Big Data is Too Small: Research Implications of Class Inequality for Online Data Collectionⁱ

Jen Schradie

L'Observatoire sociologique du changement
Sciences Po Paris
www.schradie.com

Citation

Schradie, Jen. 2017. "Big Data is Too Small: Research Implications of Class Inequality for Online Data Collection," Chapter 14, *Media and Class: TV, Film and Digital Culture*, Edited by June Deery and Andrea Press. Abingdon, UK: Taylor & Francis.

Abstract

With a growing interest in data science and online analytics, researchers are increasingly using data derived from the Internet. Whether for qualitative or quantitative analysis, online data, including "Big Data," can often exclude marginalized populations, especially those from the poor and working class, as the digital divide remains a persistent problem. This methodological commentary on the current state of digital data and methods disentangles the hype from the reality of digitally produced data for sociological research. In the process, it offers strategies to address the weaknesses of data that is derived from the Internet in order to represent marginalized populations.

Introduction

Digital technology has enabled researchers to access massive amounts of online data and analyze it almost instantaneously. Often this data comes from public and private Internet repositories that hold so many information bytes that a colloquial term is not just data, but “Big Data.”ⁱⁱ This data may be mined, scraped or downloaded from online sources for social research. Never before has so much data been available to so many people. Many types of researchers are enthralled with this gold mine. Corporations track, amass and sell information on demographics and habits of consumers. Media organizations track eyeballs and clicks. And social scientists use online data with unprecedented high speed and low cost, whether it’s using just a sliver of computer code to access millions of Tweets for a social network diagram, online news sites for automated text analysis, digital labor platforms for online surveys or online communities for ethnography 24/7.

There are many critiques of the research use of online data, such as not focusing enough on the offline world, not sharing it publicly with academic researchers, or not using the scientific method. The more mass utopian hype over Big Data also faces criticism. After all, it isn’t *that* new. The Romans and the Nazis amassed huge amounts of data on their populations. And then, of course, there is the creepy, Big Brother aspect. Another gaping hole in this digital cloud, I argue, though, is that online data analytics can leave out those excluded from the digital realm. This persistent sampling problem of online data threatens generalizability in sociological research.

Despite assertions that the “digital divide” is passé, inequality persists between those who can afford consistent and high quality Internet connectivity and those who cannot. But Big Data isn’t just gleaned from those who have *consumed* online content but also from those who have *produced* it. When it comes to creating publicly available online content, such as blogs, videos and social media posts, digital creators are still dominated by those with more income, education, and class power. While age is still a factor, class persists as one of the most significant determinants of online data production, especially as to how it intersects with gender, race, or ethnicity (Hargittai 2008; Hassani 2006; Schradie 2012; Segev 2010; Selwyn 2006).

In other words, Big Data is too small. Those who cannot afford consistent connectivity, let alone have the resources for content creation, are often absent from online analytics. Our samples, therefore are too small, not in sample “n” size but in generalizability. What mistakes, therefore, do Big Data studies make by ignoring data not visible online?

In this chapter, I introduce this question and what’s at stake in the research context, explain the state of digital inequality, offer a social class critique of using online data for research, as well as make suggestions for addressing these weaknesses. I then propose that social class be more thoroughly studied as a concept in these and other digital studies, as well as included as part of any study using Big Data analytics. Class is a contested term in the study of stratification. Many studies of (or incorporating) digital inequality examine socioeconomic divisions like income or education, rather than class power differences (Vanneman and Cannon 1987; Wright 1997; Zweig 2001). Nonetheless, not being able to control the digital means of online

content production or experiencing disempowerment in using digital technology – both can reduce how and how much poor and working-class people contribute to online data metrics (Robinson 2009; Schradie 2011; Sims 2014).

The implications for this hole in the Cloud are as vast as the deepening digital divide. Big Money is going into Big Data. Policymakers increasingly rely on online data for political decisions, so the concerns of the poor and working class may be left out of the equation (Hargittai and Jennrich 2016). Relying on Internet-based survey panel data for election predictions is less accurate without the incorporation of marginalized communities. For social scientists, our theories may only hold true for the digitally literate and in some cases the digital elite.

