THIS IS THE FINAL AUTHORS' VERSION OF A PUBLISHED BOOK CHAPTER. WHEN CITING THIS WORK, PLEASE CITE THE PUBLISHED VERSION: Foote, Jeremy D., Aaron Shaw, and Benjamin Mako Hill. 2017. "A Computational Analysis of Social Media Scholarship." In *The SAGE Handbook of Social Media*, edited by Jean Burgess, Alice Marwick, and Thomas Poell, 111–34. London, UK: SAGE.

A Computational Analysis of Social Media Scholarship

Jeremy Foote (jdfoote@u.northwestern.edu)

Aaron Shaw (aaronshaw@northwestern.edu)

Benjamin Mako Hill (makohill@uw.edu)

Abstract Data from social media platforms and online communities have fueled the growth of computational social science. In this chapter, we use computational analysis to characterize the state of research on social media and demonstrate the utility of such methods. First, we discuss how to obtain datasets from the APIs published by many social media platforms. Then, we perform some of the most widely used computational analyses on a dataset of social media scholarship we extract from the Scopus bibliographic database's API. We apply three methods: network analysis, topic modeling using latent Dirichlet allocation, and statistical prediction using machine learning. For each technique, we explain the method and demonstrate how it can be used to draw insights from our dataset. Our analyses reveal overlapping scholarly communities studying social media. We find that early social media research applied social network analysis and quantitative methods, but the most cited and influential work has come from marketing and medical research. We also find that publication venue and, to a lesser degree, textual features of papers explain the largest variation in incoming citations. We conclude with some consideration of the limitations of computational research and future directions.

INTRODUCTION

The combination of large-scale trace data generated through social media with a series of advances in computing and statistics have enabled the growth of

'computational social science' (Lazer et al., 2009). This turn presents an unprecedented opportunity for researchers who can now test social theories using massive datasets of fine-grained, unobtrusively collected behavioral data. In this chapter, we aim to introduce non-technical readers to the promise of these computational social science techniques by applying three of the most common approaches to a bibliographic dataset of social media scholarship. We use our analyses as a context for discussing the benefits of each approach as well as some of the common pitfalls and dangers of computational approaches.

The chapter walks through the entire process of computational analysis, beginning with data collection. We explain how we gather a large-scale dataset about social media research from the *Scopus* website's application programming interface. The dataset we collect contains metadata about every article in the Scopus database that includes the term 'social media' in its title, abstract, or keywords. Using this dataset, we perform multiple computational analyses. First, we use network analysis (Wasserman & Faust, 1994) on article citation metadata to understand the structure of references between the articles. Second, we use topic models (Blei, 2012), an unsupervised natural language processing technique, to describe the distribution of topics within the sample of articles included in our analysis. Third, we perform statistical prediction (James, Witten, Hastie, & Tibshirani, 2013) in order to understand what characteristics of articles best predict subsequent citations. For each analysis, we describe the method we use in detail and discuss some of its benefits and limitations.

Our results reveal several patterns in social media scholarship. Bibliometric network data reveals disparities in the degree that disciplines cite each other and illustrate that marketing and medical research each enjoy surprisingly large influence. Through descriptive analysis and topic modeling, we find evidence of the early influence of social network research. When we use papers' characteristics to predict which work gets cited, we find that publication venues and linguistic features provide the most explanatory power.

In carrying out our work in this chapter, we seek to exemplify several current best practices in computational research. We use data collected in a manner consistent with the expectations of privacy and access held by the subjects of our analysis as well as the publishers of the data source. We also make our analysis fully reproducible from start to finish. In an online supplement, we provide the full source code for all aspects of this project – from the beginning of data collection to the creation of the figures and the chapter text itself – as

a resource for future researchers.

Collecting and describing data from the web

A major part of computational research consists of obtaining data, preparing it for analysis, and generating initial descriptions that can help guide subsequent inquiry. Social media datasets vary in how they make it into researchers' hands. There are several sources of social media data which are provided in a form that is pre-processed and ready for analysis. For example, The Stanford Large Network Dataset Collection (Leskovec & Krevl, 2014) contains pre-formatted and processed data from a variety of social media platforms. Typically, prepared datasets come formatted as 'flat files' such as comma-separated value (CSV) tables, which many types of statistical software and programming tools can import directly.

More typically, researchers retrieve data directly from social media platforms or other web-based sources. These 'primary' sources provide more extensive, dynamic, and up-to-date datasets, but also require much more work to prepare the data for analysis. Typically, researchers retrieve these data from social media sites through application programming interfaces (APIs). Web sites and platforms use APIs to provide programmers with limited access to their servers and databases. Unfortunately, APIs are rarely designed with research in mind and are often inconvenient and limited for social scientists as a result. For example, Twitter's search API returns a small, non-random sample of tweets by default (what a user might want to read), rather than all of the tweets that match a given query (what a researcher building a sample would want). In addition, APIs typically limit how much data they will provide for each query and how many queries can be submitted within a given time period.

APIs provide raw data in formats like XML or JSON, which are poorly suited to most data analysis tasks. As a result, researchers must take the intermediate step of converting data into more appropriate formats and structures. Typically, researchers must also construct measures from the raw data, such as user-level statistics (e.g., number of retweets) or metadata (e.g., post length). A number of tools, such as NodeXL (Hansen, Shneiderman, & Smith, 2010), exist to make the process of obtaining and preparing digital trace data easier. However, off-the-shelf tools tend to come with their own limitations and, in our experience, gathering data amenable to computational analysis usually involves some programming work.

Compared to some traditional forms of data collection, obtaining and preparing social media data has high initial costs. It frequently involves writing and debugging custom software, reading documentation about APIs, learning new software libraries, and testing datasets for completeness and accuracy. However, computational methods scale very well and gathering additional data often simply means expanding the date range in a program . Contrast this with interviews, surveys, or experiments, where recruitment is often labor-intensive, expensive, and slow. Such scalability, paired with the massive participation on many social media platforms, can support the collection of very large samples.

Our application: The Scopus Bibliographic Database

We used a series of Scopus Bibliographic Database APIs to retrieve data about all of the publications in their database that contained the phrase 'social media' in their abstract, title, or keywords. We used the Python programming language to write custom software to download this data. First, we wrote a program to query the Scopus Search API to retrieve a list of the articles that matched our criteria. We stored the resulting list of 23,131 articles in a file. We used this list of articles as input to a second program, which used the Scopus Citations Overview API to retrieve metadata about all of the articles that cited these 23,131 articles. Finally, we wrote a third program that used the Scopus Abstract Retrieval API to download abstracts and additional metadata about the original 23,131 articles. Due to rate limits and the process of trial and error involved in writing, testing, and debugging these custom programs, it took a few weeks to obtain the complete dataset.

Like many social media APIs, the Scopus APIs returns data in JSON format. Although not suitable for analysis without processing, we stored this JSON data in the form it was given to us. Retaining the 'raw' data as it was provided by APIs allows researchers to construct new measures they might not have believed were relevant in the early stages of their research and to fix any bugs that they find in their data processing and reduction code without having to re-download raw data. Once we obtained the raw data, we wrote additional Python scripts to turn the downloaded JSON files into CSV tables which could be imported into Python and R, the programming languages we used to complete our analyses.

Results

The Scopus dataset contains a wide variety of data, and many different descriptive statistics could speak to various research questions. Here we present a sample of the sorts of summary data that computational researchers might explore. We begin by looking at where social media research is produced. Table 6.1 shows the number of papers produced by authors located in each of the six most frequently seen countries in our dataset.[1] We can immediately see that the English-language world produces much of the research on social media (which is perhaps unsurprising given that our search term was in English), but that a large amount of research comes from authors in China and Germany.

