medical information on YouTube. *Journal of the American Medical Association*, 299(12):1424, 2008.

[96] N. Heaivilin, B. Gerbert, J. E. Page, and J. L. Gibbs. Public health surveillance of dental pain via Twitter. *Journal of Dental Research*, 90(9):1047–1051, 2011.

[97] L. Herman, O. Shtayermman, B. Aksnes, M. Anzalone, A. Cormerais, and C. Liodice. The use of prescription stimulants to enhance academic performance among college

students in health care programs. *Journal of Physician Assistant Education*, 22(4): 15–22, 2011.

[98] S. H. Hernandez and L. S. Nelson. Prescription drug abuse: Insight into the epidemic. *Clinical Pharmacology & Therapeutics*, 88(3):307–317, 2010.

[99] K. Heron and J. M. Smyth. Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behavior treatments. *British Journal of Health Psychology*, 15:1–39, 2010.

[100] E. W. Hossler and M. P. Conroy. YouTube as a source of information on tanning bed use. *Archives of Dermatology*, 144(10):1395–1396, 2008.

[101] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? Privacy issues on Twitter. In *Conference of International Communication Association*, 2010.

[102] S. S. Intille, C. Kukla, R. Farzanfar, and W. Bakr. Just-in-time technology to encourage incremental, dietary behavior change. In *Proceedings of the American Medical Informatics Association Annual Symposium*, page 874, 2003.

[103] iProspect.com. iProspect search engine user behavior. Technical report, iProspect.com, Inc., April 2006.

[104] M. K. Islam, J. Merlo, I. Kawachi, M. Lindström, and U. G. Gerdtham. Social capital and health: Does egalitarianism matter? a literature review. *International Journal for Equity in Health*, 5(3), 2006.

[105] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[106] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[107] Sachin H. Jain. Practicing medicine in the age of facebook. *New England Journal of Medicine*, 361(7):649–651, 2009.

[108] J. Jashinsky, S. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. Barnes, and T. Argyle. Tracking suicide risk factors through Twitter in the U.S. *In Submission*, 2013.

[109] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. In *Proc. of 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, pages 56–65, 2007.

[110] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[111] L. D. Johnston, P. M. O'Malley, J. G. Bachman, and J. E. Schulenberg. Monitoring the future: National survey results on drug use, 1975-2009. volume i: Secondary school students. nih publication no. 10-7584. *National Institute on Drug Abuse (NIDA)*, 2010.

[112] A. B. Jonas, A. M. Young, C. B. Oser, C. G. Leukefeld, and J. R. Havens. Oxycontin® as currency: Oxycontin® use and increased social capital among rural Appalachian drug users. *Social Science & Medicine*, 74(10):1602–1609, 2012.

[113] R. Judson and S. W. Langdon. Illicit use of prescription stimulants among college students: prescription status, motives, theory of planned behaviour, knowledge and self-diagnostic tendencies. *Psychology Health and Medicine*, 14(1):97–104, 2009.

[114] C. Kadushin. The friends and supporters of psychotherapy: On social circles in urban life. *American Sociological Review*, 31(6):786–802, 1966.

[115] C. Kadushin. Power, influence and social circles: A new methodology for studying opinion makers. *American Sociological Review*, 33(5):685–699, 1968.

[116] M. N. Kamel Boulos and S. Wheelert. The emerging Web 2.0 social software: An enabling suite of sociable technologies in health and health care education. *Health Information and Libraries Journal*, 24(1):2–23, 2007.

[117] E. Katz, J. G. Blumler, and M. Gurevitch. Utilization of mass communication by the individual. In J. G. Blumler and E. Katz, editors, *The uses of mass communications: Current perspectives on gratifications research*, pages 19–32. Sage, Beverly Hills, CA, 1974.

[118] J. Keelan, V. Pavri-Garcia, G. Tomlinson, and K. Wilson. YouTube as a source of information on immunization: A content analysis. *JAMA: The Journal of the American Medical Association*, 298(21):2482–2484, 2007.

[119] S. Keeter, C. Kennedy, M. Dimock, J. Best, and P. Craighill. Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 70(5):759–779.

[120] R. Kelly. Twitter study reveals interesting results about usage–40% is "pointless babble". http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results-40-percent-pointless-babble, 2010. Accessed: 2 September 2012.

[121] R. R. Khorasgani, J. Chen, and O. R. Zaïane. Top leaders community detection approach in information networks. In *Proceedings of the 4th Workshop on Social Network Mining and Analysis*, 2010.

