HHS Public Access

Author manuscript

Demography. Author manuscript; available in PMC 2018 December 13.

Published in final edited form as:

Demography. 2018 October ; 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2.

Promises and Pitfalls of Using Digital Traces for Demographic Research

Nina Cesare[1], **Hedwig Lee**[2], **Tyler McCormick**[3,4], **Emma Spiro**[2,5], and **Emilio Zagheni**[6]

Nina Cesare: ncesare@bu.edu

[1]Department of Global Health, Boston University, Boston, MA, USA

[2]Department of Sociology, Washington University, St. Louis, MO, USA

[3]Department of Sociology, University of Washington, Seattle, WA, USA

[4]Department of Statistics, University of Washington, Seattle, WA, USA

[5]Information School, University of Washington, Seattle, WA, USA

[6]Max Planck Institute for Demographic Research, Rostock, Germany

Abstract

The digital traces that we leave online are increasingly fruitful sources of data for social scientists, including those interested in demographic research. The collection and use of digital data also presents numerous statistical, computational, and ethical challenges, motivating the development of new research approaches to address these burgeoning issues. In this article, we argue that researchers with formal training in demography—those who have a history of developing innovative approaches to using challenging data—are well positioned to contribute to this area of work. We discuss the benefits and challenges of using digital trace data for social and demographic research, and we review examples of current demographic literature that creatively use digital trace data to study processes related to fertility, mortality, and migration. Focusing on Facebook data for advertisers—a novel "digital census" that has largely been untapped by demographers—we provide illustrative and empirical examples of how demographic researchers can manage issues such as bias and representation when using digital trace data. We conclude by offering our perspective on the road ahead regarding demography and its role in the data revolution.

Keywords

Digital data; Social media; Big data; Demographic methods

Introduction

Researchers' interest in and excitement toward *Big Data*—roughly defined as data sets that are large and heterogeneous enough to make storing, managing, and analyzing data difficult

Correspondence to: Nina Cesare, ncesare@bu.edu.

[1]For a review of these data, see Ruggles (2014) in a past issue of Demography.

(Sagiroglu and Sinanc 2013)—has grown significantly in the past several years. Many forms of Big Data are social data and therefore valuable to those interested in examining behaviors, attitudes, and macro-level social processes. Our study focuses on one type of socially relevant data—namely, *digital traces*—the results of social interaction via digital tools and spaces as well as digital records of other culturally relevant materials, such as archived newspapers and Google searches (Manovich 2011), including data from popular social networking sites (such as Facebook or Twitter), personal blogs, collaborative online spaces (such as Wikipedia), and data derived from mobile phone or credit card usage. These digital traces—a term that can be attributed to Latour (2007) in an effort to spread awareness regarding the permanence and traceability of online interaction—provide valuable insight into human behavior. However, they come in a variety of structures, including text, images, videos, and networks (Lazer and Radford 2017) and were not "constructed and designed with research questions in mind" (Ang et al. 2013:39).

Although the scope of our discussion addresses digital traces—which could include many forms of digital documentation of human behavior, including e-mail, credit card transactions, cell phone records, and more—most of the examples that we provide focus on social media data. We do not address very large but systematically collected conventional data sets, although these data share some similarities with digital trace data.[1] The context for our discussion is a growing body of research that has considered the use of digital trace data to study population processes, including fertility (e.g., Billari et al. 2013), migration (e.g., Zagheni and Weber 2012), and mortality (e.g., Tamgno et al. 2013). Further, the presence of this literature reflects an intellectual environment wherein a substantial portion of demographic research using Web and social media has been published in outlets that are not traditionally accessed by demographic researchers (e.g., proceedings of computer science or social informatics conferences) and where many advances are driven by researchers not classically trained in formal demography.

We first provide background on features of digital traces that make them promising for population research, followed by a discussion of the technical, ethical, and institutional challenges for research with digital traces. We then review the emerging literature on digital demography and provide a snapshot of the state of the art. We proceed to present our perspective on some open research questions that the community of demographers is well positioned to tackle. More specifically, as an illustrative example, we discuss a new and promising data source that has been largely untapped by demographers—Facebook data for advertisers—and show how these data can be leveraged as a digital census. We conclude with a discussion of our article within the broader context of the discipline.

The Promises of Digital Traces

Not only are digital trace data geographically far-reaching and generated on a nearly continuous basis, but they also provide unique, unsolicited insight into patterns of interaction and self-expression. In considering the ease by which every move mediated by digital technology is stored, archived, and available for analysis, Latour (2007) stated, "The precise forces that mold our subjectivities and the precise characters that furnish our imaginations

are all open to inquiries by the social sciences. It is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable."

The ability to collect large quantities of data in very short periods of time from digital sources has prompted interest and enthusiasm over their use across a variety of scientific fields. A 2016 Sage Publishing survey of 9,412 social scientists, for instance, found that 33 % of them had engaged in Big Data research (which includes analysis of digital trace data) in the past year; and of those who did not, 49 % planned to do so in the near future (Metzler et al. 2016). At the root of this interest are the advantages that these data offer. Although some digital trace data (such as Facebook profile information) are difficult to access or inaccessible to academic researchers, some forms of digital trace data are publicly available for download or accessible through data-sharing agreements. Sites such as Twitter and Pinterest offer their application programming interfaces (APIs) to the public, making it possible for developers and researchers alike to stream past and/or current, up-to-the-minute (or even up-to-the-second) data. Data startups, such as Gnip (gnip.com; now owned by Twitter), help facilitate the storage and distribution of digital trace data and view social science researchers as an important part of their market base. Telecommunications providers are amenable to social research as well and often provide documented and anonymized digital trace data from their customers to researchers interested in analyzing these data (Blumenstock 2012; Blumenstock and Eagle 2010; Blumenstock et al. 2015).

Digital traces of interaction are created and can be collected in real time, allowing researchers to examine small fluctuations in attitudes or behaviors rather than observing the same group at discrete time points. The ability to represent time as "continuous, rather than bundled" (Ruppert et al. 2013:36) opens a wealth of new opportunities to researchers, such as examining real-time trends in daily activities (Golder and Macy 2014), mobility (Williams et al. 2015), attitudes (O'Connor et al. 2010), health behaviors (Heaivilin et al. 2011), and migration (Zagheni and Weber 2012; Zagheni et al. 2014). Researchers may also examine these behaviors before, during, and after crisis events, such as natural disasters (Reeder et al. 2014; Sutton et al. 2014) or terrorist attacks (Starbird et al. 2014).

Aside from being high volume, easy to collect, and generated in real time, digital traces also provide unsolicited documentation of individuals' opinions and interactions. A large body of literature has documented the difficulty of capturing attitudes and opinions related to controversial topics because of social desirability bias (Belli et al. 1999; Holbrook and Krosnick 2010; Tourangeau and Yan 2007) or selective recall (Fadnes et al. 2009). Digital traces can provide ready access to users' controversial opinions and/or disclosure of engagement in deviant behavior, which may be easy to conceal through other forms of data collection (Berinksy 1999; Marwick and Boyd 2010). Moreover, digital traces provide documentation of movement and activity (Palmer et al. 2013), which may help researchers circumvent other possible sources of data error, such as recall bias. One important drawback is that unsolicited data also contains content from bots (e.g., software that tweets automatically via the Twitter API), individuals misrepresenting themselves, and other violations of the ideal user assumption (Lazer and Radford 2017).