Country	Number of Papers
United States	7812
United Kingdom	1711
Australia	1096
China	926
Germany	787
Canada	771

Table 6.1: Top author countries by number of social media papers.

Next we look at the disciplines that publish social media research. Figure 6.1 shows the number of papers containing the term 'social media' over time. The plot illustrates that the quantity of published research on social media has increased rapidly over time. The growth appears to slow down more recently, but this may be due to the speed at which the Scopus database imports data about new articles.

Figure 6.1 shows the top ten disciplines, as categorized by Scopus. We see that the field started off dominated by computer science publications, with additional disciplines increasing their activity in recent years. This story is also reflected in the top venues, listed in Table 6.2, where we see that computer science venues have published social media research most frequently.

[1]Technically, each paper is assigned to the modal (i.e., most frequent) country among its authors. For example, if a paper has three authors with two authors located in Canada and one in Japan, the modal country for the paper would be Canada. Any ties (i.e., if more than one country is tied for most frequent location among a paper's authors) were broken by randomly selecting among the tied countries.

Publication Venue	Papers
Lecture Notes in Computer Science	935
ACM International Conference Proceeding Series	288
Computers in Human Behavior	257
CEUR Workshop Proceedings	227
Proceedings of the Hawaii International Conference on System Sciences	179
Journal of Medical Internet Research	170

Table 6.2: Venues with the most social media papers.

Title	Publication Venue	Cited by
Users of the world, unite! The challenges and opportunities of Social Media	Business Horizons	1876
Why we twitter: Understanding microblogging usage and communities	Proceedings of WebKDD / SNA-KDD 2007	645
Social media? Get serious! Understanding the functional building blocks of social media	Business Horizons	468
Social media: The new hybrid element of the promotion mix	Business Horizons	450
Role of social media in online travel information search	Tourism Management	389
Networked narratives: Understanding word-of-mouth marketing in online communities	Journal of Marketing	335

Table 6.3: Most cited social media papers.

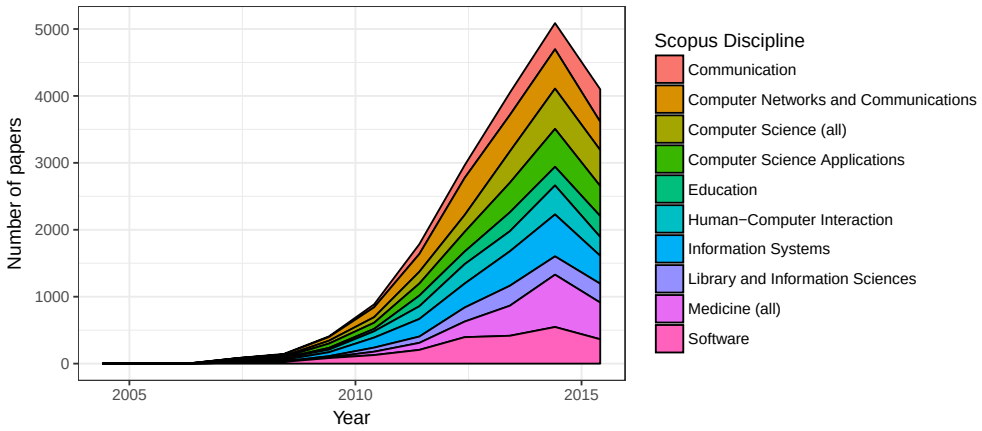

Figure 6.1: Social media papers published in the top ten disciplines (as categorized by Scopus), over time.

We then consider the impact of this set of papers as measured by the citations they have received. Like many phenomena in social systems, citation counts follow a highly skewed distribution with a few papers receiving many citations and most papers receiving very few. Table 6.3 provides a list of the most cited papers. These sorts of distributions suggest the presence of 'preferential attachment' (Barabási & Albert, 1999) or the 'Matthew effect' (Merton, 1968), where success leads to greater success.

Discussion

The summary statistics and exploratory visualizations presented above provide an overview of the scope and trajectory of social media research. We find that social media research is growing – both overall and within many disciplines. We find evidence that computer scientists laid the groundwork for the study of social media, but that social scientists, learning scientists, and medical researchers have increasingly been referring to social media in their published work. We also find several business and marketing papers among the most cited pieces of social media research even though neither these disciplines nor their journals appear among the most prevalent in the dataset.

These results are interesting and believable because they come from a comprehensive database of academic work. In most social science contexts, researchers have to sample from a population and that sampling is often biased. For example, the people willing to come to a lab to participate in a study or take a phone survey may have different attributes from those unwilling to par-

ticipate. This makes generalizing to the entire population problematic. When using trace data, on the other hand, we often have data from all members of a community including those who would not have chosen to participate. One of the primary benefits of collecting data from a comprehensive resource like Scopus is that it can reduce some types of bias in the data collection process. For example, we do not have backgrounds in education or medical research; had we tried to summarize the state of social media research by identifying articles and journals manually, we might have overlooked these disciplines.

That said, this apparent benefit can also become a liability when we seek to generalize our results beyond the community that we have data for. The large *N* of big data studies using social media traces may make results appear more valid, precise, or certain, but a biased sample does not become less biased just because it is larger (Hargittai, 2015). For example, a sample of 100 million Twitter users might be a worse predictor of election results than a truly random sample of only 1,000 likely voters because Twitter users likely have different attributes and opinions than the voting public. Another risk comes from the danger that data providers collect or filter data in ways that aren't apparent. Researchers should think carefully about the relationship of their data to the population they wish to study and find ways to estimate bias empirically.

Overall, we view the ease of obtaining and analyzing digital traces as one of the most exciting developments in social science. Although the hurdles involved represent a real challenge to many scholars of social media today, learning the technical skills required to obtain online trace data is no more challenging than the statistics training that is part of many PhD programs. Below, we present examples of a few computational analyses that can be done with this sort of data.

Network analysis

Social network analysis encompasses the most established set of computational methods in the social sciences (Wasserman & Faust, 1994). At its core, network analysis revolves around a 'graph' representation of data that tries to capture relationships (called edges) between discrete objects (called nodes). Graphs can represent any type of object and relationship, such as roads connecting a group of cities or shared ingredients across a set of recipes. Graph representations of data, and the network analytic methods built to reason using these data, are widely used across the social sciences as well as other fields

including physics, genomics, computer science, and philosophy. 'Social network analysis' constitutes a specialized branch of network analysis in which nodes represent people (or other social entities) and edges represent social relationships like friendship, interaction, or communication.

The power of network analysis stems from its capacity to reduce a very large and complex dataset to a relatively simple set of relations that possess enormous explanatory power. For example, Hausmann et al. (2014) use network data on the presence or absence of trading relationships between countries to build a series of extremely accurate predictions about countries' relative wealth and economic performance over time. By reasoning over a set of relationships in a network, Hausmann and his colleagues show that details of the nature or amount of goods exchanged are not necessary to arrive at accurate economic conclusions.

Network analysis has flourished in studies of citation patterns within scholarly literature, called 'bibliometrics' or 'scientometrics.' Bibliometric scholars have developed and applied network analytic tools for more than a half-century (Kessler, 1963; Hood & Wilson, 2001). As a result, bibliometric analysis provides an obvious jumping-off point for our tour of computational methods. Because network methods reflect a whole family of statistics, algorithms, and applications, we focus on approaches that are both well-suited to bibliometric analysis and representative of network analyses used in computational social science more broadly.

Our application: Citation networks

Our network analysis begins by representing citation information we collected from the Scopus APIs as a graph. In our representation, each node represents a paper and each edge represents a citation. Scopus provides data on incoming citations for each article. Our full dataset includes 35,620 incoming citations to the 23,131 articles in Scopus with 'social media' in their titles, abstracts, or keywords. 19,267 of these articles (83%) have not been cited even once by another article in Scopus and 18,324 (79%) do not cite any other article in our sample. The recent development of social media and the rapid growth of the field depicted in Figure 6.1 might help explain the sparseness (i.e. lack of connections) of the graph. As a result, and as is often the case in network analysis, a majority of our dataset plays no role in our analysis described in the rest of this section.