[122] K. Kim, H-J. Paek, and J. Lynn. A content analysis of smoking fetish videos on YouTube: Regulatory implications for tobacco control. *Health Communications*, 25(2): 97–106, 2010.

[123] Y. Kim, S-W. Son, and H. Jeong. Finding communities in directed networks. *Physical Review E*, 81:016103, Jan 2010.

[124] K. A. King. Developing a comprehensive school suicide prevention program. *Journal of School Health*, 71(4):132–137, 2001.

[125] M. Klein. Combating misuse and abuse of prescription drugs. FDA Consumer Health Information. Online at: http://www.fda.gov/downloads/ForConsumers/ConsumerUpdates/UCM220434.pdf, 2010. Accessed: 22 May 2013.

[126] A. B. Klomek, A. Sourander, and M. S. Gould. Bullying and suicide. *Psychiatric Times*, 28(2):1–6, 2011.

[127] D. E. Knuth. *The Art of Computer Programming: Fundamental Algorithms (Vol. 1)*. Boston: Addison-Wesley, 3rd edition, 1997.

[128] K. D. Kochanek, X. Jiaquan, S. L. Murphy, A. M. Minino, and K. Hsiang-Ching. National vital statistics report: Deaths. http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_03.pdf, 2011. Accessed: 2 September 2012.

[129] M. Kreuter, V. Strecher, and B. Glassman. One size does not fit all: The case for tailoring print materials. *Annals of Behavioral Medicine*, 21(4):276–283, 1999.

[130] L. Lamberg. Psychiatric emergencies call for comprehensive assessment and treatment. *Journal of the American Medical Association*, 288(6):686–687, 2002.

[131] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *2nd International Workshop on Cognitive Information Processing (CIP)*, pages 411–416, 2010.

[132] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):72:1–72:22, 2012.

[133] A. Lancichinetti and S. Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, 2009.

[134] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78(4):046110, 2008.

[135] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[136] R. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[137] P. G. Lange. Commenting on comments: Investigating responses to antagonism on YouTube. In *Proceedings of the 70th Annual Conference of the Society for Applied Anthropology*, 2007.

[138] T. Lappas, K. Liu, and E. Terzi. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 467–476. ACM, 2009.

[139] C. Lee, F. Reid, A. McDaid, and N. Hurley. Detecting highly overlapping community structure by greedy clique expansion. In *The 4th SNA-KDD Workshop (SNA-KDD 2010)*, pages 33–42, 2010.

[140] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, 100(11):118703, Mar 2008.

[141] C. Leukefeld, R. Walker, J. Havens, C. A. Leedham, and V. Tolbert. What does the community say: Key informant perceptions of rural prescription drug use. *Journal of Drug Issues*, 37(3):503–524, 2007.

[142] P. M. Lewinsohn, P. Rohde, and J. R. Seely. Psychosocial risk factors for future adolescent suicide attempts. *Journal of Consulting and Clinical Psychology*, 62(2): 297–305, 1994.

[143] M. A Lewis and C. Neighbors. Gender-specific misperceptions of college student drinking norms. *Psychology of Addictive Behaviors*, 18(4):334–339, 2004.

[144] M. Linkletter, K. Gordon, and J. Dooley. The choking game and YouTube: A dangerous combination. *Clinical Pediatrics*, 49(3):274–279, 2009.

[145] K. A. Lochner, I. Kawachi, R. T. Brennan, and S. L. Buka. Social capital and neighborhood mortality rates in chicago. *Social Science & Medicine*, 56(8):1797–1806, 2003.

[146] J. Lomas. Social capital and health: Implications for public health and epidemiology. *Social Science & Medicine*, 47(9):1181–1188, 1998.

[147] K. G. Low and A. E. Gendaszek. Illicit use of psychostimulants among college students: A preliminary study. *Psychology, Health & Medicine*, 7(3):283–287, 2002.

[148] F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6(4):387–400, 2008.

[149] X. Ma, G. Chen, and J. Xiao. Analysis of an online health social network. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 297–306, 2010.

[150] R. Markman. Hip-hop artist Freddy E dead in apparent suicide. http://www.mtv.com/news/articles/1699754/freddy-e-dead.jhtml, 2013. Accessed: 8 February 2013.

[151] N. McAuliffe and L. Perry. Making it safer: A health centre's strategy for suicide prevention. *Psychiatry Quarterly*, 78:295–307, 2007.

[152] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 548–556. 2012.

