1 Why Social Media Matters to the Social Sciences

In 1989, millions of protesters filled the streets of China, with some estimates putting the peak number in the hundreds of millions, though it is unlikely that true statistics will ever be known. Our minute-by-minute knowledge of those last days, and of the massacre in Tiananmen Square at their climax, is patchy at best. And what little understanding we do have is due largely to the battalion of foreign journalists who happened to be in Beijing at that exact time to cover a Sino-Soviet summit.

Thirty years later, such a paucity of preserved information is unimaginable outside of a handful of increasingly isolated cases, as protests in the most authoritarian of states are accompanied by a cavalcade of tweets, blogs, social media posts, and amateur photography. The Arab Spring and Color Revolutions highlighted this phenomenon: events tracked to the minute, electronic coordination of opposition action, publication of atrocity in real time, the names and faces of both martyrs and villains universally publicized. The Internet seems toxic to the old authoritarian standbys of social atomization and persistent secrecy.

Consider the famous photographs associated with Tiananmen Square and those from Tahrir Square. The former are grainy, taken from a vast distance at a high angle. The famous "tank man" in the iconic photograph is grabbed and escorted away moments later – whether by desperate friends or state security we have never conclusively learned – and his identity remains unknown to this day. One of the most iconic photographs in history features a man whose name we do not know, and only exists in the historical record at all because of the chance presence of foreign journalists. The world didn't see these pictures for quite some time, as they were shot through telescopic lenses peeking through closed curtains, the film hurriedly hidden inside toilet tanks minutes before the arrival of state security, and then smuggled out of the country days later.

Contrast that with the images taken in Tahrir Square during the mass uprising in Egypt amid the Arab Spring. Those photos are high resolution, perfectly crisp, and taken at street level. The photos were posted online within moments of being taken, and circulated around the world on millions of screens small and large. The famous image of a protester standing alone with upraised sign before the might of the state is in this case not an unknown revolutionary, but

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

in fact was instantly identified. You could follow him on Twitter and read his blog within moments of seeing him flicker across your screen.

More information is published by more people today than at any time in human history. The era of mass consumption of media has given way to an era of mass production. Numbers confound easy comprehension, and give way to colorful illustration: The amount of text posted to Twitter every day is the equivalent of some 8 000 copies of *War and Peace*, more photos are uploaded to Facebook every day than exist in all of human history prior to 1990. It's not that thousands of great novels are being written every day, but that a source of human communication is now being recorded, where before it was lost to historical entropy. Internet communications, recorded in an infinite proliferation of magnetic bits, are enshrining the low background noise of human society: the diaries, snippets of conversation, personal letters, and oral histories that once faded from records, but are now archived away digitally.

This is the key challenge of social media as "big data": it is not just that it represents *more* data than we have ever had to deal with before, but that it represents *different* data than we have had access to. Quantity, as Stalin might have said, is a quality all its own.

1.1 The Theory of Why It Matters

Two schools of thought have developed around the effect of the Internet and social media upon society. Some insist that the effect these technologies have is but incremental, building on print, telephones, radio, and television. Others argue that the information revolution is truly a *revolution*, a seismic shift in the balance of power between peoples and states comparable only to the fruition of Gutenberg in the nineteenth century. The reality, as many scholars have argued over the last decade, rests somewhere between the two extremes.

A slightly tangential body of work provides perhaps the best model for understanding why social media is important to us as social scientists: the literature of how mass literacy transformed society in the nineteenth century. Gellner (and to a lesser degree Anderson) argued that states needed to encourage mass literacy (and the social changes it wrought) in order to industrialize, but then reaped the consequences of increased public capacities for collective action that culminated in nationalism (Anderson, 1983; Gellner, 1983). Gellner notes that mass literacy fundamentally changed the structure of society during the course of modernization. Preliterate societies were highly stratified, the population of each strata non-interchangeable with others. State-sponsored education

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

flattened the postliterate societal structure, churning out masses of technically equal citizens, interchangeable parts in the machinery of industry. The political and social implications of this change were profound, giving birth to the age of mass movements. The economic motivation of elites in sponsoring mass literacy had the unintended consequence of the whirlwind of mass consciousness.