Finally, using digital trace data may allow researchers to access groups that are hard to reach and/or generally underrepresented by traditional survey techniques. A demographically diverse population of individuals uses social media sites to interact, track daily habits, and gather and share information on current events (Barberá 2016; Lewis et al. 2013). Recent data from the Pew Research Center (2018) (see Table 1) indicate that Internet usage among non-Hispanic blacks and Hispanics roughly parallels that of non-Hispanic whites overall, non-Hispanic blacks and Hispanics are actually more active on some social media sites than white users (Smith and Anderson 2018). These numbers show that racial/ethnic minorities are not only highly present on sites such as Instagram and Twitter but also proportionally overrepresented in some cases.

Challenges and Opportunities for Digital Research

Following the burst of initial research advocating the possibilities of using digital trace data to study social phenomena, a recent body of literature has emerged outlining ways in which these data have been misused, cautioning researchers about the unique challenges that they present (Adams and Brueckner 2015; boyd and Crawford 2012; Couldry and Powell 2014; Felt 2016; Golder and Macy 2014; González-Bailón 2013; Kitchin 2014; Lazer et al. 2014; Lewis 2015; Lohr 2012; Manovich 2011; Tufekci 2014; Zwitter 2014). We argue that these challenges are opportunities for researchers to advance the field of demography—and the social sciences in general—by finding ways to overcome them. We outline challenges and common areas of data misuse, and discuss how existing research seeks to improve applications of digital trace data and correct past mistakes.

First, given that traces are typically not originally collected for research purposes, the decision to use such data is usually somewhat opportunistic. We do not see this as inherently negative because it highlights the importance of theoretically grounding and motivating empirical findings. In a traditional survey-based framework, a researcher with a theoretically motivated question may first collect survey data by using a carefully crafted set of definitions for each item in the survey. Collecting data about a social network, for example, requires defining what it means for individuals to be “friends.” Collecting data about exercise requires bounding what types of activities constitute a workout. A study on unemployment would first define the amount of time between jobs required for a person to be considered persistently unemployed. With digital trace data, the process of operationalizing research concepts may occur in reverse: researchers observe all activity but then must map the observed data back to covariates. Both settings require the researchers to make decisions about definitions that do not map exactly to theoretical concepts, but only digital trace data allow researchers to economically revisit their choice. Gelman and Loken (2013:10) referred to the additional uncertainty that arises from this decision-making process as the “garden of forking paths.” Even with solid theoretical grounding, navigating the pathway from conception to operationalization to analysis and interpretation is difficult. In a digital context, the journey down this path may not be linear, but this challenge may also be interpreted as an opportunity to expand researchers’ investigative potential.

Another limitation of using digital trace data is that it is not typically representative of populations to which demographic researchers often seek to generalize. Some platforms

(such as cell phone technology) may be more pervasive than others, but not everyone uses—and is therefore not captured by—technologies that archive digital trace data (Graham et al. 2012). A keyword-based sample drawn from Facebook or Twitter would contain only those individuals with regular Internet access and who also elect to provide information on the topic of interest. These selection mechanisms introduce important biases into the data that limit the conclusions that may be drawn from them.

The issue of representativeness, however, is not a new problem, nor is it unique to digital traces. Bias may arise, for instance, when using standard survey procedures, such as phone-based sampling, which represents only non-institutionalized populations (Pettit 2012). Moreover, as the percentage of households with a landline is decreasing in the United States, traditional sampling methods that rely on phone-based interviews may become increasingly challenging. Because of these factors, well-used data sources such as the General Social Survey and National Health Interview Survey, among others, are subject to issues of representativeness. A growing awareness of this challenge means, however, that researchers have begun to develop poststratification techniques that allow them to draw inference about populations, using nonrepresentative data. For example, recent research related to surveying nonrepresentative Xbox users about their intentions to vote has offered promising results (Wang et al. 2015). Demographers have developed approaches to correct for bias when ground truth is known in traditional albeit imperfect data (Alkema et al. 2012). Although the statistical problems are more complex for data sampled from digital platforms where ground truth information may not exist, this challenge is an opportunity to develop new methods.

In some research contexts, the nonrepresentativeness of the data is key for the research design. For example, in a number of situations, the key question is whether new forms of media and communication reflect existing social structures or are drivers of change. For instance, we may consider whether patterns of same-race connectedness within friendship networks seen offline are perpetuated regardless of social context using data drawn from social media sites (Cesare et al. 2017). Analogously, we can study the effect of online dating websites on homophily with respect to dating, cohabitation, and marriage. Comparing behavior online with observed patterns offline could help us understand the implications of online dating websites on patterns of assortative pairing (Rosenfeld and Thomas 2012).

Mechanisms involved in data collection processes, as well as the infrastructure and design of the platform of interest, also have the potential to introduce bias into digital trace data. This limitation, although important to acknowledge, does not prohibit quality research. Querying Twitter's streaming API, for example, has historically provided users with a small sample of query results, not the full set of tweets containing that query.[2] Additionally, researchers must make decisions about how to design queries to capture data from APIs, determining which data are included and excluded. A common query strategy for those studying social processes using Twitter is to sample data based on *hashtags* (metadata that users add to categorize content). Prominent, highly used hashtags—those that researchers might be aware of and use in their query—are generally only those that gained success for one reason or another in the environment, and their use may render research vulnerable to social trends

[2]See https://developer.twitter.com/en/docs/tweets/filter-realtime/overview for an overview of Twitter's public streaming APIs.

that lie outside its focus (Tufekci 2014). Furthermore, the resulting data may not represent the entirety of relevant content or stakeholders. Researchers should acknowledge these limitations, make efforts to bound their conclusions to accommodate the data at hand, and strive to develop methods to model and understand biases.

In addition to the issue of representativeness, researchers have noted the difficulty of engaging in qualitative research using digital trace data. The large scale of digital traces (e.g., hundreds of millions of social media posts) prevents the use of many traditional qualitative methods that require human inspection of each data element. Some researchers have used automated analysis strategies, such as topic modeling, to better understand data content (Reeder et al. 2014; Zeng et al. 2016). Although convenient, these methods do raise concerns. For example, using automated text analysis techniques on large textual data provides only a partial understanding of the meaning embedded within this content. And although analyzing word counts or using machine learning techniques (i.e., methods in which algorithms are trained to recognize user traits based on patterns inferred from input data) to categorize text may provide a rough illustration of ideas covered within a corpus of data, they do not provide a deep understanding of the processes that generated them or of user intent. Similarly, a “like” on Facebook or a retweet on Twitter may mean significantly different things given the content and context of the post involved, so using these measures at a large scale may obfuscate some data richness (Tufekci 2014).

Barring the availability of a large team of coders, very large data sets of digital traces preclude a thorough qualitative analysis of their full content. However, researchers should not dismiss qualitative analysis as a technique for understanding their data. Although it may be tantalizing to use an entire corpus containing millions of tweets, it may be more helpful to qualitatively code a small, randomly selected subset of those millions of tweets in order to understand the nuance of content within this space (Andrews et al. 2016; Tufekci 2014). Likewise, it is important to note that the field of automated text analysis is advancing rapidly. Computer scientists who specialize in natural language processing (NLP) are making strides within a range of core challenges, such as text parsing, information extraction, machine translation, modeling and processing social media text, analyzing linguistic style, and jointly modeling language and vision.

The use of digital trace data presents ethical challenges as well. Despite the easy accessibility to digital trace data, their use is not always ethical (Boyd and Crawford 2012). Indeed, most social media users value online privacy (Madden and Rainie 2015) and usually do not suspect that their information will be used for research purposes (Vitak 2015). Although researchers may take steps to ensure the anonymity of users within their study, it is often easy to link fingerprint-like user metadata to specific individuals. This challenge of anonymity is illustrated most clearly by the Taste, Ties, and Time (T3) project, a data collection initiative that gathered an entire university cohort’s worth of Facebook profile data but was disrupted when those assessing the research were able to identify individual students (Zimmer 2010). Privacy and protection in data use, however, extend beyond the individual. Designating ethical standards for data use also includes ensuring that vulnerable groups are protected from identification and possible discrimination using digital trace data (Taylor et al. 2017).