Once we create our citation graph, there are many potential ways to an-

alyze it. One important application, common to bibliometrics, is the computational identification of communities or clusters within networks. In network studies, the term 'community' is used to refer to groups of nodes that are densely connected to each other but relatively less connected to other groups. In bibliometric analyses, communities can describe fields or sub-fields of articles which cite each other, but are much less likely to cite or be cited by papers in other groups. Although there are many statistical approaches to community detection in network science, we use a technique from Rosvall and Bergstrom (2008) that has been identified as appropriate for the study of bibliometric networks (Šubelj, Eck, & Waltman, 2016). By looking at the most frequently occurring journals and publication venues in each community, we are able to identify and name sub-fields of social media research as distinct communities.

A citation graph is only one possible network representation of the relationships between articles. For example, the use of common topics or terminology might constitute another type of edge. Alternatively, journals or individual authors (rather than articles) might constitute an alternative source of nodes. In bibliometric analyses, for example, it is common for edges to represent 'co-citations' between articles or authors. Using this approach, papers are said to be tied together by a co-citation if they have both been cited in a third document (Small, 1973). Due to limited space, we only present the simplest case of direct citations.

Results

As is common in social networks, the large majority of articles with any citations connect to each other in one large 'component' or sub-network. Figure 6.2 shows a visualization of this large component. The optimal way to represent network data in two-dimensional space is a topic of research and debate. Figure 6.2 uses a force-directed drawing technique (Fruchterman & Reingold, 1991), the most widely used algorithm in network visualization, using the free/open source software package Gephi (Bastian, Heymann, Jacomy, et al., 2009). The basic idea behind the algorithm is that nodes naturally push away from each other, but are pulled together by the edges between them. Shades in each graph in this section reflect the communities of documents identified by Rosvall and colleagues' 'map' algorithm (Martin Rosvall & Bergstrom, 2008; M. Rosvall, Axelsson, & Bergstrom, 2010). Although the algorithm identified several dozen communities, most are extremely small, so we have shown

Figure 6.2: Network visualization of the citation network in our dataset. The layout is 'force directed' meaning that nodes (papers) with more edges (citations) appear closer to the center of the figure.

only the largest 6 communities in Figure 6.2. Each of these communities are summarized in Table 6.4 where the right-most column lists the three most common journals for the articles included in each community.

At this point, we could look in more depth at the attributes of the different communities. For example, in a bibliometric analysis published in the journal *Scientometrics*, Kovács, Looy, and Cassiman (2015) reported summary statistics for articles in each of the major communities identified (e.g., the average number of citations) as well as qualitative descriptions of the nodes in each community. We can see from looking at Table 6.4 that the communities point to the existence of coherent thematic groups. For example, Community 1 includes biomedical research while Community 3 contains papers published in communication journals. Earlier, we relied on an existing category scheme applied to journals to create Figure 6.1; all articles published in particular journals were treated as being within one field. Network analysis, however, can

Community	Description	Journals
Community 1	biomedicine; bioinformatics	Journal of Medical Internet Research; PLoS ONE; Studies in Health Technology and Informatics
Community 2	information technology; management	Computers in Human Behavior; Business Horizons; Journal of Interactive Marketing
Community 3	communication	Information Communication and Society; New Media and Society; Journal of Communication
Community 4	computer science; network science	Lecture Notes in Computer Science; PLoS ONE; WWW; KDD
Community 5	psychology; psychometrics	Computers in Human Behavior; Cyberpsychology, Behavior, and Social Networking; Computers and Education
Community 6	multimedia	IEEE Transactions on Multimedia; Lecture Notes in Computer Science; ACM Multimedia

Table 6.4: Description of each of the citation network clusters identified by the community detection algorithm, together with a list of the three most common journals in each community.

identify groups and categories of articles in terms of who is citing whom and, as a result, can reveal groups that cross journal boundaries. PLoS ONE, for example, is a 'megajournal' that publishes articles from all scientific fields (Binfield, 2012). As a result, PLoS ONE is one of the most frequently included journals in both Community 1 and Community 4. In a journal-based categorization system, articles may be misclassified or not classified at all.

Network analysis can also reveal details about the connections between fields. Figure 6.3 shows a second network we have created in which our communities are represented as nodes and citations from articles in one community to articles in the other communities are represented as edges. The thickness of each edge represents the number of citations and the graph shows the directional strength of the relative connections between communities. For example, the graph suggests that the communication studies community (Community 3) cites many papers in information technology and management (Community 2) but that this relationship is not reciprocated.

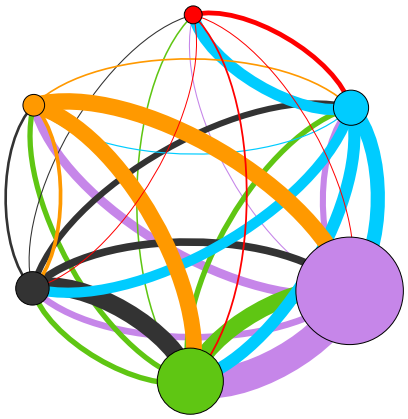

Figure 6.3: Graphical representation of citations between communities using the same mapping described in Table 6.4. The size of the nodes reflects the total number of papers in each community. The thickness of each edge reflects the number of outgoing citations. Edges are directional, and share the color of their source (i.e., citing) community.

Discussion

Like many computational methods, the power of network techniques comes from representing complex relationships in simplified forms. Although elegant and powerful, the network analysis approach is inherently reductive in nature and limited in many ways. What we gain in our ability to analyze millions or billions of individuals comes at the cost of speaking about particular individuals and sub-groups. A second limitation stems from the huge number of relationships that can be represented in graphs. A citation network and a co-citation network, for example, represent different types of connections and these differences might lead an algorithm to identify different communities. As a result, choices about the way that edges and nodes are defined can lead to very different conclusions about the structure of a network or the influence of particular nodes. Network analyses often treat all connections and all nodes as similar in ways that mask important variation.

Network analysis is built on the assumption that knowing about the relationships between individuals in a system is often as important, and sometimes more important, than knowing about the individuals themselves. It inherently recognizes interdependence and the importance of social structures. This perspective comes with a cost, however. The relational structure and interdependence of social networks make it impossible to use traditional statistical methods. SNA practitioners have had to move to more complex modeling

strategies and simulations to test hypotheses.

Text analysis

Social media produces an incredible amount of text, and social media researchers often analyze the content of this text. For example, researchers use ethnographic approaches (Kozinets, 2002) or content analysis (Chew & Eysenbach, 2010) to study the texts of interactions online. Because the amount of text available for analysis is far beyond the ability of any set of researchers to analyze by hand, scholars increasingly turn to computational approaches. Some of these analyses are fairly simple, such as tracking the occurrence of terms related to a topic or psychological construct (Tausczik & Pennebaker, 2010). Others are more complicated, using tools from natural language processing (NLP). NLP includes a range of approaches in which algorithms are applied to texts, such as machine translation, optical character recognition, and part-of-speech tagging. Perhaps the most common use in the social sciences is sentiment analysis, in which the affect of a piece of text is intuited based on the words that are used (Asur & Huberman, 2010). Many of these techniques have applications for social media research.