In the modern setting, we see a parallel paradox. Power-holding elites exist in a tension between wanting to open up communication (in order to assess public support, efficiently set placating policy, and reap the benefits of globalized trade that depends on these new communications technologies) while simultaneously wanting to repress communication in order to prevent the organization of opposition. The pressure of the International Telecommunications Union during the 1990s to open up telecoms industries globally was premised on promises of the economic benefits that it would bring regimes otherwise disinclined toward openness or liberalization. And similar to mass literacy's unintended consequences, the opening of telecommunications laid the groundwork for worldwide internet access. While it is hopelessly utopian to suggest that the Internet is uncensorable or a magic guaranteer of free speech in the face of authoritarianism, it has undeniably altered the capacity of populations to communicate. Communication, the flow of information, is the heart blood of society. It is what determines how and why our societies organize, who ends up with power, what our social structure looks like. And so something that drastically changes the way communication works, inevitably alters society itself.

In preliterate societies, information could cheaply flow within a particular social stratum, but crossing strata was more expensive. This had two effects. First, social movements in preliterate societies tended to be intrastrata. Second, the flow of information between ruler and ruled was more limited than that within strata. The natural governmental equilibrium during this period would be despotic, because the poor flow of information between strata would make interstrata collective action less likely. The strongest individual stratum could maintain control by force over the other strata, and because of the high cost of interstrata collective action, a noncoercive equilibrium was unlikely.

By collapsing the boundaries between strata, the advent of mass literacy altered the flow of information in society. Social movements could draw from the bulk of the population, leading variously to democratic or authoritarian outcomes. The key feature of this era was that while receiving information became universal and cheap, the transmission of information, while cheaper than previously, was still proportionately much more expensive than receiving. This era is characterized by universal reception of information, but with the transmission of information centralized in a small number of power holders. While social movements could become more massive, they also became less

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

precise. In Tilly's many works on contentious politics and social movements throughout history, he identifies in particular the way that repertoires of social action evolved with the changing nature of the state during this period (Tilly, 2002, 2003).

The spread of the Internet and new communications technologies changes the flow of information again, and in doing so changes the resulting solutions to collective action problems. While receiving information gets even cheaper with widespread internet access, transmitting information also becomes cheap. In a general sense, the effect of modern communications is to effectively eliminate the transaction costs of both receiving and sending information.

To derive an understanding of how the Internet changes society, it is first necessary to think through how exactly the Internet differs from previous iterations of communications technologies. One way to compare different communications technologies is by comparing the relative costs to send and receive information, and how that cost scales to larger numbers of individuals with each technology. Figure 1.1 displays the number of people being reached by