[153] S. E. McCabe, J. R. Knight, Christian J. Teter, and H. Wechsler. Non-medical use of prescription stimulants among US college students: Prevalence and correlates from a national survey. *Addiction*, 100(1):96–106, 2005.

[154] S. E. McCabe, C. J. Teter, and C. J. Boyd. Medical use, illicit use and diversion of prescription stimulant medication. *Journal of Psychoactive Drugs*, 38(1):43–56, 2006.

[155] A. K. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[156] J. McLean, M. Maxwell, S. Platt, F. Harris, and R. Jepson. Risk and protective facts for suicide and suicidal behavior: A literature review. http://www.scotland.gov.uk/Publications/2008/11/28141444/0, 2008. Accessed: 2 September 2012.

[157] A. D. McNiel, K. B. Muzzin, J. P. DeWald, A. L. McCann, E. D. Schneiderman, J. Scofield, and P. R. Campbell. The nonmedical use of prescription stimulants among dental and dental hygiene students. *Journal of Dental Education*, 75(3):365–376, 2011.

[158] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[159] D. Mechanic. The influence of mothers on their children's health attitudes and behavior. *Pediatrics*, 33(3):444–453, 1964.

[160] R. M. Merchant, S. Elmer, and N. Lurie. Integrating social media into emergency-preparedness efforts. *New England Journal of Medicine*, 365(4):289–291, 2011.

[161] E. A. Miller and A. Pole. Diagnosis blog: Checking up on health blogs in the blogosphere. *American Journal of Public Health*, 100(8):1514–1519, 2010.

[162] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42, 2007.

[163] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 251–260, 2010.

[164] M. R. Morris, J. Teevan, and K. Panovich. A comparison of information seeking using search engines and social networks. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 291–294, 2010.

[165] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, pages 1739–1748, 2010.

[166] S. T. Moturu, H. Liu, and W. G. Johnson. Trust evaluation in health information on the world wide web. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, pages 1525–1528, 2008.

[167] K. Murugiah, A. Vallakati, K. Rajput, A. Sood, and N. Challa. YouTube as a source of information on cardiopulmonary resuscitation. *Resuscitation*, 82(3):332–334, 2011.

[168] National Center on Addiction and Substance Abuse at Columbia University. You've got drugs! Prescription drug pushers on the internet. http://www.casacolumbia.org/articlefiles/531-2008%20You've%20Got%20Drugs2007. Accessed: 14 May 2013.

[169] National Center on Addiction and Substance Abuse at Columbia University. National survey of american attitudes on substance abuse xvi: Teens and parents. http://www.casacolumbia.org/upload/2011/20110824teensurveyreport.pdf, 2011. Accessed: 14 May 2013.

[170] National Health Expenditure Accounts. National health expenditures 2010 highlights. http://www.cms.gov/NationalHealthExpendData/downloads/highlights.pdf, 2011. Accessed: April 2012.

[171] National Institute of Mental Health. Suicide in the U.S.: Statistics and prevention. http://www.nimh.nih.gov/health/publications/suicide-in-the-us-statistics-and-prevention/index.shtml, 2012. Accessed: 2 September 2012.

[172] National Institute on Drug Abuse. Research report series: Prescription drugs abuse and addiction. http://www.dhp.virginia.gov/dhp_programs/pmp/docs/NIDA%20Research%20Report%20on%20Prescription%20Drugs.pdf, 2005. Accessed: 23 May 2013.

[173] National Public Health Performance Standards Program. 10 essential public health services. http://www.cdc.gov/nphpsp/essentialservices.html. Accessed: April 2012.

[174] National Survey on Drug Use and Health. National findings, substance abuse mental health services administration (SAMHSA). http://www.samhsa.gov/data/nsduh/2k10MH_Findings/2k10MHResults.pdf. Accessed: 22 May 2013.

[175] B. L. Neiger, R. Thackeray, S. H. Burton, C. G. Giraud-Carrier, and M. C. Fagen. Evaluating social media's capacity to develop engaged audiences in health promotion settings: Use of Twitter metrics as a case study. *Health Promotion Practice*, 14:157–162, 2013.

[176] A. Neustein. Sequence package analysis: A new natural language understanding method for intelligent mining of recordings of doctor-patient interviews and health-related blogs.

In *Fourth International Conference on Information Technology (ITNG)*, pages 431–438, 2007.

[177] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133, 2004.

[178] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[179] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[180] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.

[181] M.E.J. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, 2011.