One natural language processing technique—topic modeling—is used increasingly often in computational social science research. Topic modeling seeks to identify topics automatically within a set of documents. In this sense, topic modeling is analogous to content analysis or other manual forms of document coding and labeling. However, topic models are a completely automated, unsupervised computational method—i.e., topic modeling algorithms do not require any sort of human intervention, such as hand-coded training data or dictionaries of terms. Topic modeling scales well to even very large datasets, and is most usefully applied to large corpora of text where labor-intensive methods like manual coding are simply not an option.

When using the technique, a researcher begins by feeding topic modeling software the texts that she would like to find topics for and by specifying the number of topics to be returned. There are multiple algorithms for identifying topics, but we focus on the most common: *latent Dirichlet allocation* or LDA (Blei, Ng, & Jordan, 2003). The nuts and bolts of how LDA works are complex and beyond the scope of this chapter, but the basic goal is fairly simple: LDA identifies sets of words that are likely to be used together and calls these sets 'topics.' For example, a computer science paper is likely to use words like 'algorithm', 'memory', and 'network.' While a communication ar-

ticle might also use 'network,' it would be much less likely to use 'algorithm' and more likely to use words like 'media' and 'influence.' The other key feature of LDA is that it does not treat documents as belonging to only one topic, but as consisting of a mixture of multiple topics with different degrees of emphasis. For example, an LDA analysis might characterize this chapter as a mixture of computer science and communication (among other topics).

LDA identifies topics inductively from the observed distributions of words in documents. The LDA algorithm looks at all of the words that co-occur within a corpus of documents and assumes that words used in the same document are more likely to be from the same topic. The algorithm then looks across all of the documents and finds the set of topics and topic distributions that would be, in a statistical sense, most likely to produce the observed documents. LDA's output is the set of topics: ranked lists of words likely to be used in documents about each topic, as well as the distribution of topics in each document. DiMaggio, Nag, and Blei (2013) argue that while many aspects of topic modeling are simplistic, many of the assumptions have parallels in sociological and communication theory. Perhaps more importantly, the topics created by LDA frequently correspond to human intuition about how documents should be grouped or classified.

The results of topic models can be used many ways. Our dataset includes 73 publications with the term 'LDA' in their abstracts. Some of these papers use topic models to conduct large-scale content analysis, such as looking at the topics used around health on Twitter (Prier, Smith, Giraud-Carrier, & Hanson, 2011; Ghosh & Guha, 2013). Researchers commonly use topic modeling for prediction and machine learning tasks, such as predicting a user's gender or personality type (Schwartz et al., 2013). Papers in the dataset also use LDA to predict transitions between topics (Wang, Agichtein, & Benzi, 2012), to recommend friends based on similar topic use (Pennacchiotti & Gurumurthy, 2011), and to identify interesting tweets on Twitter (Yang & Rim, 2014).

Our application: Identifying topics in social media research

We apply LDA to the texts of abstracts in our dataset in order to identify topics in social media research. We show how topics are extracted and labeled and then use data on topic distributions to show how the focus of social media research has changed over time. We begin by collecting each of the abstracts for the papers in our sample. Scopus does not include abstract text for 2,801 of the 23,131 articles in our sample. We examined a random sample of the en-

tries with missing abstracts by hand, and found that abstracts for many simply never existed (e.g., articles published in trade journals or books). Other articles had published abstracts, but the text of these abstracts, for reasons that are not clear, were not available through Scopus.[2] We proceed with the 20,330 articles in our sample for which abstract data was available. The average abstract in this dataset is 177 words long, with a max of 1,353 words and a minimum of 5 ("The proceedings contain 15 papers.").

We then remove 'stop words' (common words like 'the,''of,' etc.) and tokenize the documents by breaking them into unigrams and bigrams (one-word and two-word terms). We analyze the data using the Python *LatentDirichletAllocation* module from the *scikit-learn* library (Pedregosa et al., 2011). Choosing the appropriate number of topics to be returned (typically referred to as k) is a matter of some debate and research (e.g., Arun, Suresh, Madhavan, & Murthy, 2010). After experimenting with different values of k, plotting the distribution of topics each time in a way similar to the graphs shown in Figure 6.4, we ultimately set k as twelve. At higher values of k, additional topics only rarely appeared in the abstracts.

Results

Table 6.5 shows the top words for each of the topics discovered by the LDA model, sorted by how common each topic is in our dataset. At this point, researchers typically evaluate the lists of words for coherence and give names to each of the topics. For example, after looking at the words associated with Topic 1 we gave it the name 'Media Use.' Of course, many other names for this topic could be chosen. We might call it 'Facebook research' because it is the only topic which includes the term 'facebook.' Researchers often validate these names by looking at some of the texts which score highest for each topic and subjectively evaluating the appropriateness of the chosen name as a label for those texts. For example, we examined the abstracts of the five papers with the highest value for the 'Media Use' topic and confirmed that we were comfortable claiming that they were examples of research about media use. In this way, topic modeling requires a mixture of both quantitative and qualitative interpretation. The computer provides results, but making sense of those results requires familiarty with the data.

[2]This provides one example of how the details of missing data can be invisible or opaque. It is easy to see how missing data like this could impact research results. For example, if certain disciplines or topics are systematically less likely to include abstracts in Scopus, we will have a skewed representation of the field.

Media Use	Social Network Analysis	Consumer Analsyis	Education	Quantitative Analysis	Information Spread
social	social	media	students	based	twitter
media	data	social	learning	approach	tweets
"social media"	media	"social media"	knowledge	method	time
use	"social media"	new	research	proposed	information
study	information	marketing	education	data	messages
online	users	2015	technology	model	events
facebook	network	business	social	text	public
research	networks	brand	use	images	videos
communication	user	communication	media	results	crisis
public	paper	information	"social media"	media	users
political	web	consumers	design	user	mobile
article	analysis	companies	tools	search	data
findings	based	organizations	technologies	using	event
"use social"	online	management	development	image	location
2014	"social networks"	consumer	digital	topic	used
people	different	customer	2015	propose	2014
new	research	services	student	paper	emergency
results	content	strategies	educational	algorithm	disaster
networking	"social network"	customers	paper	problem	real
using	people	service	project	detection	youtube

Health	Sentiment Analysis	News	HCI	Influence	Methodology
health	content	news	systems	2015	purpose
information	springer	women	information	model	value
use	sentiment	study	privacy	influence	implications
patients	analysis	facebook	papers	al	findings
medical	user	posts	based	intention	limited
care	results	articles	music	et	paper
methods	negative	sexual	security	"et al"	methodology
results	generated	participants	personality	factors	approach
patient	online	page	cloud	perceived	publishing
participants	positive	young	alcohol	smoking	"publishing limited"
using	study	men	model	tobacco	emerald
related	opinion	stories	online	"2015 elsevier"	"emerald group"
internet	reviews	gender	include	satisfaction	"group publishing"
reported	comments	journalists	using	theory	practical
conclusions	switzerland	online	software	structural	originality
support	opinions	significantly	management	variables	design
used	"sentiment analysis"	female	proceedings	intentions	research
clinical	quality	group	discussed	equation	group
healthcare	users	exposure	contain	study	"originality value"
risk	media	pages	analysis	addiction	"methodology approach"

Table 6.5: Top 20 terms for each topic. Topics are presented in the order of their frequency in the corpus of abstracts.