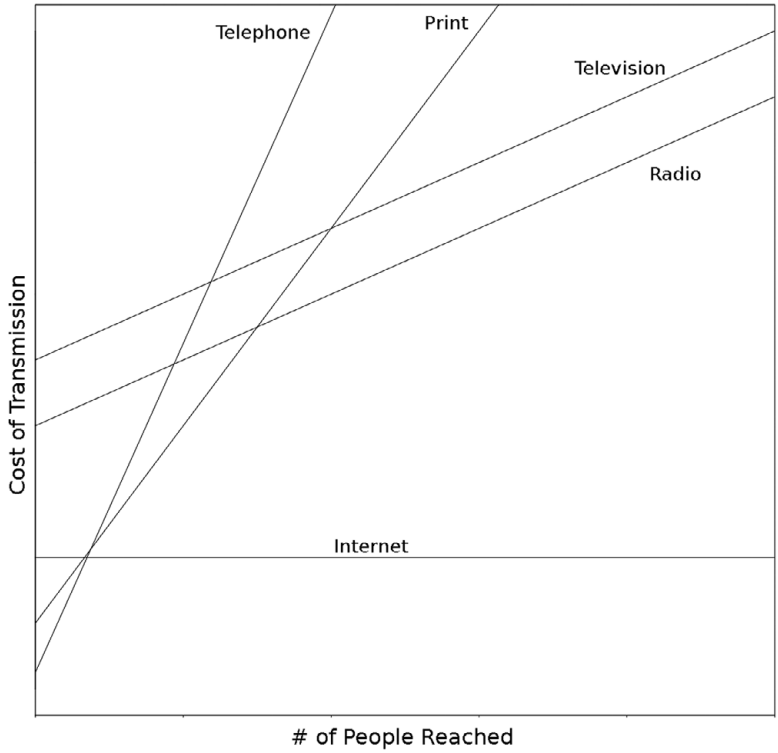

Figure 1.1 Cost of transmission as a function of number of recipients

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

a particular transmission on the x-axis, and the cost required to do so by the transmitting individual on the y-axis. The cost curves of each technology demonstrate both the fixed-cost barriers of entry into particular types of transmission (represented by the y-intercepts of each curve) and the marginal cost associated with using that transmission technology (the slopes of the curves). I have plotted theoretical cost curves for several important communications technologies. Television, requiring the enormous fixed cost of transmitters, has the highest fixed cost and its curve increases monotonically but relatively slowly. Radio, a very similar technology to television, has a similar curve, but shifted to a lower cost.

Telephone and print are instructive cases, in that the former involves extremely low fixed costs for the end user, but scales very badly due to its essentially one-to-one design. Print media begins with a higher fixed cost than telephony, and scales better, but does not scale as well as television and radio for very large numbers of recipients. The novelty of the Internet is that while traditional computers entail a moderate fixed cost (moderate in comparison to building a television transmitter, that is), they scale almost infinitely well with a marginal cost of nearly zero to reach more and more individuals with each transmission. As the cost of computers decreases, that horizontal curve shifts downward. The value of smartphones in this context is that they graft the minimal fixed costs of telephony to the minuscule marginal costs of transmission associated with computers.

This perfect scaling effect is what makes internet communications distinct from every previous communications technology, and is the theoretical basis for why social media emerged in the contemporary world. Social media represents infinitely scaling, low fixed-cost communication taken to its logical extreme. Every person on the planet able to communicate with no effort or expense with every other person on the planet.

1.2 Using Social Media Data

Early social media research focused on studying what exactly it was, and how it affected existing variables of interest. That is, this is literature in which social media is the dependent or independent variable. The greatest body of this early literature focused on the connection between social media and mass mobilization, in both authoritarian and democratic contexts (Farrell, 2012; Howard et al., 2011; Tufekci and Wilson, 2012; Wilson, 2017). In addition to the extensive article-length works, mobilization and social media have been explored in standout books from Clay Shirky (*Here Comes Everybody*) and

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

Manuel Castells (*Networks of Outrage and Hope*) (Castells, 2015; Shirky, 2009). The use of the Internet and social media in general in politics and society has been well covered by Cristian Vaccari's *Digital Politics in Western Democracies*, Christian Fuchs' *Social Media: A Critical Introduction*, Persily and Tucker's *Social Media and Democracy*, and Internet and Politics handbooks from both Routledge and Oxford (Chadwick and Howard, 2010; Dutton, 2013; Fuchs, 2017; Persily and Tucker, 2020; Vaccari, 2013). While constantly maturing, the study of social media as a phenomenon has evolved alongside the technology itself. For example, disinformation (and misinformation) campaigns via bots and trolls on social media have intensified, and have been made especially salient by their prominence in the last few American electoral cycles (Starbird, 2019).