[182] N. P. Nguyen, T. N. Dinh, D. T. Nguyen, and M. T. Thai. Overlapping community structures and their detection on social networks. In *Proceedings of 3rd IEEE International Conference on Social Computing*, pages 35–40, 2011.

[183] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03024, 2009.

[184] R. Niska, F. Bhuiya, and J. Xu. National hospital ambulatory medical care survey: 2007 emergency department visits. http://198.246.98.21/nchs/data/nhsr/nhsr026.pdf, 2010. Accessed: 2 September 2012.

[185] Office of National Drug Control Policy and US Executive Office of the President. Epidemic: Responding to america's prescription drug abuse crisis. https://www.ncjrs.gov/pdffiles1/ondcp/rx_abuse_plan.pdf, 2011. Accessed: 22 May 2013.

[186] H-J. Paek, K. Kim, and T. Hove. Content analysis of antismoking videos on YouTube: Message sensation value, message appeals, and their relationships with viewer responses. *Health Education Research*, 25(6):1085–1099, 2010.

[187] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[188] A. Pandey, N. Patni, M. Singh, A. Sood, and G. Singh. Youtube as a source of information on the H1N1 influenza pandemic. *American Journal of Preventive Medicine*, 38(3):e1–3, 2010.

[189] J. C. Paolillo. Structure and network in the youtube core. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pages 156–165, 2008.

[190] S. Papadopoulos, A. Skusa, A. Vakali, Y. Kompatsiaris, and N. Wagner. Bridge bounding: A local approach for efficient community discovery in complex networks. *Arxiv preprint arXiv:0902.0871*, 2009.

[191] K. Patrick, S. S. Intille, and M. F. Zabinski. An ecological framework for cancer communication: Implications for research. *Journal of Medical Internet Research*, 7(3): e23, 2005.

[192] M. J. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[193] S. A. Paul, L. Hong, and E. H. Chi. What is a question? Crowdsourcing tweet categorization. In *Proceedings of the CHI 2011 Workshop on Crowdsourcing and Human Computation*, 2011.

[194] S. A. Paul, L. Hong, and E. H. Chi. Is Twitter a good place for asking questions? A characterization study. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 578–581, 2011.

[195] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A. J. Valleron. More diseases tracked by using Google trends. *Emerging Infectious Diseases*, 15(8):1327, 2009.

[196] A. Pentland, D. Lazer, D. Brewer, and T. Heibeck. Using reality mining to improve public health and medicine. *Studies in Health Technology and Informatics*, 149:93–102, 2009.

[197] H. W. Perkins, P. W. Meilman, J. S. Leichliter, J. R. Cashin, and C. A. Presley. Misperceptions of the norms for the frequency of alcohol and other drug use on college campuses. *Journal of American College Health*, 47(6):253–258, 1999.

[198] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448, 2008.

[199] P. Pons and M. Latapy. Computing communities in large networks using random walks. In *Proceedings of the 20th International Symposium on Computer and Information Sciences (LNCS 3733)*, pages 284–293, 2005.

[200] C. Poulin. Medical and nonmedical stimulant use among adolescents: From sanctioned to unsanctioned use. *Canadian Medical Association Journal*, 165(8):1039–1044, 2001.

[201] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *Proceedings of the 4th International Conference on Social Computing, Behavioral-cultural Modeling and Prediction*, pages 18–25, 2011.

[202] H. Qin, T. Liu, and Y. Ma. Mining user's real social circle in microblog. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 348–352, 2012.

[203] D. L. Rabiner, A. D. Anastopoulos, E. J. Costello, R. H. Hoyle, S. E. McCabe, and H. S. Swartzwelder. Motives and perceived consequences of nonmedical adhd medication use by college students are students treating themselves for attention problems? *Journal of Attention Disorders*, 13(3):259–270, 2009.

[204] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76:036106, 2007.

[205] S. C. Ratzan. Our new "social" communication age in health. *Journal of Health Communication*, 16(8):803–804, 2011.

[206] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74:016110, 2006.

[207] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of 20th International Conference on World Wide Web*, pages 695–704, 2011.

[208] M. Rosenberg. *Society and the adolescent self-image (rev).* Wesleyan University Press, 1989.

[209] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[210] T. D. Ruder, G. M. Hatch, G. Ampanozi, M. J. Thali, and N. Fischer. Suicide announcement on Facebook. *Crisis*, 32(5):280–282, 2011.

[211] Melodie Rydalch. Prescription drug abuse: A conversation we should keep having. http://www.justice.gov/usao/ut/press/releases/2012/September /PrescriptionDrugAbuse_Mel09232012.htm, 2012. Accessed: April 2013.