Ethical management of digital trace data is complicated by a varying landscape of data ownership. For example, Facebook electively shares user data with approved third-party partners from domains including advertising, law enforcement and academia.[3] Twitter, on the other hand, makes clear that users are responsible for their public information and that these data are accessible to anyone via Twitter's API.[4] The premise of withholding data may be to protect the privacy of the individuals who produce it, but such nondisclosure also inhibits exploration from the scientific community. Project OPAL[5] (Open Algorithms) is an excellent example of an initiative designed to balance the need for individual privacy with the provision of scientific opportunity. This project seeks to provide access to transparent algorithms and secure, fully anonymized, formatted data that, given their size and nature, may leave users vulnerable to breaches of privacy but that could benefit researchers and policymakers because of the content. The suite of tools developed by OPAL will be replicated by other organizations.

Beyond privacy and protection, the use of digital trace data invites novel concerns regarding procedural standards for ethical human subjects research. Standards such as informed consent may be impossible to implement when managing sets of participants that range in the tens of thousands or millions. Likewise, given recent evolution in how individuals view and understand digital privacy, it may be difficult to assess the risks and benefits of research that uses these novel data sources until after the research is conducted or published. The National Research Council has proposed modifying the definition of human subjects research to "a systematic investigation designed to develop or contribute to generalizable knowledge by obtaining data about a living individual directly through interaction or intervention, or by obtaining identifiable private information about an individual" (NRC 2014:40; recommendation 2.1). However, rules regarding informed consent do not apply to data that are anonymized or collected via a third party and may not change the ethical management of many Big Data sources (Lazer and Radford 2017). Researchers who use digital trace data for social research and are well trained in the ethics of human subjects research must be aware of these challenges and actively contribute to discussions regarding their ethical use.

Finally, collecting, storing, and managing digital trace data sets can present formidable barriers for many demographers. In particular, using such data in research activities requires technical skills not currently offered as part of most graduate training in the social sciences. As mentioned previously, representativeness and sampling pose significant challenges for researchers interested in using these data, and these biases limit the applicability of popular probabilistic statistical techniques to these data. Although we believe that these skills can be incorporated into existing pedagogy, they are often learned as a result of isolated researchers' initiatives to obtain skills through self-directed study. Many institutions, however, are taking steps to promote interaction between demographers and computer scientists as well as to promote awareness of modern data science techniques for reproducible research. One example of this effort is the International Union for the Scientific

[3]See Facebook's Data Policy: https://www.facebook.com/policy.php.
[4]See Twitter's Privacy Policy for more information: https://twitter.com/en/privacy.
[5]See https://www.opalproject.org/about-opal for more information.

Study of Population (IUSSP) scientific panel "Big Data and Population Processes," which currently offers training workshops at population and social media/informatics conferences.[6]

Overall, it is critical that social and demographic researchers engage in dialog regarding the proper use and application of digital trace data. A survey of more than 9,000 social scientists conducted by Sage Publishing found that the majority of scientists (81 %) believe that finding collaborators whose skills and interests complement their own is the greatest barrier toward completing digital data research (Metzler et al. 2016). Resources must be available to ensure researchers are (1) adept at programing and computational methods, (2) willing to be transparent about their methods in order to ensure reproducibility, and (3) able to work and communicate within an interdisciplinary setting. Some universities have created environments that welcome social scientists interested in enhancing their understanding of computational methods. The eScience Institute at the University of Washington and the Matrix at the University of California–Berkeley are examples of innovative centers designed to foster social science collaboration and promote innovative approaches to analyzing social data. Centers like these play a critical role in researchers' collective ability to overcome the methodological and ethical challenges that the use of digital trace data presents.

Digital Traces in Demographic Research: Existing Work and Areas of Development

Interest in the use of digital trace data among demographers is growing, as evidenced by multiple sessions in recent Population Association of America (PAA) meetings (Blumenstock and Toomet 2014; Cesare et al. 2015; Kashyap et al. 2017; Massey 2016; Mateos and Durand 2014; Reeder et al. 2014; Rosello and Filgueira 2016; Williams et al. 2015; Zagheni et al. 2017); recent publications (Blumenstock 2012; Blumenstock and Eagle 2012; Malik and Pfeffer 2016; Mendieta et al. 2016; Palmer et al. 2013; Stevenson 2014; Willekens et al. 2016; Zagheni and Weber 2012; Zagheni et al. 2014); and special issues of relevant social science journals, such as *Social Science Research*. Big Data research appears within the journal *Demography* as well, as illustrated by Barry's (2006) analysis of interracial friendship using wedding photos posted online and Palmer et al.'s (2013) work on spatial mobility with data collected via a smartphone app. Although work such as Barry's (2006) stands out in its innovation and novelty, little research has directly built on this contribution. Overall, the technical capacity to use digital traces for population studies among demographers lags years behind similar work in other fields, such as computer science. Relatedly, relevant demographic research has often appeared in outlets that are not traditional demographic journals, such as conference proceedings in the area of social informatics. In this section, we briefly review the emerging literature on digital demography and provide a snapshot of the state of the art.

Fertility, Mortality, and Migration

Researchers have begun to use digital trace data to examine topics traditionally discussed within the context of demography—such as fertility, mortality, and migration—in new ways.

[6]See http://iussp.org/en/panel/big-data-and-population-processes for more information on IUSSP workshop events.

In regard to fertility, existing work has found that search data provide a reliable and accurate means of monitoring the fertility patterns of hard-to-reach populations. Reis and Brownstein (2010), for example, compared the volume of abortion-related searches in a particular area and the number of restrictions imposed upon abortions in the area. They found an inverse relationship between these measures, suggesting that those who live in areas where abortion is prohibited turn to the Internet to find out how to access these services elsewhere. These data would likely not be captured in traditionally collected data regarding abortion rates in a given area. Similarly, Billari et al. (2013) found that adjusted measures of Google search data (based on queries such as *ovulation or pregnancy*) can be used to make short-term predictions about national fertility trends. Ojala et al. (2017) combined data from Google Correlate/Google Trends and the American Community Survey (ACS) to study socioeconomic differences on the circumstances surrounding pregnancy and birth. Studies of fertility using digital traces need not limit themselves to search data, however. Blogs and microblogs such as Twitter provide unsolicited information about maternal and reproductive health as well (De Choudhury et al. 2013a, 2013b).

Some studies of mortality have leveraged digital traces. Tomlinson et al. (2009) suggested that sending short surveys via mobile devices may be an effective means of tracking health behaviors and instances of mortality among difficult to reach—often rural—populations. Tamgno et al. (2013) showed that cell phones may be used as a tool for conducting verbal autopsies and understanding mortality conditions in hard-to-reach populations. Similar to studies of fertility, analyses of this sort need not be limited to one data source. Details related to health and mortality can potentially be explored via other sources, such as search queries, social media data (from sites such as Twitter, Tumblr, or Facebook) (Eichstaedt et al. 2015), or other forms of archived digital data (Tourassi et al. 2016).