The second wave of social media research emerged focusing on the use of it as data for the measurement of offline phenomena. It treats social media usage and content not as an independent or dependent variable itself, but as a new method for measurement alongside surveys, interviews, and other tried and true methods. This second genre of literature is still in its infancy relative to these established methods, with projects targeting a plethora of different research goals across the social sciences. These include applications as varied as Carlos Castillo's book *Big Crisis Data* using social media to detect when and where natural disasters occur (Castillo, 2016), comparison of traditional media content versus social media content during natural disasters (Vieweg et al., 2010), the study of information diffusion through a networked population (Jansen et al., 2009; Lerman and Ghosh, 2010; Romero et al., 2011), the prediction of shifts in the stock market based on tweet analysis (Bollen, Mao, and Zeng, 2011; Si et al., 2013), and how social media affects informal communication in the workplace (Zhao and Rosson, 2009).

In addition, a number of long-term and collaborative research projects are working on using social media as general measurement tools. In this arena, one standout is the Center for Social Media and Politics (CSMaP) at New York University led by Bonneau, Nagler, and Tucker, which has produced dozens of publications, datasets, policy white papers, and research tools. This scholarship runs the gamut from misinformation on social media (Guess, Nagler, and Tucker, 2019; Munger et al., 2022; Sanderson et al., 2021; Tucker et al., 2018), to political mobilization and protest on social media (Jost et al., 2018; Metzger and Tucker, 2017; Munger et al., 2019), to extensive software tools in both Python and R made available for free to the scientific community (CSMaP, 2022; Padmakumar and Terechshenko, 2020).

Additional significant institutional collaborations include the Social Media Lab (SoMe Lab) at the University of Washington. Led by Mason, Spiro, and Starbird, the lab has published exceptional scholarship, especially in the usage

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

of network analysis techniques across social media datasets (Reed, Spiro, and Butts, 2016; Wilson and Starbird, 2020). And at Northeastern, David Lazer's lab has published widely as well with dozens of academic articles and the seminal big-picture piece of scholarship *Meaningful Measures of Human Society in the Twenty-First Century* (Lazer et al., 2021).

Finally, the cross-university effort of the Digital Society Project has focused on large-n cross-country measurements of a variety of aspects of social media, including the ability of governments to censor and monitor social media, the prevalence of misinformation on the social media of each country, and the usage of social media to mobilize offline action of various sorts (Mechkova et al., 2019; Wilson, Lindberg, and Tronvoll, 2021; Wilson and Wiysonge, 2020).

Despite the exciting opportunities afforded by this new form of data, its usage involves a number of methodological challenges, both in terms of collection and measurement. Ruths and Pfeffer's article "Social Media for Large Studies of Behavior" in *Science* outlined the major obstacles to using social media as a direct source of data (Ruths and Pfeffer, 2014). Zachary Steinert-Threlkeld's *Twitter as Data* provides an excellent primer for the methods needed to get started collecting and analyzing tweets (Steinert-Threlkeld, 2018).

While scholars have made individual advances in resolving methodological challenges for their own particular projects, there has yet to be a general text that pulls together the methodologies of social media research into a cohesive framework. And problematically, many of the techniques for collecting social media data have remained locked behind programming technical expertise not particularly common in the social sciences. As such, this book provides an "on-ramp" for social scientists who want to use social media data in their research, regardless of what level of computer science expertise they currently possess. This book functions as a guide for scholars to thinking about the methodological problems inherent in social media data, and how to rigorously compensate for those problems such that the data can be held to the same scientific standards as surveys and other similar data.

This book is intended as a text for advanced undergraduates, graduate students, and researchers in the social sciences. Four broad uses of it as a text are envisioned. First, it serves as the primary textbook for courses explicitly focused on the usage of social media in the social sciences (such as Internet and Politics courses). Second, it provides a methodological supplement for substantive courses in Comparative and American Politics, allowing students to personally collect and utilize real data for their research topics using this new source of data. Throughout the book, example applications are pulled from a variety of global contexts, from the populations of the developing world to the social media activity of elites in the American government. Third, it functions as an applied toolkit to be plugged into research methods courses. Fourth, it

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

offers a reference for researchers who want to utilize social media data in their research, but currently lack the technical ability to get started on their own.