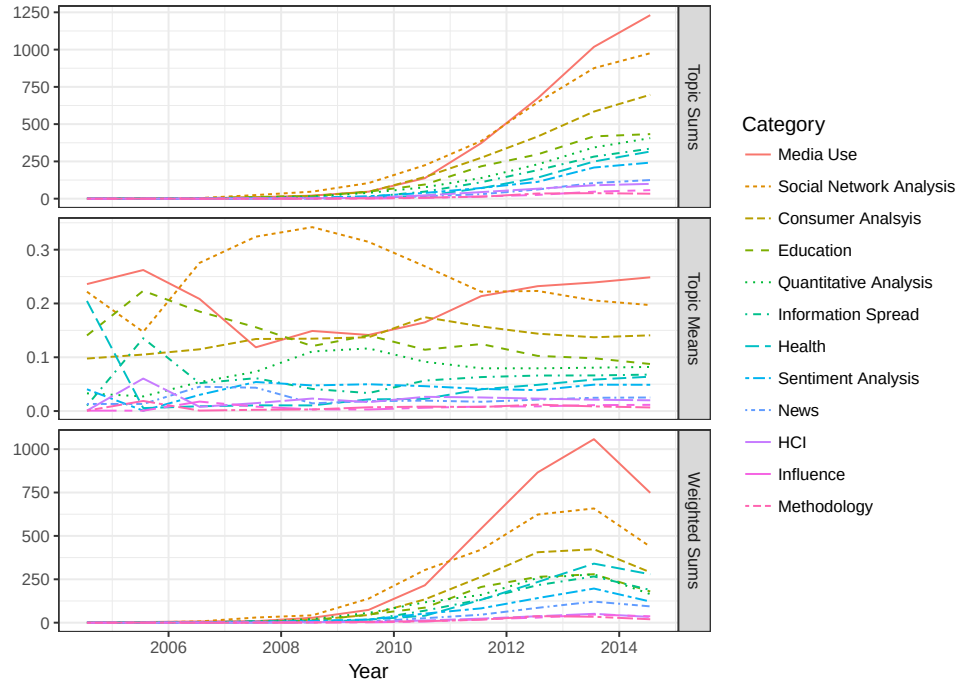

Figure 6.4: Statistics from our LDA analysis, over time. The top panel shows topic sums which capture the amount that each topic is used in abstracts, by year. The middle panel shows topic means which are the average amount that each topic is used in a given abstract. The bottom panel shows the amount that each topic is used in abstracts, by year, weighted by citation count.

The top panel of Figure 6.4 shows how the distribution of topics identified by LDA in our analysis has changed over time. The LDA algorithm gives each abstract a probability distribution over each of the topics, such that it sums to 1 (e.g., a given abstract may be 80% 'Social Network Analysis,' 20% 'Education,' and 0% everything else). To construct Figure 6.4, we sum these percentages for all of the documents published in each year and plot the resulting prevalence of each topic over time.[3]

The figures provide insight into the history and trajectory of social media research. Looking at the top figure, it appears that the 'Social Network Analysis' topic was the early leader in publishing on social media, but was overtaken by the 'Media Use' topic around 2012. This pattern is even more apparent when we look at the mean amount that each topic was used each year (the middle panel of Figure 6.4). In the bottom panel, we take a third

[3]More complex approaches such as dynamic LDA (Blei & Lafferty, 2006) are often better suited to identify the temporal evolution of topics.

look at this data by weighting the topics used in each paper by the log of the number of citations that the paper received. This final statistic characterizes how influential each topic has been. The overall story is similar, although we see that the 'Health' topic and the 'Media Use' topic are more influential than the non-weighted figures suggest.

Discussion

Some of the strengths of topic modeling become apparent when we compare these LDA-based analyses with the distribution of papers by discipline that we created earlier (Figure 6.1). In our earlier attempt, we relied on the categories that Scopus provided and found that early interest in social media was driven by computer science and information systems researchers. Through topic modeling, we learn that these researchers engaged in social network analysis (rather than interface design, for example). While some of our topics match up well with the disciplines identified by Scopus, a few are more broad (e.g., 'Media Use') and most are more narrow (e.g., 'Sentiment Analysis'). This analysis provides a richer sense of the topics of interest to social media researchers. Finally, these topics emerged inductively without any need for explicit coding, such as classifying journals into disciplines. This final feature is a major benefit in social media research where text is rarely categorized for researchers ahead of time.

Topic modeling provides an intuitive, approachable way of doing large-scale text analysis. Its outputs can be understandable and theory-generating. The inductive creation of topics has advantages over traditional content analysis or 'supervised' computational methods that require researchers to define labels or categories of interest ahead of time. While topic models clearly lack the nuance and depth of understanding that human coders bring to texts, the method allows researchers to analyze datasets at a scale and granularity that would take a huge amount of resources to code manually.

There are, of course, limitations to topic modeling. Many of LDA's limitations have analogues in manual coding. One we have already mentioned is that researchers must choose the number of topics without any clear rules about how to do so. Although a similar problem exists in content analysis, the merging and splitting of topics can be done more intuitively and intentionally when using traditional methods. An additional limitation is that topic modeling tends to work best with many long documents. This can represent a stumbling block for researchers with datasets of short social media posts or

comments; in these cases posts can be aggregated by user or by page to produce meaningful topics. The scope of documents can also affect the results of topic models. If, in addition to using abstracts about 'social media,' we had also included abstracts containing the term 'gene splicing,' our twelve topics would be divided between the two fields and each topic would be less granular. To recover topics similar to those we report here, we would have to increase the number of topics created.

As with network analysis, a goal of LDA is to distill large, messy, and noisy data down to much simpler representations in order to find patterns. Such simplification will always entail ignoring some part of what is going on. Luckily, human coders and LDA have complementary advantages and disadvantages in this regard. Computational methods do not understand which text is more or less important. Humans are good at seeing the meaning and importance of topics, but may suffer from cognitive biases and miss out on topics that are less salient (DiMaggio et al., 2013). Topic models work best when they are interpreted by researchers with a rich understanding of the texts and contexts under investigation.

Predicting citation

A final computational technique is statistical prediction. Statistical prediction can come in handy in situations where researchers have a great deal of data, including measures of an important, well-defined outcome they care about, but little in the way of prior literature or theory to guide analysis. Prediction has become a mainstream computational approach that encompasses a number of specific statistical techniques including classification, cross validation, and machine learning (also known as statistical learning) methods (Tibshirani, 1996). Arguably made most famous by Nate Silver (2015), who uses the technique to predict elections and sporting event outcomes, prediction increasingly colors public discourse about current events (Domingos, 2015).

There are many approaches to prediction. We focus on regression-based prediction because it offers a reasonably straightforward workflow. Begin by breaking a dataset into two random subsets: a large subset used as 'training' data and a small subset as 'holdout' or 'test' data. Next, use the training data to construct a regression model of a given outcome (dependent variable) that incorporates a set of features (independent variables) that might explain variations in the outcome. Apply statistical model selection techniques to determine the best weights (coefficients) to apply to the variables. Evaluate the per-

formance of the model by seeing how accurately it can predict the outcome on the test data. After selecting an appropriate model, assess and interpret the items that most strongly predict the outcome. One can even compare the performance of different or nested sets of features by repeating these steps with multiple groups of independent variables.

Interpreting the results of statistical prediction can be less clear-cut. The term 'prediction' suggests a deep knowledge of a complex social process and the factors that determine a particular outcome. However, statistical prediction often proves more suitable for exploratory analysis where causal mechanisms and processes are poorly understood. We demonstrate this in the following example that predicts whether or not papers in our dataset get cited during the period of data collection. In particular, we try to find out whether textual features of the abstracts can help explain citation outcomes. Our approach follows that used by Mitra and Gilbert (2014), who sought to understand what textual features of Kickstarter projects predicted whether or not projects were funded.

Our application: Predicting paper citation

We use multiple attributes of the papers in our dataset, including text of their abstracts, to predict citations. About 42% of the papers (9,713 out of 23,131) received one or more citations ($\mu = 3$; $\sigma = 19$). Can textual features of the abstracts explain which papers receive citations? What about other attributes, such as the publication venue or subject area? A prediction analysis can help evaluate these competing alternatives.