Digital traces have been used extensively to examine migration as well. Blumenstock (2012) used mobile phone data to track within-country migration in rural Rwanda as a means of improving the reach and application of social programs in that country. Similarly, Deville et al. (2014) proposed methods of calibrating cell phone data that produced information on intra- and international mobility patterns that is as detailed or more detailed than traditionally collected survey data. Taking a different approach, Palmer et al. (2013) used cell phone surveys to study micro-interactions and examine social processes within activity spaces rather than residential census units. Although cell phone data are valuable for studies of migration, other geotagged digital traces may also be used to examine human mobility: georeferenced Yahoo! e-mail data to estimate profiles of international migration by age and sex (Zagheni and Weber 2012), geolocated Twitter tweets for the study of short-term migrations in OECD countries (Zagheni et al. 2014), LinkedIn information about professional histories to evaluate trends in international migrations of professionals (State et al. 2014), and networks of Skype calls to track international migrations (Kikas et al. 2015).

Augmenting Traditional Survey Data With Digital Trace Data

Some researchers have advocated that combining digital trace data with systematically collected survey data can add much-needed dimensionality to data-rich but variable-poor digital traces. Snijders et al. (2012), for example, proposed combining digital traces with

survey information drawn from individuals within the sample and/or general information about the platform from which the data were collected in order to better understand the micro-level social processes that produced the data collected. Lazer et al. (2014) found that combining Google flu trends data and Centers for Disease Control (CDC) data would produce more accurate flu predictions than either source alone. De Choudhury et al. (2013a, 2013b) and De Choudury et al. (2016) combined digital data with other data sources to predict depression and food insecurity, respectively. Blumenstock and Eagle (2012) combined call record data with household survey data to examine disparities in mobile phone access and usage.

Additionally, combining survey data with social media data can then be used to validate coding methods when no surveys are available. For example, Cesare et al. (2015) used survey data containing self-reported demographic information linked to Twitter data to examine trends in self-presentation. Similarly, Moreno et al. (2012) correlated photo displays and reports of drinking behavior on Facebook with self-reported alcohol consumption indicators from a linked survey. We believe that this approach of augmenting digital traces with other, more traditional sources of data is a promising direction. However, we also emphasize the importance of being aware of methodological issues that may arise in matching the units of analysis between the data sources used.

Adding Demographic Dimensions to Digital Trace Data

Researchers can aggregate and disaggregate information embedded within digital traces in unique and interesting ways. Profile photos contained within big data sets, for example, often contain demographic information generally not reported on individuals' profiles, such as the age, race, and gender of a user. Existing work has found that crowd-sourced human intelligence can be used to accurately and reliably extract valuable information from these photos (McCormick et al. 2015). Similarly, Zagheni et al. (2014) used facial recognition software to add demographic information to Twitter data as a means of tracking demographic trends in international and internal migration. Others have combined metadata containing users' first and/or last names with other data sources such as the U.S. Census to estimate users' demographic characteristics (e.g., Mislove et al. 2011).

Accounting for Bias

A number of scholars have been developing methods to account for the bias created by the use of nonrepresentative samples when baseline population data are both known and unknown. In regard to the former, Zagheni and Weber (2012) used e-mail data to measure rates of international migration. To work with these nonrepresentative data, they developed a method of scaling their estimates to account for bias introduced by variability in Internet penetration rates across space and demographic groups. In the context of nonrepresentative polls, Wang et al. (2015) used multilevel regression and poststratification based on respondents' demographic characteristics to predict election outcomes, using Xbox as their survey tool. Their predictions were extremely similar to results from nationally representative data both nationwide and state by state, thus illustrating that data drawn from convenience samples—such as sources of digital data—can provide valuable information in a quicker and more cost-effective way than traditional survey methods. Developing methods

for drawing conclusions from abundant and unsolicited yet unrepresentative digital trace sources, however, is an area with significant room for methodological innovations from social scientists and statistical demographers.

Beyond the opportunity for methodological contributions, researchers can directly address the limitations of digital trace data by appropriately bounding the conclusions drawn from them. When analyzing data from a particular social media site, for instance, a researcher should specify that conclusions may be context-dependent and in some way connected to the characteristics of the platform used. However, sampling from one context is not a challenge unique to the use of digital trace data. Social scientists have gained extensive insight on the role of neighborhoods for multiple behavioral and social outcomes using data from only a few cities, such as Chicago (Park and Burgess 1925; Shaw and McKay 1942). These cities are not representative of the entire United States, but the authors noted that the behaviors and attitudes of these individuals can provide insights on the behaviors and attitudes of other individuals in similar contexts across the country.

“Digital Census”: Facebook Ads Manager Data as a Case Study

In this section, we present an illustrative example of the use of digital trace data by examining Facebook data for advertisers. We discuss how existing work has used these data, address the characteristics of methods needed to extract meaningful information from digital traces such as these, and share an example of demographic analysis that leverages digital trace data. Although some forms of digital data, such as Facebook advertisement data, are new in terms of format and content, most of their associated challenges are similar to those of data sources that demographers have analyzed in the past.

The Facebook Ads Manager[7] enables advertisers to select detailed demographic characteristics of the users to whom the ads should be shown. Before the ad is launched and the advertiser is billed, Facebook offers an estimate of the selected audience size. This information is (as of this writing) free and can be accessed in a programmatic way via the Facebook API.[8] Such a cost estimate is useful for advertisers when planning ad campaigns and developing an appropriate budget, or when deciding whether to narrow or broaden a target audience. Given that online advertisers are primarily interested in understanding the characteristics of their user base, the same information is also useful for researchers who can access what is essentially a digital census of more than 2 billion Facebook users and freely obtain aggregate-level measures of demographic characteristics as well as topical interests.

Facebook’s Ads Manager provides population estimates from large nonrepresentative samples, but bias analysis in the estimation of demographic quantities is at the core of the discipline of demography. Many models and techniques have been developed to address issues that range from measurement error to stochasticity, undercounting, and various dimensions of data imperfection. Estimating and correcting for bias in Facebook data is a crucial step toward extracting information from these data. Zagheni et al. (2017) used data

[7]Find more at http://www.facebook.com/business/.
[8]See https://developers.facebook.com/docs/marketing-api/audiences-api.

from Facebook Ads Manager to estimate stocks of migrants in the United States and to understand biases in the population of Facebook users. For each combination of age, sex, country of origin, and U.S. state of destination, they examined the difference between the fraction of foreign-born individuals estimated by the ACS and the respective quantity for Facebook users. The discrepancy between the two estimates (i.e., the bias) was then modeled using a linear regression framework to evaluate the extent to which patterns emerged. For example, using a model in which the bias was regressed against a series of indicator variables for different demographic groups, countries of origin, and U.S. states of destination, Zagheni et al. (2017) found important regularities in profiles of migrants by age and sex across U.S. states of destination or countries of origin. The authors then leveraged these regularities to improve predictions. Their approach relied on combining traditional data sources (e.g., the ACS) and new emerging ones (e.g., Facebook data for advertisers) to generate timely and geographically granular estimates in developed countries. In the context of developing countries, sparse data could be triangulated to potentially improve estimates of demographic rates.

Although bias adjustment in the context of social media data analysis is relatively new, approaches for evaluating and correcting biases have been used by demographers in nonsocial media contexts. For example, Alkema et al. (2012) used a regression model with indicator variables for data quality to estimate trends in total fertility rates using imperfect data from West Africa. Ševčíková et al. (2007) proposed a statistical model to evaluate and correct for biases in simulation outcomes. Similar approaches could be repurposed in the context of demographic estimation with social media data. We believe that there is room for the development of appropriate Bayesian models that allow researchers to combine a number of sources of information within a solid statistical framework while also borrowing strength across groups with similar features and leveraging the overall structure of the data, which is often hierarchical.