While every study has limitations in terms of its generalizability to the general population, if researchers increasingly rely on data culled from the Internet, empirical findings and subsequent theories will increasingly be skewed away from marginalized communities and toward those who can afford, know how, and are empowered to post online.

Big Data and the Scientific Method

In 2008, Chris Anderson, editor-in-chief of *Wired Magazine*, wrote a provocative article entitled, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (2008). He argued that hypothesis testing is no longer necessary with Google’s petabytes of data. Rather than needing first to come up with research questions, hypotheses, and *then* collect data, patterns that would organically emerge from massive data sets that would guide social scientific discoveries. The data would

provide all of the answers to how society works. Correlation would “supersede” causation:

“This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”

An easy strawman, Anderson’s piece generated a host of articles in academic journals decrying his claim. The overall consensus, to no one’s surprise, was that the scientific method, such as hypothesis testing, is far from over (e.g. Pigliucci, 2009). Some focused on the debate around “correlation is not causation” and that correlation can lead you in the wrong direction with spurious noise. Still, many implicitly pointed to what Box (1976) articulated so well pre-Big Data – that science is an iterative process in which correlation is useful in that it can trigger research which uses hypothesis testing. Most argued as Pigliucci (2009:534) articulated, “But, if we stop looking for models and hypotheses, are we still really doing science? Science, unlike advertising, is not about finding patterns—although that is certainly part of the process—it is about finding explanations for those patterns.”

For many social scientists, this issue may simply seem like fodder for an undergraduate lesson in the scientific method. Yet what may unite all scholars is that throwing online data up in the air to see what lands fails to carry out the mission of academic work: to generate research questions situated in existing scholarly work –

and in the process, challenge and build on each other's research to generate knowledge.

The problem is not simply the lack of hypothesis testing but the absence of social theory and normative value in research questions. In fact, the first part of the title of Anderson's *Wired* article was, "The End of Theory." Is research driven by available Big Data from corporations or is it driven by theoretical questions of how society operates? The dearth of social theory in data science is a wider problem than this chapter can tackle, but inherent in evaluating how society operates theoretically and in practice is how inequality functions.

Some may simply fail to acknowledge its generalizability problems. The risk of taking online data at face value, without a broader societal and historical context, has only accelerated since Anderson's controversial article. First, online data permeate many academic disciplines, from new data science majors to digital humanities programs. In 2016, 95% of nationally ranked universities in the U.S. had data analytics program. (Tableau 2016). Next, journals have even emerged on this specific topic, such as *Big Data and Society*. At the same time, access to online data, as well as research techniques using it, have also become much easier and faster to download, analyze and visualize. But the digital divide has largely been overlooked with an increasing rate of scholarship that often analyzes publicly produced online data.

Since the initial firestorm over the scientific method, researchers have decried pitfalls that researchers face with Big Data. Some scholars have tackled the question of representation with online data, but the debate often focuses on whether or not it tells a better story than smaller, micro-level analog data. For instance, Karpf (2012)

contended that what is happening offline and the context of Big Data are essential to understanding it while Snijders et al. claimed that the mathematical models used to mine data do not reflect actual online behavior (2012). Nonetheless, Hargittai and Jennrich did argue that policy needs to reflect inequalities derived from online data (2016). As a result, the focus on our “digital trace” does not tell a complete story of society. Despite these limitations of Big Data, scant attention has been paid to how social stratification may limit online data’s representativeness in research.

Digital Production Inequality – What’s Missing?

While the debate over scholarly Big Data is abundant, another critical area, I argue, is *who* specifically may be more likely to be left out of research data: the poor and working class. They are less apt to be online content users and creators, which can bias the digital data that researchers collect. Therefore, scholarship using these data may result in theories skewed toward metrics from middle to upper class digital participants. If online data often captures both the public’s consumption and production of content,ⁱⁱⁱ it is helpful to analyze these two critical areas of digital inequality to understand how, exactly, Big Data is too small.