Chapter 2 introduces the building blocks of an infrastructure for collecting social media data. It includes a summary of what data is available via Twitter, and how we can best structure a collection system in general for any number of social science applications. In addition, it walks through how to collect a worldwide sample of all posted tweets in real time, along with database and compression tools for ensuring that the infrastructure can be used for long-term data collection projects encompassing millions of data points.

The next three chapters provide detailed discussion of three of the major methods used for analyzing social media data. Chapter 3 focuses on content analysis and introduces the collection of data from Twitter by either select keywords or languages. It then develops computerized content analysis techniques for use on tweets, covering the particular challenges of adapting these techniques for usage on the text from social media (for instance, dealing with the often very short passages of text, the especially dense usage of colloquialisms, and the frequent mixing of different languages within a particular source of social media text data). It also covers the download of other forms of content (such as video and images) and the handling of meta-objects such as mentions and hashtags.

One of the most exciting types of social media data are geolocated data, which include the source location of the post, based on the GPS capabilities of the posting device. Chapter 4 discusses the particular advantages offered by this data, including the capacity to perform extremely fine-grained subnational studies impossible with traditional sources of data. In addition, the chapter provides software for processing geocoded social media data in order to efficiently identify the country and subnational unit of every tweet in a collection, including an example application collecting *all* geocoded tweets in the world.

Chapter 5 introduces network analysis. Social media data frequently has elements that are amenable to network analysis, including friend/follower networks and retweet networks. This chapter addresses how to collect and operationalize this data into measures appropriate for network analysis. It shows how to collect en masse the timelines of a given set of users, in addition to traversing their friend and follower networks. In addition, it demonstrates how to do so by collecting all tweets of all members of Congress in real time. Finally, it demonstrates in applied form how to identify automated accounts (bots) among the data being collected.

Finally, the book concludes with Chapter 6's discussion of the particular ethical concerns raised by using social media data. This includes data privacy concerns, for instance the need in some contexts to anonymize unique user identifiers in all stored tweets so that not even the researchers have access to user

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press

names and such. In addition, the chapter reviews concerns frequently raised by institutional review boards in terms of human subjects research, and some of the thorny issues raised by the terms of use of social media sites with regard to data sharing and replicability. In particular, it will walk through what scholars need to know about the limitations imposed by Twitter's terms of use, what use cases are considered acceptable use (sharing data among researchers on the same project), and strategies for common scholarly needs that fall within gray areas (e.g. providing word frequency matrices so that content analysis can be fully replicated, but the terms of use conditions regarding republication of tweets are not violated).

Throughout, the book includes sample data and fully functional Python code in order to allow readers to proceed with their own collection and analysis of social media data. In particular, all source code is explained in full, and made fully available via the book's GitHub repository. Most importantly, all of the code in this book is built for use on real data, often in large quantities. Applications that work on small artificial datasets tend to break down and not scale when confronted with real data projects involving big data. As such, this book avoids using "toy" data and applications as examples throughout in order to ensure that the techniques are always explained in the context of real applications for real research.

By necessity of brevity, it focuses on using Twitter as the social media platform of choice simply because the public application programming interface and widespread support of the platform makes it the easiest to connect to and use for the broadest range of social science projects. However, the infrastructure constructed throughout the book, and the concepts and methods developed, are universal in the sense that they can be adapted and repurposed for any particular social media site.

Social media has put mass communication in the hands of normal people on an unprecedented scale, and has also given social scientists the tools necessary to *listen* to the everyday speech of everyday people around the world. This books aims to give social scientists the skills necessary to leverage that opportunity, and transform social media's infinite stream of information into social science data.

https://doi.org/10.1017/9781108677561.002 Published online by Cambridge University Press