Facebook Ads Manager can also be used to survey hard-to-reach populations or groups for which there is not a register or clear sampling frame. Pötzschke and Braun (2016), for instance, used Facebook ads to sample Polish migrants in Austria, Ireland, Switzerland, and the United Kingdom. Facebook users who matched specific criteria were targeted with Facebook ads that invited them to participate in a survey. With a sample of more than 1,100 individuals who completed an extensive questionnaire, they showed that their approach was cost-effective and efficient. More generally, this is an example of a survey that uses nonrepresentative samples and requires poststratification techniques in order to weight the respondents and make statistical inferences about the underlying population. This is an area of active research where demographers and social scientists can make important contributions (see, e.g., Wang et al. 2015). Moreover, this research presents a challenge for which there is precedent given that it is related to issues that traditional phone-based surveys currently face: decreasing response rates that are nonrandom, and increasing numbers of households without landlines and are thus excluded from the samples.

A tool like Facebook Ads Manager can be used to design experimental setups in order to gain insights into the processes that drive population health. For example, Araújo et al. (2017) tracked the size of audiences in Facebook with interests that could be markers of

tobacco use, obesity, or diabetes. Although their results were negative—meaning that they found that differences in interest audiences were only weakly indicative of the corresponding prevalence rates—they developed an analysis approach that can be used in other contexts. More specifically, they compared differences in a specific set of interests related to health, with differences in other placebo interests unrelated to health. The premise of creating a baseline was to control for the amount of time and number of searches people generally conduct on Facebook. In other words, Araújo et al. (2017) used a form of normalization for behavioral features of Facebook users. Developing tools to standardize compositional changes in the population of Facebook users is another area where demographers can develop methodological advances.

For a concrete example of how traditional demographic methods may be applied to digital trace data, consider the following illustrative case. Say that a researcher is interested in determining whether differences in the educational attainment of Facebook and LinkedIn users is driven by the age composition or the degree rate schedules of each site. If LinkedIn users appear to be more highly educated than Facebook users, how much of that difference is attributable to LinkedIn users falling into a different age range than Facebook users versus to true educational differences between LinkedIn and Facebook users? If each site is considered a population, then a simple age decomposition analysis provides the answer.

To illustrate, we used Facebook Ads Manager and LinkedIn Campaign Manager[9] to obtain age-specific population estimates for each site (see Table 2). Facebook allows advertisers to specify the type of campaign they wish to design (e.g., designed to increase store visits, video views, or clicks). Given our aim of selecting as broad an audience as possible within our criteria, we opted for a “reach” campaign, an awareness-based campaign designed for audience breadth. We then requested population estimates for users located within the United States within specific age ranges as well as subsets of these groups that Facebook identifies as college graduates. We conducted a similar selection process on LinkedIn, but because LinkedIn provides more detailed educational information than Facebook, we selected users who have any one of a variety of undergraduate degrees (e.g., BA, BS, BFA). Using the estimates obtained, we generated a crude educational attainment rate for both sites: 0.33 for Facebook and 0.61 for LinkedIn. We then used the educational rate schedules for each age group and the proportion of users who fall within each age group as input for an age decomposition analysis, which we based on methods outlined by Preston et al. (2001).[10] Results indicate that the difference in educational attainment across these sites is mostly attributable to their differences in educational rate schedules (~98.9 %) and only partially to age composition of users (~1.1 %).

Although most digital traces are forms of imperfect data, ignoring them would be a missed opportunity for demographers (Billari and Zagheni 2017). We expect that these data will become routinely used in research. As demographers understand the problems and opportunities connected with these data, the use of digital traces in demographic research will become common practice.

[9]See https://www.linkedin.com/ad/accounts.
[10]See Preston et al. (2001:28–30) for details on conducting an age decomposition analysis.

Discussion

Demographers have always been data scientists and have a history of using innovative and creative techniques to work with challenging data. John Graunt, largely considered the father of demography, produced the very first life tables in the seventeenth century by leveraging data collected for marketing purposes to estimate the age distribution of potential customers in London. Today, new types of repurposed digital traces provide opportunities for advances in the field. In this article, we review the state of the art uses of digital trace data in demography and highlight what we believe are the current challenges and opportunities within this research arena.

Digital traces have captured the attention of social scientists and demographers for many reasons. In a global context of civil registration systems where an estimated two-thirds of all annual deaths and almost one-half of the world's children are not registered, digital traces offer some hope that alternative data could complement existing ones to provide important estimates about fundamental demographic processes, such as fertility, mortality, and migration. These hopes are sustained by the observation that certain types of technology, such as mobile phones, are ubiquitous even in developing countries. Analogously, Internet penetration rates are likely to increase at a faster pace than the development of mature civil registration systems. Because demographers have traditionally dealt with imperfect data in the context of developing countries, they are well suited to lead advances in the development of methods to leverage digital trace data. The work of Brass (1976) related to indirect estimation in Africa is a testament to past achievements as well as a source of inspiration for future developments in a new data landscape.

Our objective in this commentary is to highlight the importance of digital traces within social science and demographic research, summarize common critiques offered against the use of these data, address which of these critiques are most salient to those interested in using digital trace data for demographic research, and discuss how researchers might overcome the challenges raised within these critiques. The use of digital trace data for social science and demographic research has incredible potential, and many points raised against these data sources are not insurmountable challenges but are instead opportunities to advance these fields methodologically.

We believe that graduate and postdoctoral training and interdisciplinary collaboration is key to increasing the accessibility of digital trace data to demographic researchers. As González-Bailón (2013:158) suggested, " … social scientists can no longer do research on their own: the scale of the data that we can now analyze, and the methods required to analyze them, can only be developed by pooling expertise with colleagues from other disciplines." Using digital traces requires a strong and varied set of technical and computational skills, but these skills alone cannot effectively leverage these data for demographic research. A true collaborative effort requires the input of researchers who can develop creative and effective ways of using digital trace data to answer relevant and interesting social science questions.

Digital trace data have the potential to answer long-standing questions in new and innovative ways. For instance, because social data are traditionally gathered using surveys, we do not

have a clear understanding of the speed at which behaviors and attitudes change or the sorts of social factors that influence this change. With sources of digital traces such as social media sites, however, we are able to not only document ties between individuals but also examine how these patterns of ties change on a minute-by-minute basis. Indeed, the real-time generation of digital traces has been used already to examine time-sensitive trends, such as how moods change over time (Golder and Macy 2011) or reactions to crisis events (Andrews et al. 2016; Starbird et al. 2016), and there are opportunities to use the capacity of these data to examine how networks change over time. Leveraging digital traces as a tool for examining how patterns of connectedness evolve and developing ways of extracting information about the users within these associative networks has the potential to expand opportunities for demographic research in relevant and innovative ways.