First, consider consumption, or access, the Internet. Data metrics may capture consumers use of various platforms. One might expect that the digital connectivity divide will eventually disappear because of what classic theories of diffusion (Coleman, Katz, and Menzel 1957) have to say about adoption rates of successful innovations, such that most everyone should be online soon-after the early adopters from the 1990s. Despite increasing adoption rates in the United States and in many countries, research shows a persistent class-based divide (Anderson 2017; Zhang

2014) of consumers. African-Americans, Latinx and people with lower educational and income levels are less likely than non-Latinx, white and college educated Americans to go online and have connectivity, although the gap is closing slightly for Latinx (Hargittai, E. & Walejko 2008; Jones et al. 2009; Schradie 2015). As of 2016, 13% of Americans did not use the Internet, with income and education as major contributors to that gap (Anderson and Perrin 2016). But the likelihood that people with a high school education are online is half as much as those with more education, and scholars have found that stating in a survey that one has been online may only indicate occasional access (Robinson 2009), rather than consistent high quality connectivity.

But who is creating and posting online content into the digital cloud? Specifically, are the poor and working class posting to Twitter, blogging or uploading videos at the same rate as people from higher socioeconomic classes? In a word, no. Inequality in connectivity also affects what people are able to *do* online, specifically production of digital content. Certainly, not all wealthy, privileged people compose social media posts. However, working class and poor people are doubly disadvantaged. They are less likely to be online in the first place – and then among those who make it into the online pool, they are less likely to be content creators online. For instance, people with a high school education are half as likely to create online content than those with a college degree, *even among people who are online* (Schradie 2011). Simply, consumption of online content does not automatically lead to production of it. Content may not always differ based on the social class

background of the creator, yet a body of scholarship has demonstrated that class does shape media content (e.g. Kendall 2005).

And this class-driven gap in online content creation shows persistence, rather than closing over time. The digital divide is not over nor are there signs that it will ever completely disappear, given that new hardware and software applications are continuously introduced, so playing catch-up to the early adopters is a treadmill, rather than a finish line. Being able to afford and control digital gadgets and Internet connectivity are key, but it is also a question of having the skills, confidence, and empowerment to post online (Van Deursen and Van Dijk 2011; van Dijk 2005; Haight, Quan-Haase, and Corbett 2014; Schradie 2017). Certainly, the “leapfrog” idea of people using cheaper mobile devices, such as smart phones, and bypassing more expensive laptops or desktops, does exist, as studies do point to people with lower incomes relying on these types of gadgets (Pew Research Center 2017). However, the consistent use of these gadgets still requires that someone can purchase a phone, pay for a data plan, is the only user of it, knows how and is confident in using it, and has time to do so. Fully controlling the means of digital production is key. Inexpensive mobile phones may be useful for consumption or making short online posts but not practical for longer or more complicated digital content creation.

But what of the intersection of class with race? On the one hand, in the United States, Blacks are up to twice as likely to post to social media than whites (Hargittai and Litt 2011; Schradie 2012). Like other media formats, scholars have found a racial divide in different forms of media (Murthy, Gross, and Pensavalle 2016; Watkins 2011). Given the prevalence of #BlackTwitter in which African-Americans use the

platform for support, community and other solidarity purposes (Clark 2015), one might expect that poor African-Americans may be an exception to the class gap. On the other hand, while Blacks may *produce* some forms of social media content more than whites, this is only true among those who are online, as they are less likely to *consume* online content in the first place. Simply, the intersection of race and class is critical – poor African-Americans are dramatically less likely to post online content than wealthy whites. African-Americans are less likely than whites to be online in the first place, and social class is a critical part of that gap.

How is Big Data Too Small?

If, as sociologists, we and other social scientists consistently use Big Data samples with a blind eye toward generalizability, then we are heading down a non-robust path that favors simplicity over the messiness of understanding the weaknesses of online data that can leave out the marginalized. In the process, it weakens rigorous methodological and theoretical analysis. Social stratification, particularly class divisions, matter with Big Data, but how? Substantive methodological methods illustrate how class makes a difference in Big Data analytics.