[212] R. L. T. Santos, B. P. S. Rocha, C. G. Rezende, and A. A. F. Loureiro. Characterizing the YouTube video-sharing community. http://security1.win.tue.nl/ ∼bpontes/pdf/yt.pdf, 2006. Accessed: May 2011.

[213] N. Savage. Twitter as medium and message. *Communications of the ACM*, 54(3):18–20, 2011.

[214] D. Scanfeld, V. Scanfeld, and E. L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3): 182–188, 2010.

[215] M. E. Shaw. *Group dynamics: the psychology of small group behavior*. McGraw-Hill, New York, 1976.

[216] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza a H1N1 pandemic. *PLoS One*, 6(5):e19467, 2011.

[217] A. Smith and J. Brenner. Twitter use 2012. http://pewinternet.org/Reports/2012/Twitter-Use-2012/Findings/Twitter-use.aspx, 2012. Accessed: June 20, 2012.

[218] M. Smith, C. Giraud-Carrier, and N. Purser. Implicit affinity networks and social capital. *Information Technology and Management*, 10(2):123–134, 2009.

[219] S. Smyth and S. White. A spectral clustering approach to finding communities in graphs. In *Proceedings of the 5th SIAM International Conference on Data Mining*, pages 76–84, 2005.

[220] A. Sood, S. Sarangi, A. Pandey, and K. Murugiah. Youtube as a source of information on kidney stone disease. *Urology*, 77(3):558–562, 2011.

[221] M. Sozio and A. Gionis. The community-search problem and how to plan a successful cocktail party. In *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 939–948, 2010.

[222] A. Stanoev, D. Smilkov, and L. Kocarev. Identifying communities by influence dynamics in social networks. *Physical Review E*, 84:046102, 2011.

[223] P. L. Steinberg, S. Wason, J. M. Stern, L. Deters, B. Kowal, and J. Seigne. Youtube as source of prostate cancer information. *Urology*, 75(3):619–622, 2010.

[224] S. A. Stoddard, J. A. Bauermeister, D. Gordon-Messer, M. Johns, and M. A. Zimmerman. Permissive norms and young adults' alcohol and marijuana use: The role of online communities. *Journal of Studies on Alcohol and Drugs*, 73(6):968–975, 2012.

[225] T. Ståhl, A. Rütten, D. Nutbeam, A. Bauman, L. Kannas, T. Abel, G. Lüschen, D. J. A. Rodriquez, J. Vinck, and J. van der Zee. The importance of the social environment for physically active lifestyle — results from an international study. *Social Science & Medicine*, 52(1):1–10, 2001.

[226] V. Strecher, C. Wang, H. Derry, K. Wildenhaus, and C. Johnson. Tailored interventions for multiple risk behaviors. *Health Education Research*, 17(5):619–626, 2002.

[227] V. J. Strecher, J. B. McClure, G. L. Alexander, B. Chakraborty, V. N. Nair, J. M. Konkel, and et al. Web-based smoking-cessation programs: Results of a randomized trial. *American Journal of Preventive Medicine*, 34(5):373–381, 2008.

[228] Substance Abuse and Mental Health Services Adminstration. The NSDUH report: Nonmedical use of Adderall among full-time college students. http://oas.samhsa.gov/2k9/adderall/adderall.htm, 2009. Accessed: 17 December 2012.

[229] L. K. Suzuki and J. P. Calzo. The search for peer advice in cyberspace: An examination of online teen bulletin boards about health and sexuality. *Journal of Applied Developmental Psychology*, 25(6):685–698, 2004.

[230] Teen Drug Rehab. Prescription drug and otc abuse among teens in iowa. http://teen-drug-rehab.blogspot.com/2011/02/prescription-drug-and-otc-abuse-among.html. Accessed: 16 February 2013.

[231] C. J. Teter, S. E. McCabe, C. J. Boyd, and S. K. Guthrie. Illicit methylphenidate use in an undergraduate student sample: prevalence and risk factors. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 23(5):609–617, 2003.

[232] C. J. Teter, S. E. McCabe, J. A. Cranford, C. J. Boyd, and S. K. Guthrie. Prevalence and motives for illicit use of prescription stimulants in an undergraduate student sample. *Journal of American College Health*, 53(6):253–262, 2005.

[233] R. Thackeray, S. H. Burton, C. Giraud-Carrier, S. Rollins, and C. R. Draper. Using social media for breast cancer prevention: An analysis of breast cancer awareness month. *In Submission*, 2013.