To begin, we generate a large set of features for each paper from the Scopus data. Our measures include the year, month, and language of publication as well as the number of citations each paper contains to prior work. We also include the modal country of origin of the authors as well as the affiliation of the first author. Finally, we include the publication venue and publication subject area as provided by Scopus. Then, we build the textual features by taking all of the abstracts and moving them through the following sequence of steps similar to those we took when performing LDA: we lowercase all the words; remove all stop words; and create uni-, bi-, and tri-grams.

We also apply some inclusion criteria to both papers and features. To avoid subject-specific jargon, we draw features only from those terms that appear across at least 30 different subject areas. To avoid spurious results, we also exclude papers that fall into unique categories. Specifically, we remove papers

which are the only publications in a given language, journal, or subject area. These sorts of unique cases can cause problems in the context of prediction tasks because they may predict certain outcomes perfectly. As a result, it is often better to focus on datasets and measures that are less 'sparse' (i.e., characterized by rare, one-off observations). Once we drop the 8,494 papers that do not meet these criteria, we are left with 14,126 papers.

We predict the dichotomous outcome variable *cited*, which indicates whether a paper received any citations during the period covered by our dataset (2004-2016). We use a method of *penalized logistic regression* called the least absolute shrinkage and selection operator (also known as the *Lasso*) to do the prediction work. Although, the technical details of Lasso models lie beyond the scope of this chapter, it, and other penalized regression models work well on data where many of the variables have nearly identical values (sometimes called collinear variables because they would sit right around the same line if you plotted them) and/or many zero values (this is also called 'sparse' data) (Friedman, Hastie, & Tibshirani, 2010; James et al., 2013). In both of these situations, some measures are redundant; the Lasso uses clever math to pick which of those measures should go into your final model and which ones should be, in effect, left out.[4] The results of a Lasso model are thus more computationally tractable and easier to interpret.

We use a common statistical technique called cross-validation to validate our models. Cross-validation helps solve another statistical problem that can undermine the results of predictive analysis. Imagine fitting an ordinary least squares regression model on a dataset to generate a set of parameter estimates reflecting the relationships between a set of independent variables and some outcome. The model results provide a set of weights (the coefficients) that represent the strength of the relationships between each predictor and the outcome. Because of the way regression works (and because this is a hypothetical example and we can assume data that does not violate the assumptions of our model), the model weights are the best, linear, unbiased estimators of those relationships. In other words, the regression model fits the data as well as possible. However, nothing about fitting this one model ensures that the same regression weights will provide the best fit for some new data from the same population that the model has not seen. A model may be overfit if it excellently predicts the dataset it was fitted on but poorly predicts new

[4]To put things a little more technically, a fitted Lasso model *selects* the optimal set of variables that should have coefficient values greater than zero and *shrinks* the rest of the coefficients to zero without sacrificing goodness of fit (Tibshirani, 1996).

data. Overfitting in this way is a common concern in statistical prediction. Cross-validation addresses this overfitting problem. First, the training data is split into equal-sized groups (typically 10). Different model specifications are tested by iteratively training them on all but one of the groups, and testing how well they predict the final group. The specification that has the lowest average error is then used on the full training data to estimate coefficients.[5] This approach ensures that the resulting models not only fit the data that we have, but that they are likely to predict the outcomes for new, unobserved results. For each model, we report the mean error rate from the cross-validation run which produced the best fit.

Our analysis proceeds in multiple stages corresponding to the different types of measures we use to predict citation outcomes. We start by estimating a model that includes only the features that correspond to paper and author-level attributes (year, month, and language of publication, modal author country). We then add information about the first author's affiliation. Next, we include predictors that have more to do with research topic and field-level variations (publication venue and subject area). Finally, we include the textual features (terms) from the abstracts.

Results

Table 6.6 summarizes the results of our prediction models. We include goodness-of-fit statistics and prediction error rates for each model as we add more features. A 'better' model will fit the data more closely (i.e., it will explain a larger percentage of the deviance) and produce a lower error rate. We also include a supplementary error rate calculated against the 'holdout' data created from a random subset of 10% of the original dataset that was not used in any of our models. An intuitive way to think about the error rate is to imagine it as the percentage of unobserved papers for which the model will correctly predict whether or not it receives any citations. The two error rate statistics are just this same percentage calculated on different sets of unobserved papers. Unlike a normal regression analysis, we do not report or interpret the full battery of coefficients, standard errors, t-statistics, or p-values. In part, we do not report this information because the results of these models are unwieldy – each model has over 2,000 predictors and most of those predictors have coefficients of zero! Additionally, unlike traditional regression results,

[5] For our Lasso models, cross-validation was used to select λ, a parameter that tells the model how quickly to shrink variable coefficients. We include this information for those of you who want to try this on your own or figure out the details of our statistical code.

Model	N features	Deviance (%)	CV error (%)	Hold-back error (%)
Controls	98	7	38	37
+ Affiliation	1909	23	39	37
+ Subject	2096	28	37	34
+ Venue	3902	55	34	30
+ Terms	4411	72	29	27

Table 6.6: Summary of fitted models predicting citation. The 'Model' column describes which features were included. The N features column shows the number of features included in the prediction. 'Deviance' summarizes the goodness of fit as a percentage of the total deviance accounted for by the model. 'CV error' (cross-validation error) reports the prediction error rates of each model in the cross-validation procedure conducted as part of the parameter estimation process. 'Holdout error' shows the prediction error on a random 10% subset of the original dataset not included in any of the model estimation procedures.

coefficient interpretation and null hypothesis testing with predictive models remain challenging (for reasons that lie beyond the scope of this chapter). Instead, we focus on interpreting the relative performance of each set of features. After we have done this, we refer to the largest coefficients to help add nuance to our interpretation.

The results reveal that all of the features improve the goodness of fit, but not necessarily the predictive performance of the models. As a baseline, our controls-only model has a 37% classification error on the holdout sample. This level of precision barely improves with the addition of both the author affiliation and subject area features. We observe substantial improvements in the prediction performance when the models include the publication venue features and the abstract text terms. When it comes to research about social media, it appears that venue and textual content are the most informative features for predicting whether or not articles get cited.

To understand these results more deeply, we explore the non-zero coefficient estimates for the best-fitting iteration of the full model. Recall that the Lasso estimation procedure returns coefficients for a subset of the parameters that produce the best fit and shrinks the other coefficients to zero. While it does not make sense to interpret the coefficients in the same way as traditional regression, the non-zero coefficients indicate what features the model identified as the most important predictors of the outcome. First, we note that among the 1,482 features with non-zero coefficients, only 2% are control

Feature	Type	Coefficient
Multiple Sclerosis Journal	venue	-2.969
Nature Communications	venue	2.871
Journal of Information Technology	venue	2.762
CrossTalk	venue	2.543
21	term	-2.472
NICTA Victoria Research Laboratory	affiliation	-2.260
The Department of Education, Sookmyung Women's University	affiliation	-2.196
20th ITS World Congress Tokyo 2013	venue	-2.191
Electronics and Communications in Japan	venue	-2.085
British Journal of Nursing	venue	2.077

Table 6.7: Feature, variable type, and beta value for top 10 non-zero coefficients estimated by the best fitting model with all features included. Note that the outcome is coded such that positive coefficients indicate features that positively predict the observed outcome of interest (getting cited) while negative coefficients indicate features that negatively predict the outcome.

measures (country, language, month, and year of publication). Similarly, 3% are subject features. In contrast, 15% are affiliation features, 34% are venue features, and a whopping 44% are textual terms. Once again, we find that publication venue and textual terms do the most to explain which works receive citations.