1. Network Analysis

When Anderson wrote his article in 2008, conducting social network analysis required the use of clunky software and cleaning up databases. Twitter, for instance, was just emerging on the social media landscape. Now, with a touch of a button, a day's Twitter conversation is instantaneously displayed creating a sociogram of people's network ties. But that Twitter data is only robust as the people who are tweeting, favorite-ing, or replying. In 2016, 20% of Americans with a high school

education (or less) *and* who are already online used Twitter while 29% of college educated Americans did, but this does not account for the people who are not online at all, who are more likely to be more working class.¹ So the gap is even wider in the general population. And just like the class differences difference between consumption and production of content more generally, everyone who has a Twitter account (consumption) is not necessarily Tweeting (production) to create nodes of Tweets and re-Tweets, for instance.

It is possible that digital inequality may not impact the robustness of any analysis of social structural networks with online data because social classes are already divided and siloed into their own stratum, but one cannot make this assumption without examining class differences. Furthermore, we are not seeing how particularly vulnerable and marginalized populations fit into our so-called network society. Perhaps the lower levels of working class people who are online are, therefore, part of “weak” versus “strong” ties if an analysis is culled from online data of the general population. Central nodes may be missing, as well. If that is the case, then the results lose more than generalizability. They are simply not valid. To overcome these weaknesses in social network analyses with digitally derived data, one could do diligence in determining the social class background of the nodes in the network, or at least an approximation. This may require two levels of assessment. One is by general Internet use and the other is online participation on specific platforms under study. Obtaining social class information may require offline data collection or,

¹ <http://www.pewinternet.org/2016/11/11/social-media-update-2016/>

perhaps, partnering with platforms, if one is able to access this data without compromising privacy concerns.

A generalizability limitation example is with digital activism networks. This type of analysis is common with studies of digital activism, analyzing how people are spreading information or communicating protest events. Research has shown class inequalities with social media usage for activism, in general (Pickerill 2003; Schradie 2017), but less is known about digital activist network inequality. Therefore, if one uses Twitter data to analyze the network ties of a protest and working class participants are present but not tweeting as much, or at all, then the network ties resulting from this Twitter analysis would skew toward more privileged users and fail to incorporate working class network ties from that event. To continue with the Twitter example, social class is one of the least understood demographic variables of participants on the platform because it is difficult to identify based only on the platform's data. If social class does not impact the research question at all, then one could proceed without knowing this information but only after eliminating this factor. Other platforms, such as Facebook, have more demographic data, such as educational background, so distinguishing between college and high school educated people is one option to approximate class, but this may be difficult with the Silicon Valley company's increasing restrictions on researcher access to data through their Application Programming Interface (API). Analyses of social networks would benefit from a full examination of how the results might vary with different social classes, rather than in a footnote. If it is not possible to determine people's class identity, at the very minimum, scholars should acknowledge this limitation.

2. Automated Text Analysis

Automated text analysis evaluates content. With digitally produced content accelerating, the options for examining text appear to be boundless. Some approaches are basic sentiment analyses using predetermined ways to examine text. More sophisticated techniques employ topic modeling with a way to “teach” the algorithm with human intervention, often called machine learning. The general argument, like much of the digitally enabled research using online data, is that text mining for content analysis is now not only more efficient but more accurate than hand-coding (Mohr and Bogdanov 2013). Just as with network analysis, a variety of disciplines use automated text analysis, from history and film studies to computer science and sociology. But whose voice (and whose text) is being analyzed? While some scholars, especially historians have gone out of their way to digitize handwritten or non-computer generated text (Nelson 2013), the tendency is increasingly toward analyzing what is already online, as that is also more efficient. But at what cost when working class people are less likely to post online content in the first place?

One example is the study of so-called citizen journalism. Scholars have extensively studied the digital news landscape as an example of what citizen journalists are now able to do outside the confines of traditional news outlets (Gilmor 2004). Prior to social media adoption and diffusion, traditional news media did not accurately represent the poor and working class, and when they did it was often mediated by voices of so-called experts, rather than by the population itself (Kendall 2005). In the early social media era, a common view was that professional journalists

from traditional media outlets got in the way of average citizens reporting directly through the likes of social media feeds, blogs and other online posts (Wall 2015). In this way, a more pluralist journalism would emerge. Yet, it was not simply the onslaught of so-called “fake news” that got in the way of more working class people having representation in these formats. Bloggers, for instance, tend to be of the same privileged social class as those of traditional reporters – there are just more of them (Hindman 2009). And if scholars rely on citizen journalist data, they may find that the citizen skews toward the elite.