It may be said that the structure and size of digital trace data—as well as the manner in which they are generated—has changed the relationship between theory and data in regard to how they drive scientific discovery. Typically, demographic research follows a strict two-stage process: (1) a *discovery* stage, which uncovers unique patterns within population data; and (2) an *explanation* stage, which hypothesizes and tests how behavior creates the population patterns observed in the second (Billari and Zagheni 2017). The introduction of new data sources, however, has the potential to disrupt the interaction of these stages. Digital trace data are decentralized and produced in real time, which means that researchers with very different backgrounds and training can access them. They are also logistically difficult to manage and analyze and are subject to biases reflective of the users and platforms generating them. Managing these unique traits invites a dialog between those who are generating and testing hypotheses about patterns observed (something similar to the explanation stage) and those with the skills to analyze and illustrate patterns in the data (something similar to the discovery stage). This largely parallels broader patterns in the field of data science described by Blei and Smyth (2017), in which more theoretically motivated statistical approaches to data analysis and data-driven computational approaches now complement one another. As a result of the availability of digital data, the pipeline of acquiring, processing, visualizing, and analyzing data is less linear and requires greater human discretion than before (Blei and Smyth 2017)

Demography as a field clearly stands to benefit from the use of digital trace data. What is less obvious is that demographers could make contributions that go beyond the boundaries of their discipline. It may be said that the users of digital tools form populations. New social media users are “born” when they sign up for a service, and they “die” when they stop using it. By adopting this conceptualization, standard demographic tools can be adapted and standardized to gain insights into populations of digital spaces that are of interest to disciplines such as media studies or communication. For example, multistate life tables could be built to quantify dynamics of survival within a platform. Similarly, the growth rate of the user base of a specific service could be estimated from a sample of users for whom we know the “age,” expressed in years since the date they signed up for the service (e.g., for work on estimating population growth from the U.S. Census, see Keyfitz and Caswell 2005). Feehan and Cobb (2017) illustrated this possibility by taking a census of Internet users and by surveying Facebook users about which of their friends are also online. Overall, many

classic demographic tools could be applied to understand populations of digital objects, with influences on fields beyond demography.

Conclusion

Most demographers are interdisciplinary by design: they have one foot in the area of demographic methods and one foot in a different but related discipline such as sociology, economics, geography, statistics, public health, public policy, anthropology, and others. Historically, the field of demography has been invigorated by exchange and collaboration with many other disciplines. Demographers have drawn ideas from, and made substantive contributions to, a number of academic fields. However, the relationship between demographers and data scientists has not fully developed yet. We believe that the data science revolution is opening new doors for mutually rewarding collaborations among demographers, computer scientists, and researchers broadly involved in the area of social informatics.

This commentary is a response to the growing use of digital traces, particularly social media data and cell phone data, in demography as well as work providing critiques of these applications. Our intent is to interpret how the field might evolve to accommodate the use of these data. We argue that demographers are well positioned to address the main challenges presented by digital trace data and to seize important methodological opportunities that these challenges open. For one, in a data-driven world, demographers possess the skills needed to develop methods to extract useful information from large but often noisy, messy, and nonrepresentative data. Additionally, demographers can draw on an arsenal of classic demographic methods to study digital traces that represent subsets of populations. In sum, we argue that demographers, who have been relatively slow to contribute to the study of digital trace data, could become primary innovators in this area.

Acknowledgments

This work is supported by Grants DMS-1737673 and SES-1559778 from the National Science Foundation and K01 HD078452 from the National Institute of Child Health and Human Development (NICHD). This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-12-1-0379, and by the Washington Research Foundation. We also appreciate the support of the Earl and Edna Stice lectureship in the Social Sciences; the University of Washington Information School, Center for Statistics and the Social Sciences; eScience Institute; and Sociology Department for supporting speakers at the frontier of data science in demographic research.

References

Adams J, & Brueckner H (2015). Wikipedia, sociology, and the promise and pitfalls of Big Data. Big Data & Society, 2(2), 1–5. https://doi.org/10.1177/2053951715614332

Alkema L, Raftery AE, Gerland P, Clark SJ, & Pelletier F (2012). Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa. Demographic Research, 26(article 15), 332–361. https://doi.org/10.4054/DemRes.2012.26.15

Andrews C, Fichet E, Ding Y, Spiro ES, & Starbird K (2016). Keeping up with the Tweet-dashians: The impact of “official” accounts on online rumoring In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (pp. 452–465). New York, NY: ACM https://doi.org/10.1145/2818048.2819986

Ang C, Bobrowicz A, Schiano D, & Nardi B (2013). Data in the wild: Some reflections. Interactions, 20(2), 39–43. https://doi.org/10.1145/2427076.2427085

Araújo M, Mejova Y, Weber I, & Benevenuto F (2017). Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations In Proceedings of the 2017 ACM on Web Science Conference (pp. 253–257). New York, NY, ACM https://doi.org/10.1145/3091478.3091513

Barberá P (2016). Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data (Working paper). New York: Center for Data Science, New York University Retrieved from http://pablobarbera.com/static/less-is-more.pdf

Barry B (2006). Friends for better or for worse: Interracial friendships in the United States as seen through wedding photos. Demography, 43, 491–510. [PubMed: 17051824]

Belli RF, Traugott MW, Young M, & McGonagle KA (1999). Reducing vote overreporting in surveys: Social desirability, memory failure, and source monitoring. Public Opinion Quarterly, 63, 90–108.

Berinsky AJ (1999). The two faces of public opinion. American Journal of Political Science, 43, 1209–1230.

Blei DM, & Smyth P (2017). Science and data science. Science, 114, 8689–8692.

Billari FC, D'Amuri F, & Marcucci J (2013, 4). Forecasting births using Google Paper presented at the annual meeting of the Population Association of America, New Orleans, LA.

Billari FC, & Zagheni E (2017). Big data and population processes: A revolution? In Petrucci A & Verde R (Eds.), SIS 2017. Statistics and Data science: New challenges, new generations. 28–30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society (pp. 167–178). Firenze, Italy: Firenze University Press https://doi.org/10.17605/OSF.IO/F9VZP

Blumenstock J, Cadamuro G, & On R (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350, 1073–1076. [PubMed: 26612950]

Blumenstock J, & Eagle N (2010). Mobile divides: Gender, socioeconomic status, and mobile phone use in Rwanda In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (pp. 6–10). New York, NY: ACM https://doi.org/10.1145/2369220.2369225

Blumenstock JE (2012). Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda. Information Technology for Development, 18, 107–125.

Blumenstock JE, & Eagle N (2012). Divided we call: Disparities in access and use of mobile phones inRwanda. Information Technologies and International Development, 8(2), 1–16.

Blumenstock JE, & Toomet O (2014, 5). Segregation and "silent separation": Using large-scale network data to model the determinants of ethnic segregation Paper presented at the annual meeting of the Population Association of America, Boston, MA.

Boyd D, & Crawford K (2012). Critical questions for big data. Information, Communication & Society, 15,662–679.

Brass W (1976). Indirect methods of estimating mortality illustrated by application to Middle East and North African data In Population Bulletin of the United Nations Economic Commission for Western Asia. Amman, Jordan: UNECWA.

Cesare N, Spiro E, & Lee H (2015, 4). Self-presentation and information disclosure on Twitter: Understanding patterns and mechanisms along demographic lines Paper presented at the annual meeting of the Population Association of America, San Diego, CA.

Cesare N, Lee H, McCormick TH, & Spiro ES (2017). Redrawing the silent "color line": Examining racial segregation in associative networks on Twitter Unpublished manuscript, University of Washington, Seattle, WA Retrieved from arXiv:1705.04401.

Couldry N, & Powell A (2014). Big Data from the bottom up. Big Data & Society, 1(2), 1–5. https://doi.org/10.1177/2053951714539277

De Choudhury M, Counts S, & Horvitz E (2013a). Predicting postpartum changes in emotion and behavior via social media In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 3267–3276). New York, NY: ACM.

De Choudhury M, Gamon M, Counts S, & Horvitz E (2013b). Predicting depression via social media In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (pp. 128–137). Palo Alto, CA: AAAI Press Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351

De Choudhury M, Sharma S, & Kiciman E (2016). Characterizing dietary choices, nutrition, and language in food deserts via social media In Proceedings of the 19th ACM Conference on

Computer-Supported Cooperative Work & Social Computing (pp. 1157–1170). New York, NY: ACM https://doi.org/10.1145/2818048.2819956

Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, … Tatem AJ (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111, 15853–15854.

Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, … Seligman MEP (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological Science, 26, 159–169. [PubMed: 25605707]

Fadnes L, Taube A, & Tylleskär T (2009). How to identify information bias due to self-reporting in epidemiological research. Internet Journal of Epidemiology, 7(2), 1–8.

Feehan D, & Cobb C (2017, 7). How many people have access to the Internet? Estimating Internet adoption around the world using Facebook. Paper presented at the International Conference on Computational Social Science, Cologne, Germany.

Felt M (2016). Social media and the social sciences: How researchers employ Big Data analytics. Big Data & Society, 3(1), 1–15. https://doi.org/10.1177/2053951716645828

Gelman A, & Loken E (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time Unpublished manuscript, Department of Statistics, Columbia University, New York, NY Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Golder SA, & Macy MW (2011). Diurnal and seasonal mood vary with work, sleep and daylength across diverse cultures. Science 30, 1878–1881.

Golder SA, & Macy MW (2014). Digital footprints: Opportunities and challenges for online social research. Annual Review of Sociology, 40, 129–152.

González-Bailón S (2013). Social science in the era of big data. Policy and the Internet, 5, 147–160.

Graham M, Hale S, & Stephens M (2012). Featured graphic: Digital divide: The geography of Internet access. Environment and Planning, 44, 1009–1010.

Heaivilin N, Gerbert B, Page JE, & Gibbs JL (2011). Public health surveillance of dental pain viaTwitter. Journal of Dental Research, 90, 1047–1051. [PubMed: 21768306]

Holbrook AL, & Krosnick JA (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly, 74, 37–67.

Kashyap R, Billari FC, Cavalli N, Quian E, & Weber I (2017, 4). Ultrasound technology and “missing women” in India: Analyses and now-casts based on Google searches. Paper presented at the annual meeting of the Population Association of America, Chicago, IL.

Keyfitz N, & Caswell H (2005). The matrix model framework In Keyfitz N & Caswell H (Eds.), Applied mathematical demography (3rd ed., pp. 47–70). New York, NY: Springer.

Kikas R, Dumas M, & Saabas A (2015). Explaining international migration in the Skype network In SIdEWayS ‘15: Proceedings of the 1st ACM Workshop on Social Media World Sensors (pp. 17–22). New York, NY: ACM https://doi.org/10.1145/2806655.2806658

Kitchin R (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 1–12. https://doi.org/10.1177/2053951714528481

Latour B (2007). Beware, your imagination leaves digital traces. Times Higher Literary Supplement, 6(4). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Beware+,+your+imagination+leaves+digital+traces#0

Lazer D, Kennedy R, King G, & Vespignani A (2014). The parable of Google flu: Traps in Big Data analysis. Science, 343, 1203–1205. [PubMed: 24626916]

Lazer D, & Radford J (2017). Data ex machina: Introduction to big data. Annual Review of Sociology, 49,19–39.

Lee H, Cesare N, McCormick TH, Morris J, & Shojaie A (2014, 5). Redrawing the “color line”: Examining racial homophily of associative networks in social media. Paper presented at the annual meeting of the Population Association of America, Boston, MA.

Lewis K (2015). Three fallacies of digital footprints. Big Data & Society, 2(2), 1–4. https://doi.org/10.1177/2053951715602496

Lewis SC, Zamith R, & Hermida A (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. Journal of Broadcasting & Electronic Media, 57, 1–33. https://doi.org/10.1080/08838151.2012.761702

Lohr S (2012, 2 11). The age of Big Data. The New York Times Retrieved from https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Madden M, & Rainie L (2015). Americans' attitudes about privacy, security and surveillance (Report). Washington, DC: Pew Research Center Retrieved from http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

Malik MM, & Pfeffer J (2016, 3). Social media data and computational models of mobility: A review for demography. Paper presented at the ICWSM Workshop on Social Media and Demographic Research, Cologne, Germany Retrieved from http://www.pfeffer.at/papers/2016_demography.pdf

Manovich L (2011). Trending: The promises and the challenges of big social data In Gold MK (Ed.), Debates in the digital humanities (pp. 460–476). Minneapolis: University of Minnesota Press.

Marwick AE, & Boyd D (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society, 13, 114–133. https://doi.org/10.1177/1461444810365313

Massey D (2016, 3). Measuring racial prejudice using Google trends. Paper presented at the annual meeting of the Population Association of America, Washington, DC.

Mateos P, & Durand J (2014, 5). Netnography and demography: Mining Internet discussion forums on migration and citizenship. Paper presented at the annual meeting of the Population Association of America, Boston, MA.

McCormick TH, Lee H, Cesare N, Shojaie A, & Spiro ES (2015). Using Twitter for demographic and social science research: Tools for data collection and processing. Sociological Methods & Research, 46, 390–421. [PubMed: 29033471]

Mendieta J, Su S, Vaca C, Ochoa D, & Vergara C (2016). Geo-localized social media data to improve characterization of international travelers In Proceedings of the 2016 Third International Conference on eDemocracy & eGovernment (ICEDEG) (pp. 126–132). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Metzler K, Kim DA, Allum N, & Denman A (2016). Who is doing computational social science? Trends in big data research (White paper). London, UK: SAGE Publishing https://doi.org/10.4135/wp160926

Mislove A, Lehmann S, & Ahn Y (2011). Understanding the demographics of Twitter users In Proceedings of the Fifth International Conference on Weblogs and Social Media (pp. 554–557). Menlo Park, CA: AAAI Press Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Moreno MA, Christakis DA, Egan KG, Brockman LN, & Becker T (2012). Associations between displayed alcohol references on Facebook and problem drinking among college students. Archives of Pediatrics & Adolescent Medicine, 166, 157–163. [PubMed: 21969360]

National Research Council (NRC) (2014). Proposed revisions to the common rule for the protection of human subjects in the behavioral and social sciences (Report). Washington, DC: National Academies Press.

O'Connor B, Balasubramanyan R, Routledge BR, & Smith NA (2010). From tweets to polls: Linking text sentiment to public opinion time series In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (pp. 122–129). Palo Alto, CA: AAAI Press Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1536/1842

Ojala J, Zagheni E, Billari FC, & Weber I (2017). Fertility and its meaning: Evidence from search behavior In Proceedings of the Eleventh International AAAI Conference on Web and Social Media (pp. 640–643). Palo Alto, CA: AAAI Press.

Palmer JRB, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, & Li K (2013). New approaches to human mobility: Using mobile phones for demographic research. Demography, 50, 1105–1128. [PubMed: 23192393]

Park RE, & Burgess EW (1925). The city. Chicago, IL: University of Chicago Press.

Pettit B (2012). Invisible men: Mass incarceration and the myth of black progress. New York, NY: Russell Sage Foundation.

Pew Research Center. (2018). Internet/broadband fact sheet. Washington, DC: Pew Research Center Retrieved from http://www.pewinternet.org/fact-sheet/internet-broadband/

Pötzschke S, & Braun M (2016). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. Social Science Computer Review, 35, 633–653.

Preston SH, Heuveline P, & Guillot M (2001). Demography: Measuring and modeling population processes. Oxford, UK: Blackwell.

Reeder H, McCormick TH, & Spiro E (2014). Online information behaviors during disaster events: Roles, routines, and reactions (Working Paper No. 144). Seattle, WA: Center for Statistics and the Social Sciences.