Closer scrutiny of the features with the largest coefficients adds further nuance to this interpretation. Table 6.7 shows the ten features with the largest coefficients in terms of absolute value. The Lasso model identified these coefficients as the most informative predictors of whether or not papers in our dataset get cited. Here we see that the majority of these most predictive features are publication venues. The pattern holds across the 100 features with the largest coefficients, of which 75 are publication venues and only 2 are textual terms from the abstracts. In other words, variations in publication venue predict which work gets cited more than any other type of feature.

Discussion

The results of our prediction models suggest that two types of features – publication venue and textual terms – do the most to explain whether or not papers on social media get cited. Both types of features substantially improve model fit and reduce predictive error in ten-fold cross-validation as well as on a hold-out sub-sample of the original dataset. However, the venue features appear

to have a much stronger relationship to our outcome (citation), with the vast majority of the most influential features in the model coming from the venue data (75 of the 100 largest coefficients).

As we said at the outset of this section, statistical prediction offers an exploratory, data-driven, and inductive approach. Based on these findings, we conclude that the venue where research on social media gets published better predicts whether that work gets cited than the other features in our dataset. Textual terms used in abstracts help to explain citation outcomes across the dataset, but the relationship between textual terms and citation only becomes salient in aggregate. On their own, hardly any of the textual terms approach the predictive power of the venue features. Features such as author affiliation and paper-level features like language or authors' country provide less explanatory power overall.

The approach has several important limitations. Most important, statistical prediction only generates 'predictions' in a fairly narrow, statistical sense. Language of prediction often sounds like the language of causality and inferring process, but these methods do not guarantee that anything being studied is causal or explanatory in terms of mechanisms. We do not claim that a paper's publication venue or the phrases in its abstract *cause* people to cite it. Rather, we think these attributes of a paper likely index specific qualities of an article that are linked to citation outcomes. Just because something is predictive does not mean it is deterministic or causal. We also note that the sort of machine learning approach we demonstrate here does not support the types of inferences commonly made with frequentist null hypothesis tests (the sort that lead to p-values and stars next to 'significant' variables in a regression model). Instead, the interpretation of learning models rests on looking closely at model summary statistics, objective performance metrics (such as error rates), and qualitative exploration of model results.

Conclusion

In this chapter, we have described computational social scientific analysis of social media by walking through a series of example analyses. We began with the process of collecting a dataset of bibliographic information on social media scholarship using a web API similar to those provided by most social media platforms. We then subjected this dataset to three of the mostly widely used computational techniques: network analysis, topic modeling, and statistical prediction. Most empirical studies would employ a single, theoretically-

motivated analytic approach, but we compromised depth in order to illustrate the diversity of computational research methodologies available. As we have shown, each approach has distinct strengths and limitations.

We believe our examples paint a realistic picture of what is involved in typical computational social media research. However, these analyses remain limited in scope and idiosyncratic in nature. For example, there are popular computational methods we did not cover in this chapter. Obvious omissions include other forms of machine learning, such as decision trees and collaborative filtering (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), as well as simulation-based techniques such as agent-based modeling (Macy & Willer, 2002; Wilensky & Rand, 2015).

Despite our diffuse approach, we report interesting substantive findings about the history and state of social media research. We discovered a number of diverse communities studying social media. We used different tools to identify these communities, including the categories provided by Scopus, the results of a community detection algorithm applied to the citation network, and the topics identified by topic modeling. Each analysis provided a slightly different picture of social media research. We learned that the study of social media related to media use and medical research is on the rise. We also learned that social network research was influential at the early stages of social media research, but that it is not as influential in the citation network. All of these findings are complicated by our final finding that subject area is not as good a predictor of whether a paper will receive a citation as the publication venue and the terms used in the abstract.

In the process of describing our analyses, we tried to point to many of the limitations of computational research methods. Although computational methods and the promise of 'big data' elicit excitement, this hype can obscure the fact that large datasets and fast computers do nothing to obviate the fundamentals of high quality social science: researchers must understand their empirical settings, design studies with care, operationalize concepts in ways that are valid and honest, take steps to ensure that their findings generalize, and ask tough questions about the substantive impacts of observed relationships. These tenets extend to computational research as well.

Other challenges go beyond methodological limitations. Researchers working with passively collected data generated by social media can face complex issues around the ethics of privacy and consent as well as the technical and legal restrictions on automated data collection. Computational analyses of

social media often involve datasets gathered without the sort of active consent considered standard in other arenas of social scientific inquiry. In some cases, data is not public and researchers access it through private agreements or employment arrangements with companies that own platforms or proprietary databases. In others, researchers obtain social media data from public or semi-public sources, but the individuals creating the data may not consider their words or actions public and may not even be aware that their participation generates durable digital traces (boyd & Crawford, 2012). A number of studies have been criticized for releasing information that researchers considered public, but which users did not (Zimmer, 2016). In other cases, researchers pursuing legitimate social inquiry have become the target of companies or state prosecutors who selectively seek to enforce terms of service agreements or invoke broad laws such as the federal Computer Fraud and Abuse Act (CFAA).[6]

We advise computational researchers to take a cautious and adaptive approach to these issues. Existing mechanisms such as Institutional Review Boards and federal laws have been slow to adjust to the realities of online research. In many cases, the authority and resources to anticipate, monitor, or police irresponsible behaviors threaten to impose unduly cumbersome restrictions. In other cases, review boards' policies greenlight research that seems deeply problematic. We believe researchers must think carefully about the specific implications of releasing specific datasets. In particular, we encourage abundant caution and public consultation before disseminating anything resembling personal information about individual social media system users. Irresponsible scholarship harms both subjects and reviewers and undermines the public trust scholars need to pursue their work.

At the same time, we remain excited and optimistic about the future of computational studies of social media. As we have shown, the potential benefits of computational methods are numerous. Trace data can capture behaviors that are often difficult to observe in labs and that went unrecorded in offline interactions. Large datasets allow researchers to measure real effects obscured by large amounts of variation, and to make excellent predictions using relatively simple models. These new tools and new datasets provide a real opportunity to advance our understanding of the world. Such opportunities

[6]See Sandvig's (2016) blogpost, "Why I am Suing the Government," for a thoughtful argument against the incredibly vague and broad scope of the CFAA as well as a cautionary tale for those who write software to conduct bulk downloads of public website data for research purposes.

should not be undermined by overly-broad laws or alarmist concerns.

Finally, much of computational social science, including this chapter, is data-focused rather than theory-focused. We would encourage others to do as we say, and not as we do. The great promise of computational social science is the opportunity to test and advance social science theory. We hope that readers of this chapter will think about whether there are theories they are interested in which might benefit from a computational approach. We urge readers with a stronger background in theory to consider learning the tools to conduct these types of analyses and to collaborate with technically minded colleagues.

Reproducible research

Computational research methods also have the important benefit of being extraordinarily reproducible and replicable (Stodden, Guo, & Ma, 2013). Unlike many other forms of social research, a computational researcher can theoretically use web APIs to collect a dataset identical to one used in a previous study. Even when API limits or other factors prohibit creating an identical dataset, researchers can work to make data available alongside the code they use for their analysis, allowing others to re-run the study and assess the results directly. Making code and data available also means that others can analyze and critique it. This can create uncomfortable situations, but we feel that such situations serve the long-term interests of society and scholarly integrity. Although not every computational researcher shares their code (Stodden et al., 2013) there are movements to encourage or require this (LeVeque, Mitchell, & Stodden, 2012; Stodden et al., 2013; Bollen et al., 2015).

We have tried to follow emerging best practices with regards to reproducibility in this chapter. We have released an online copy of all of the code that we used to create this chapter. By making our code available, we hope to make our unstated assumptions and decisions visible. By looking at our code, you might find errors or omissions which can be corrected in subsequent work. By releasing our code and data, we also hope that others can learn from and build on our work. For example, a reader with access to a server and some knowledge of the Python and R programming languages should be able to build a more up-to-date version of our dataset years from now. Another reader might create a similar bibliographic analysis of another field. By using our code, this reader should able to produce results, tables, and figures like those in this chapter. Data repositories, such as the Harvard Dataverse, make

storing and sharing data simple and inexpensive. When thinking of the opportunities for openness, transparency, and collaboration, we are inspired by the potential of computational research methods for social media. We hope that our overview, data, and code can facilitate more of this type of work.