Before the dawn of digital technology, scholars studied the artefacts of the elite – from 16th century books to the *New York Times*. However, there was little pretense that this was representative of all of society. A common view of Big Data and online data, however, is that this digital information is more egalitarian, yet without metrics and offline comparisons, it is difficult to make that claim.

3. Survey data

This trend of using online data in research is not just limited to digital content creation. Scholars harness the internet to conduct surveys and polls. For years, the robustness of traditional telephone polling data has been on the decline with the increase in unlisted numbers and mobile phone lines, as well as fewer landline numbers and less public trust in the participation of polling. Two trends have emerged in response. One is the use of self-selecting users of Mechanical Turk (MTurk), an online labor system. MTurk workers are paid a nominal fee for doing low wage tasks, including taking polls. Studies have shown that they tend to skew towards digitally savvy and slightly higher socioeconomic levels (Berinsky, Huber, and Lenz

2012) although other studies of MTurk workers' demographics have not reported socioeconomic details well enough to evaluate, often focusing on other characteristics (Huff and Tingley 2015). The other survey approach has been with panel data. This is also a paid survey system in which people agree to participate in polls over a lengthy time period, and researchers pay companies to manage the polls. Since these surveys are conducted via the Internet, the participants are also less likely to be from lower income or educational levels. Digital inequality has been given scant attention with online panel data. One analysis comparing MTurk and panel data in terms of robustness said that these can be used by "anyone with Internet access" (Weinberg, Freese, and McElhattan 2014), and it found little difference in outcome with the two approaches. The problem, though, is that not everyone has sufficiently regular and consistent access in order to participate. Moreover, research has shown that many non-responders to survey questions are from lower classes (Laurison 2015). The skill level required may also be a challenge for people without substantial new media literacy.

Social scientists have a few ways to work around this problem, in general, with representation problems in survey data. First, they can use a stratified sample in that they can ensure that those less represented are sought-after to be included in the sample. Related is an approach to "weigh" people who are less represented, to give their findings more emphasis so that the sample ends up being representative of the general population under study. But if one only gives more statistical weight to poor people who are online this technique does not address any differences between people who are and are not online to participate in the panel. For instance, it is

possible that some “poor” participants are college-aged youth, more likely to be online, and therefore have a temporary low income, rather than those who come from a working-class background. All of these companies who employ digital workers, whether with M-Turk-like platforms or with panel data, need more rigorous evaluation and transparency to ensure their representativeness of class divisions.

4. Online ethnography

The use of online data is not just limited to quantitative analysis. Researchers use qualitative techniques such as digital ethnography as a way to understand thick descriptions of the online world. This can include participant observation in an online community, such as listservs or games. Qualitative studies have a different approach from quantitative analyses in that they offer in-depth mechanisms and explanations of how society operates. Qualitative researchers who zoom in for precision (Becker 1998) and thick description (Geertz 1973) on middle to upper class online users may inadvertently create findings only relevant to that class. Just as scholars have found that class differences shape how people interpret older forms of media (Press 1991), social class differences are reproduced with online engagement. For instance, having an information habitus or social milieu that is more middle to upper class shaped how some youth sought information online, with higher income users being more playful online, what Bourdieu called “skhole;” lower income students, though, lacked the time and autonomy to do more than tasks set before them (Robinson 2009).

Online ethnographies could benefit from digging in the direction of how social class differences may shape online dialogue, postings, or other content – if it is there at all. And with online ethnography, like with any method, it is particularly important

to acknowledge how reflexivity, or the class perspective and position of the ethnographer, may shape observation. Qualitative researchers who have overcome the limitations of social class differences in how much people are online have conducted both online and offline ethnography to go deeper into the everyday lives of the poor and working class in online engagement. For instance, Sims delved into the contextualized lives of students at a school designed to overcome technological inequalities in the first place, finding that even skill and access remedies resulted in less beneficial online use (2014). The result is that understanding class differences in an ethnographic online context is similar to the other methods – that inequalities can be incorporated into the research design stage.