Reis BY, & Brownstein JS (2010). Measuring the impact of health policies using Internet search patterns: The case of abortion. BMC Public Health, 10(article 514). https://doi.org/10.1186/1471-2458-10-514

Rosello JLD, & Filgueira F (2016, 4). Big data in a small country: Integrating birth, maternal and child statistics in Uruguay. Paper presented at the annual meeting of the Population Association of America, Washington, DC.

Rosenfeld MJ, & Thomas RJ (2012). Searching for a mate: The rise of the Internet as a social intermediary. American Sociological Review, 77, 523–547.

Ruggles S (2014). Big microdata for population research. Demography, 51, 287–297. [PubMed: 24014182]

Ruppert E, Law J, & Savage M (2013). Reassembling social science methods: The challenge of digital devices. Theory, Culture and Society, 30(4), 22–46.

Sagiroglu S, & Sinanc D (2013). Big Data: A review In 2013 International Conference on Collaboration Technologies and Systems (CTS) (pp. 42–47). Piscataway, NY: Institute of Electrical and Electronics Engineers https://doi.org/10.1109/CTS.2013.6567202

Ševčíková H, Raftery AE, & Waddell PA (2007). Assessing uncertainty in urban simulations using Bayesian melding. Transportation Research, Part B: Methodological, 41, 652–669.

Shaw CR, & McKay HD (1942). Juvenile delinquency and urban areas. Chicago, IL: University of Chicago Press.

Smith A, & Anderson M (2018). Social media use in 2018. Washington, DC: Pew Research Center Retrieved from http://assets.pewresearch.org/wp-content/uploads/sites/14/2018/03/01105133/PI_2018.03.01_Social-Media_FINAL.pdf

Snijders C, Matzat U, & Reips U-D (2012). Big data: Big gaps of knowledge in the field of Internet science. International Journal of Internet Science, 7, 1–5. Retrieved from http://www.ijis.net/ijis7_1/ijis7_1_editorial_pre.html

Starbird K, Maddock J, Orand M, Achterman P, & Mason RM (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing In Kindling M & Greinfeneder E (Eds.), iConference 2014 proceedings (pp. 654–662). Urbana-Champaign, IL: iSchools.

Starbird K, Spiro E, Edwards I, Zhou K, Maddock J, & Narasimhan S (2016). Could this be true? I think so! Expressed uncertainty in online rumoring In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 360–371). New York, NY: ACM.

State B, Rodriguez M, Helbing D, & Zagheni E (2014). Migration of professionals to the U.S.: Evidence from LinkedIn data In Aiello LM & McFarland D (Eds.), 6th International Conference on Social Informatics, SocInfo 2014 (pp. 531–543). Cham, Switzerland: Springer.

Stevenson AJ (2014). Finding the Twitter users who stood with Wendy. Contraception, 90, 502–507. [PubMed: 25129330]

Sutton J, Spiro ES, Johnson B, Fitzhugh S, Gibson B, & Butts CT (2014). Warning tweets: Serial transmission of messages during the warning phase of a disaster event. Information, Communication & Society, 17, 765–787.

Tamgno JK, Faye RM, & Lishou C (2013). Verbal autopsies, mobile data collection for monitoring and warning causes of deaths In 14th International Conference on Advanced Communication Technology, Technical Proceedings, 2013 (pp. 495–501). Piscataway, NJ: Institute of Electrical and Electronics Engineers Retrieved from https://ieeexplore.ieee.org/document/6488236/

Taylor L, Floridi L, & van der Sloot L (Eds.). (2017). Group privacy: New challenges of data technologies.Cham, Switzerland: Springer.

Tomlinson M, Solomon W, Singh Y, Doherty T, Chopra M, Ijumba P, … Jackson D (2009). The use of mobile phones as a data collection tool: A report from a household survey in South Africa. BMC Medical Informatics and Decision Making, 9(article 51). https://doi.org/10.1186/1472-6947-9-51

Tourangeau R, & Yan T (2007). Sensitive questions in surveys. Psychological Bulletin, 133, 859–883. [PubMed: 17723033]

Tourassi G, Yoon HJ, & Xu S (2016). A novel web informatics approach for automated surveillance of cancer mortality trends. Journal of Biomedical Informatics, 61, 110–118. [PubMed: 27044930]

Tufekci Z (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (pp. 505–514). Palo Alto, CA: AAAI Press.

Vitak J (2015). I like it…. Whatever that means: The evolving relationship between disclosure, audience, and privacy in networked spaces [SlideShare presentation]. Retrieved from https://www.slideshare.net/jvitak/i-like-itwhatever-that-means-the-evolving-relationship-between-disclosure-audience-and-privacy-in-networked-spaces

Wang W, Rothschild D, Goel S, & Gelman A (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31, 980–991.

Willekens F, Massey D, Raymer J, & Beauchemin C (2016). International migration under the microscope. Science, 352, 897–899. [PubMed: 27199405]

Williams NE, Thomas TA, Dunbar M, Eagle N, & Dobra A (2015). Measures of human mobility using mobile phone records enhanced with GIS data. PLoS One, 10(7), 1–16. https://doi.org/10.1371/journal.pone.0133630

Zagheni E, Garimella VRK, Ingmar W, & State B (2014). Inferring international and internal migration patterns from Twitter data In Proceedings of the 23rd International Conference on World Wide Web (pp. 439–444). New York, NY: ACM Press https://doi.org/10.1145/2567948.2576930

Zagheni E, & Weber I (2012). You are where you e-mail: using e-mail data to estimate international migration rates In Proceedings of the 4th Annual ACM Web Science Conference (pp. 348–351). New York, NY: ACM https://doi.org/10.1145/2380718.2380764

Zagheni E, Weber I, & Gummadi K (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. Population and Development Review, 43, 721–734.

Zeng L, Starbird K, & Spiro ES (2016). Rumors at the speed of light? Modeling the rate of rumor transmission during crisis In 49th Hawaii International Conference on System Sciences (HICSS), 1969–1978. Piscataway, NJ: Institute of Electrical and Electronics Engineers https://doi.org/10.1109/HICSS.2016.248

Zimmer M (2010). But the data is already public: On the ethics of research in Facebook. Ethics and Information Technology, 12, 313–325.

Zwitter A (2014). Big data ethics. Big Data & Society, 1(2), 1–6. https://doi.org/10.1177/2053951714559253

Table 1

Social media use by racial/ethnic category (from Smith and Anderson 2018)

Race	Use Internet (%)[a]	Facebook (%)	Twitter (%)	Instagram (%)	YouTube (%)	WhatsApp (%)
Non-Hispanic White	89	67	24	32	71	14
Non-Hispanic Black	87	70	26	43	76	21
Hispanic	88	73	20	38	78	49

[a]Internet penetration rates are from Pew Research Center (2018).

Table 2

Age and educational attainment within Facebook and LinkedIn: U.S. data only

Website	Age Interval	Total Population	Population With a College Degree	Degree Rate	Proportion of Users in Age Interval
LinkedIn	18–24	9,200,000	6,400,000	.696	.190
LinkedIn	25–34	16,000,000	9,600,000	.600	.330
LinkedIn	34–54	16,000,000	9,800,000	.613	.330
LinkedIn	55+	7,300,000	3,900,000	.534	.151
Facebook	18–24	39,000,000	8,600,000	.221	.171
Facebook	25–34	60,000,000	23,000,000	.383	.263
Facebook	34–54	81,000,000	28,000,000	.346	.355
Facebook	55+	48,000,000	16,000,000	.333	.211

Notes: Contribution of age compositional differences = –0.0032. Contribution of rate schedule differences = –0.2776. Proportion of total contribution attributable to difference in age composition = 0.0113. Proportion of total contribution attributable to differences in rate schedules = 0.9887.