[234] H. Tong and C. Faloutsos. Center-piece subgraphs: Problem definition and fast solutions. In *Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 404–413, 2006.

[235] Twitter. 200 million tweets per day. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html, 2011. Accessed: June 27, 2012.

[236] U.S. Food and Drug and Aministration. FDA warns consumers about counterfeit version of Teva's Adderall. http://www.webcitation.org/6CzHsIlpU, 2012. Accessed: 17 December 2012.

[237] U.S. Public Health Service, Department of Health and Human Services. The surgeon general's call to action to prevent suicide. http://www.surgeongeneral.gov/library/calltoaction/default.htm, 1999. Accessed: 2 September 2012.

[238] T. W. Valente. *Social Networks and Health*. Oxford University Press, New York, NY, 2010.

[239] K. Vance, W. Howe, and R. P. Dellavalle. Social internet sites as a source of public health information. *Dermatologic Clinics*, 27(2):133–136, 2009.

[240] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *Proceedings of the 10th IEEE International Conference on Data Mining*, pages 569–578, 2010.

[241] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[242] C. Wei. Formation of norms in a blog community. In L. Gurak, S. Antonijevic, L. Johnson, and C. Ratliff, editors, *Into the Blogosphere: Rhetoric, Community, and Culture in Weblogs*. University of Minnesota, 2004.

[243] J. West, P. C. Hall, C. Hanson, R. Thackeray, M. Barnes, B. Neiger, and E. McIntyre. Breastfeeding and blogging: Exploring the utility of blogs to promote breastfeeding. *American Journal of Health Education*, 43(2):106–115, 2011.

[244] J. H. West, P. C. Hall, K. Prier, C. L. Hanson, C. Giraud-Carrier, E. S. Neeley, and M. D. Barnes. Temporal variability of problem drinking on Twitter. *Open Journal of Preventive Medicine*, 2(1):43–48, 2012.

[245] B. P. White, K. A. Becker-Blease, and K. Grace-Bishop. Stimulant medication use, misuse, and abuse in an undergraduate and graduate student sample. *Journal of American College Health*, 54(5):261–268, 2006.

[246] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In *Proceedings of the 5th SIAM International Conference on Data Mining*, pages 274–285, 2005.

[247] Christopher Paul Wild. The exposome: From concept to utility. *International Journal of Epidemiology*, 41(1):24–32, 2012.

[248] K. Wilson and J. S. Brownstein. Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal*, 180(8):829, 2009.

[249] World Health Organization. Public health surveillance. http://www.who.int/topics/public_health_surveillance/en/, 2012. Accessed: Apr 2012.

[250] Y. Xiang, D. Fuhry, K. Kaya, R. Jin, Ü. Çatalyürek, and K. Huang. Merging network patterns: A general framework to summarize biomedical network data. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, pages 1–14, 2012.

[251] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger. SCAN: A structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 824–833. ACM, 2007.

[252] Yahoo! Inc. Yahoo! place finder API. http://developer.yahoo.com/geo/placefinder/, 2011. Accessed December 1, 2011.

[253] J. Yang and J. Leskovec. Structure and overlaps of communities in networks. In *Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis*, 2012.

[254] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the IEEE International Conference on Data Mining*, pages 745–754, 2012.

[255] J. Yang and J. Leskovec. Community-affiliation graph model for overlapping network community detection. In *Proceedings of the IEEE International Conference on Data Mining*, pages 1170–1175, 2012.

[256] I. H. Yen and S. L. Syme. The social environment and health: A discussion of the epidemiologic literature. *Annual Review of Public Health*, 20(1):287–308, 1999.

[257] S. Young. White house launches effort to combat soaring prescription drug abuse. CNN. http://www.cnn.com/2011/HEALTH/04/19/drug.abuse/index.html. Accessed: 16 February 2013.

[258] D. L. Zahl and K. Hawton. Repetition of deliberate self-harm and subsequent suicide risk: Long-term follow-up study of 11 583 patients. *The British Journal of Psychiatry*, 185:70–75, 2004.

[259] W. Zhang. *State-space Search: Algorithms, Complexity, Extensions, and Applications*. Springer: New York, 1999.

[260] X. Zhang, H. Fuehres, and P. A. Gloor. Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26:55–62, 2011.

[261] K. Zickuhr and A. Smith. 28% of American adults use mobile and social location-based services. http://pewinternet.org/∼/media/Files/Reports/2011/PIP_Location-based-services.pdf, 2011. Accessed December 6, 2011.