Implications of Small Data

What is the upshot methodologically? Certainly, any data set is imperfect and has limitations as to its generalizability. The efficiency and other advantages to using online data are clear. However, digitally created online data almost *always* have the same limitation: it does not include people not online or do not post as frequently as others. Therefore, it is rarely generalizable to the population as a whole.

Of course, if scholars are just interested in what is happening online, then the Big-Data-is-too small concept is not as relevant, but if we want to understand what is happening in all of society – both online and offline--we cannot just examine those who are online 24-7.

Ultimately, then, theory suffers from Big Data being too small.

A well-established debate within Science and Technology Studies is about ways to move beyond technological determinism (Fischer 1992), which is an

outdated theory that technology unilaterally shapes society and culture. Nonetheless, there is still an enthusiasm for using Big Data without incorporating class-based and other structural limitations of this data, such that we know little about how the reverse of technological determinism -- how society determines technology. More precisely, how societal differences in the form of class divisions shapes how and how much this shapes the voices, nodes or text within Big Data derived from online sources. While these broad theoretical questions are not new, it would behoove the academy to push for class-based contextual understandings of online data collection. Castells said, "...Of course, technology does not determine society....nor does society script the course of technical change ...technology *is* society, and society cannot be understood without its technological tools (2010: 10)." Still, what is missing is a definition of society that recognizes and includes all social classes. Robust theories generated from online data need to incorporate more than digital influencers.

Castells (2010) also argued that new technologies and a ripe societal context (i.e. the right historical time) converged to create a new global network society, although he fluctuated between arguing that this is not a revolutionary shift to broader claims that this new era is more profound than the industrial revolution in terms of its effect on the economy, culture and society. He and others (i.e. Raine and Wellman 2012) have suggested that we now have a network society because of the digital age, yet the question remains, a network society for whom? Is it changing for everyone equally? Broader theories of digital society would benefit from including and analyzing all social classes, not just those who are digitally networked. Most of the theory building and knowledge production from scholars of digital media will sit

in more meso-level (or smaller) theoretical contributions, but even then, it is critical to have the most accurate understanding of one's data – and what is lacking – to build theory in a contextual, rather than an a-contextual, direction. For instance, with the citizen journalism example, a half-century earlier, Mills described the hierarchy of the power of the media in which the working class is largely left out of the equation (1958). Presented with the digital era, especially in light of heavily funded “fake” news, as well as the digital divide, Mills might be dubious about the power of working class citizens in the digital era.

Simply, the relationship between inequality and Big Data needs further analysis and attention on its own, particularly by analyzing class relations, rather than only socioeconomic variables. Nonetheless, studies which do make broad theoretical claims about online data need to include questions of inequality and sampling limitations in more than a footnote. I am not arguing that Big Data, therefore, should be discarded. Quite the opposite, it is not only a literal goldmine of information for academia but it is also fascinating and fun to analyze. Instead, we need to figure out ways to use information from the digital cloud in a methodological and theoretical way that represents all of society, not just those online, or at a minimum, recognize that it is smaller than we might have thought.