Online supplements

All of the code used to generate our dataset, to complete our analyses, and even to produce the text of this chapter, is available for download on the following public website: https://communitydata.cc/social-media-chapter/

Because the Scopus dataset is constantly being updated and changed, reproducing the precise numbers and graphs in this chapter requires access to a copy of the dataset we collected from Scopus in 2016. Unfortunately, like many social media websites, the terms of use for the Scopus APIs prohibit the re-publication of data collected from their database. However, they did allow us to create a private, access-controlled, replication dataset in the Harvard Dataverse archive at the following URL: http://dx.doi.org/10.7910/DVN/W31PH5. Upon request, we will grant access to this dataset to any researchers interested in reproducing our analyses.

References

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010, June 21). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining* (pp. 391–402). Lecture Notes in Computer Science. Pacific-asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-13657-3_43

Asur, S. & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT)* (Vol. 1, pp. 492–499). 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT). doi:10.1109/WI-IAT.2010.63

Barabási, A.-L. & Albert, R. (1999, October 15). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. doi:10.1126/science.286.5439.509

Bastian, M., Heymann, S., Jacomy, M., et al. (2009). Gephi: An open source software for exploring and manipulating networks. *ICWSM*, *8*, 361–

362. Retrieved July 20, 2016, from http://www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154/1009/

Binfield, P. (2012, February 29). *PLoS ONE and the rise of the open access Mega-Journal*. The 5th SPARC Japan Seminar 2011, National Institute for Informatics. Retrieved July 20, 2016, from http://www.nii.ac.jp/sparc/en/event/2011/pdf/20120229_doc3_binfield.pdf

Blei, D. M. (2012, April). Probabilistic topic models. *Commun. ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120). ACM. Retrieved April 21, 2016, from http://dl.acm.org/citation.cfm?id=1143859

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. Retrieved December 3, 2015, from http://dl.acm.org/citation.cfm?id=944937

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Krosnick, J. A., Olds, J. L., & Dean, H. (2015, May). *Social, behavioral, and economic sciences perspectives on robust and reliable science*. National Science Foundation. Retrieved from http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

boyd, d. & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Chew, C. & Eysenbach, G. (2010, November 29). Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLOS ONE*, *5*(11), e14118. doi:10.1371/journal.pone.0014118

DiMaggio, P., Nag, M., & Blei, D. (2013, December). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, *41*(6), 570–606. doi:10.1016/j.poetic.2013.08.004

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York, New York: Basic Books.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1–22. Retrieved July 20, 2016, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

Fruchterman, T. M. J. & Reingold, E. M. (1991, November 1). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. doi:10.1002/spe.4380211102

Ghosh, D. (& Guha, R. (2013, March 1). What are we 'tweeting' about obesity? mapping tweets with topic modeling and geographic information

system. *Cartography and Geographic Information Science*, *40*(2), 90–102. doi:10.1080/15230406.2013.776210

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL: Insights from a connected world*. Burlington, Massachusetts: Morgan Kaufmann.

Hargittai, E. (2015, May 1). Is bigger always better? potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63–76. doi:10.1177/0002716215570866

Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., Simoes, A., & Yildirim, M. A. (2014, January 17). *The atlas of economic complexity: Mapping paths to prosperity*. MIT Press.

Hood, W. W. & Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics*, *52*(2), 291–314. doi:10.1023/A:1017919924342

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. New York: Springer.

Kessler, M. M. (1963, January 1). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. doi:10.1002/asi.5090140103

Kovács, A., Looy, B. V., & Cassiman, B. (2015, June 20). Exploring the scope of open innovation: A bibliometric review of a decade of research. *Scientometrics*, *104*(3), 951–983. doi:10.1007/s11192-015-1628-0

Kozinets, R. V. (2002, February 1). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research*, *39*(1), 61–72. doi:10.1509/jmkr.39.1.61.18935

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009, February 6). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Leskovec, J. & Krevl, A. (2014, June). *SNAP datasets: Stanford large network dataset collection*. Retrieved from http://snap.stanford.edu/data

LeVeque, R. J., Mitchell, I. M., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science and Engineering*, *14*(4), 13–17.

Macy, M. W. & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143–166. Retrieved July 20, 2016, from http://www.jstor.org/stable/3069238

Merton, R. K. (1968). The matthew effect in science. *Science*, *159*(3810), 56–63. Retrieved September 27, 2014, from http://www.unc.edu/ ~ fbaum/teaching/PLSC541_Fall06/Merton_Science_1968.pdf

Mitra, T. & Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on computer supported cooperative work & social computing* (pp. 49–

61). CSCW '14. New York, NY, USA: ACM. doi:10.1145/2531602.2531656

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011, October). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. bibtex: pedregosa_scikit-learn:_2011. Retrieved June 7, 2016, from http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pennacchiotti, M. & Gurumurthy, S. (2011). Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on world wide web* (pp. 101–102). WWW '11. New York, NY, USA: ACM. doi:10.1145/1963192.1963244

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). 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* (pp. 18–25). SBP'11. Berlin, Heidelberg: Springer-Verlag. Retrieved July 19, 2016, from http://dl.acm.org/citation.cfm?id=1964698.1964702

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on computer supported cooperative work* (pp. 175–186). CSCW '94. New York, NY, USA: ACM. doi:10.1145/192844.192905

Rosvall, M. [M.], Axelsson, D., & Bergstrom, C. T. (2010, April 17). The map equation. *The European Physical Journal Special Topics*, *178*(1), 13–23. doi:10.1140/epjst/e2010-01179-1

Rosvall, M. [Martin] & Bergstrom, C. T. (2008, January 29). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. doi:10.1073/pnas.0706851105

Sandvig, C. (2016, July 1). Why i am suing the government [Social media collective research blog]. Retrieved October 23, 2016, from https://socialmediacollective.org/2016/07/01/why-i-am-suing-the-government/

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013, September 25). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, *8*(9), e73791. doi:10.1371/journal.pone.0073791

Silver, N. (2015). *The signal and the noise: Why so many predictions fail–but some don't*. New York, New York: Penguin Books.

Small, H. (1973, July 1). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. doi:10.1002/asi.4630240406

Stodden, V., Guo, P., & Ma, Z. (2013, June 21). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLOS ONE*, *8*(6), e67111. doi:10.1371/journal.pone.0067111

Šubelj, L., Eck, N. J. v., & Waltman, L. (2016, April 28). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE*, *11*(4), e0154404. doi:10.1371/journal.pone.0154404

Tausczik, Y. R. & Pennebaker, J. W. (2010, March 1). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. doi:10.1177/0261927X09351676

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved July 20, 2016, from http://www.jstor.org/stable/2346178

Wang, Y., Agichtein, E., & Benzi, M. (2012). TM-LDA: Efficient online modeling of latent topic transitions in social media. (p. 123). ACM Press. doi:10.1145/2339530.2339552

Wasserman, S. & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.

Wilensky, U. & Rand, W. (2015). *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*. Cambridge, Massachusetts: MIT Press.

Yang, M.-C. & Rim, H.-C. (2014, July). Identifying interesting twitter contents using topical analysis. *Expert Syst. Appl. 41*(9), 4330–4336. doi:10.1016/j.eswa.2013.12.051

Zimmer, M. (2016, May 14). OkCupid study reveals the perils of big-data science. *WIRED*. Retrieved August 31, 2016, from https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/