References

- Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *Wired Magazine*. Retrieved June 7, 2012 (<https://www.wired.com/2008/06/pb-theory/>).
- Anderson, Monica. 2017. "Digital Divide Persists Even as Lower-Income Americans Make Gains in Tech Adoption." Washington D.C.: Pew Research Center. Retrieved April 1, 2017. (<http://www.pewresearch.org/fact-tank/2017/03/22/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption/>).
- Anderson, Monica and Andrew Perrin. 2016. "13% of Americans Don't Use the Internet. Who Are They?" Washington D.C.: Pew Research Center. Retrieved December 1, 2016. (<http://www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/>).
- Becker, Howard Saul. 1998. *Tricks of the Trade: How to Think about Your Research While You're Doing It*. Chicago, IL: University of Chicago Press.
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20:351–68.
- Box, G. E. P. 1976. "Science and Statistics." *Journal of the American Statistical Association* (71):791–99.
- Castells, Manuel. 2010. *The Rise of the Network Society*. West Sussex, UK: Blackwell.
- Clark, Meredith. 2015. "Black Twitter: Building Connection through Cultural Conversation." Pp. 205–18 in *Hashtag Publics: The Power and Politics of Discursive Networks*, edited by Nathan Rambukkana. New York: Peter Lang Publishing Inc.
- Coleman, James, Elihu Katz, and Herbert Menzel. 1957. "The Diffusion of an Innovation Among Physicians." *Sociometry* 20(4):253–70.
- Fischer, Claude S. 1992. *America Calling: A Social History of the Telephone to 1940*. Berkeley: University of California Press.
- Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York, NY: Basic Books.
- Gilmor, Dan. 2004. *We the Media*. Santa Rosa, CA: O'Reilly Media.
- Haight, Michael, Anabel Quan-Haase, and Bradley Corbett. 2014. "Revisiting the Digital Divide in Canada: The Impact of Demographic Factors on Access to the Internet, Level of Online Activity, and Social Networking Site Usage." *Information, Communication & Society* 17(4):503–19.
- Hargittai, Eszter. 2008. "Whose Space? Differences Among Users and Non-Users of Social Network Sites." *Journal of Computer-Mediated Communication* 13(1):276–97.
- Hargittai, Eszter and Kaitlin Jennrich. 2016. "The Online Participation Divide." Pp. 199–213 in *The Communication Crisis in America, And How to Fix It*, edited by Mark Lloyd and Lewis A. Friedland. New York: Palgrave Macmillan.
- Hargittai, Eszter and Eden Litt. 2011. "The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption among a Diverse Group of Young Adults." *New Media and Society* 13(5):824–42.
- Hargittai, E. & Walejko, G. 2008. "The Participation Divide: Content Creation and

- Sharing in the Digital Age." *Information, Communication & Society* 11(2):239–56.
- Hassani, Sara Nephew. 2006. "Locating Digital Divides at Home, Work, and Everywhere Else." *Poetics* 34(4-5):250–72.
- Hindman, Matthew. 2009. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.
- Huff, Connor and Dustin Tingley. 2015. "Who Are These People? Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research and Politics* July-Sept:1–12.
- Jones, Steve, Camille Johnson-Yale, Sarah Millermaier, and Francisco Seoane Pérez. 2009. "U.S. College Students' Internet Use: Race, Gender and Digital Divides." *Journal of Computer-Mediated Communication* 14(2):244–64.
- Karpf, David. 2012. "Social Science Research Methods in Internet Time." *Information, Communication & Society* 15(5):639–61.
- Kendall, Diana. 2005. *Framing Class: Media Representations of Wealth and Poverty in America*. New York: Rowman & Littlefield.
- Laurison, Daniel. 2015. "The Willingness to State an Opinion: Inequality, Don't Know Responses, and Political Participation." *Sociological Forum* 30(4):925–48.
- Mills, C. Wright. 1958. "The Structure of Power in American Society." *The British Journal of Sociology* 9(1):29–41.
- Mohr, John and Petko Bogdanov. 2013. "Topic Models: What They Are and Why They Matter." *Poetics* 41(6):545–770.
- Murthy, Dhiraj, Alexander Gross, and Alexander Pensavalle. 2016. "Urban Social Media Demographics: An Exploration of Twitter Use in Major." *Journal of Computer-Mediated Communication* 21:33–49.
- Nelson, Laura. 2014. "Enduring Political Logics as Cultural 'Memory': Comparing Women's Organizations in Chicago and New York City." Dissertation. UC Berkeley. Berkeley, CA,
- Pew Research Center. 2017. "Mobile Fact Sheet." Washington D.C.: Pew Research Center.
- Pickerill, Jenny. 2003. *Cyberprotest: Environmental Activism Online*. Manchester, UK: Manchester University Press.
- Pigliucci, Massimo. 2009. "The End of Theory in Science." *EMBO reports* (10):534.
- Press, Andrea. 1991. "Women Watching Television: Gender, Class And Generation In The American Television Experience." Philadelphia, PA: University of Pennsylvania Press.
- Raine, Lee and Barry Wellman. 2012. *Networked: The New Social Operating System*. Cambridge, MA: MIT.
- Robinson, Laura. 2009. "A Taste for the Necessary." *Information, Communication & Society* 12(4):488–507.
- Schradie, Jen. 2011. "The Digital Production Gap: The Digital Divide and Web 2.0 Collide." *Poetics* 39(2):145–68.
- Schradie, Jen. 2012. "The Trend of Class, Race, and Ethnicity in Social Media Inequality: Who Still Can't Afford to Blog?" *Information, Communication & Society* 15(4):1–17.
- Schradie, Jen. 2015. "Silicon Valley Ideology and Class Inequality: A Virtual Poll Tax on Digital Politics." Pp. 67–84 in *Handbook of Digital Politics*, edited by Dean

- Freelon and Stephen Coleman. Cheltenham, UK.: Edward Elgar.
- Schradie, Jen. 2017. "The Digital Activism Gap: How Class and Costs Shape Online Collective Action." *Social Problems* Forthcoming.
- Segev, Elad. 2010. *Google and the Digital Divide: The Biases of Online Knowledge (Chandos Internet Series)*. Oxford: Chandos Publishing.
- Selwyn, Neil. 2006. "Digital Division or Digital Decision? A Study of Non-Users and Low-Users of Computers." *Poetics* 34(4-5):273–92.
- Sims, Christo. 2014. "From Differentiated Use to Differentiating Practices: Negotiating Legitimate Participation and the Production of Privileged Identities." *Information, Communication & Society* 17(6):670–82.
- Snijders, Chris, Uwe Matzat, & Ulf-Dietrich Reips. 2012. "'Big Data': Big Knowledge Gaps in the Field of Internet Science." *International Journal of Internet Science* 7(1):1–5.
- Tableau. 2016. "The State of Data Education in 2016." Seattle, WA: Tableau Inc.
- van Deursen, A. J. A. M. and J. A. G. M. van Dijk. 2011. "Internet Skills and the Digital Divide." *New Media & Society* 13(6):893–911.
- van Dijk, Jan A. G. M. 2005. *The Deepening Divide: Inequality in the Information Society*. Thousand Oaks, CA: Sage Publications.
- Vanneman, Reeve and Lynn Weber Cannon. 1987. *The American Perception of Class*. Philadelphia, PA: Temple University Press.
- Wall, Melissa. 2015. "Citizen Journalism: A Retrospective on What We Know, an Agenda for What We Don't." *Digital Journalism* 3(6):797–813.
- Watkins, S. Craig. 2011. "Digital Divide: Navigating the Digital Edge." *International Journal of Learning and Media* 3(2):1–12.
- Weinberg, Jill, Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsourced-Recruited Sample." *Sociological Science* 1(August):292–310.
- Wright, Erik Olin. 1997. *Class Counts: Comparative Studies in Class Analysis*. New York: Cambridge University Press.
- Zhang, PH. 2014. "Digital Divides and Socio-Demographic Factors: A Longitudinal Quantitative Study of Internet Users in U.S. from 2000 to 2010." Dissertation. Capella University, Ann Arbor, MI.
- Zweig, Michael. 2001. *The Working Class Majority*. Ithaca, NY: Cornell University Press.

ⁱ This paper received support from the Jacob K. Javits Foundation and the Berkeley Center for New Media. Support through the ANR Labex IAST is gratefully acknowledged. Additional guidance provided by Abigail Andrews, Mike Hout, and Kim Voss, but the author is solely accountable for any omissions or errors. Please send any comments to Jen Schradie at jen.schradie@gmail.com.

ⁱⁱ The word, "data" is plural but I use the term "Big Data" in its singular form to represent the concept.

ⁱⁱⁱ Of course, Big Data is not limited to drawing just from online data from consumers' consumption or production of it.