Title: Promises and Pitfalls of Using Digital Traces for Demographic Research

Authors: Nina Cesare[1], Hedwig Lee[2], Tyler McCormick[1], Emma Spiro[1] and Emilio Zagheni[1]

[1] University of Washington
[2] Washington University in St. Louis

Abstract

The digital traces 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 – 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 review examples of current demographic literature that creatively utilizes 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 (Sagiroglu and Sinac 2013) – has grown significantly in the past several years. Many forms of "big data" are social data, and are valuable to those interested in examining behaviors, attitudes and macro-level social processes. This paper focuses on one type of socially relevant data, *digital traces*. We define digital traces as 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). This may include data from popular social networking sites such as Facebook or Twitter, personal blogs, collaborative online spaces such as Wikipedia, as well as 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, Bobrowicz, Schiano and Nardi 2013: 39).

While the scope of this 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 we provide focus on social media data. We do not address very large but systematically collected conventional data sets, though 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, D'Amuri and Marcucci 2013) and migration (e.g., Zagheni and Weber 2012). Further, it reflects an intellectual environment where 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.

The article is organized as follows. The first section provides background on features of digital traces that make them promising for population research. The second section discusses the technical, ethical and institutional challenges for research with digital traces. The third section reviews the emerging literature on digital demography and provides a snapshot of the state of the art. The fourth section presents 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 how these data can be leveraged as a "digital census". The last two sections offer a discussion of our article in 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, 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) states: "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 percent 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, Kim, Allum and Denman, 2016). At the root of this interest are the

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

advantages that these data offer. While 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 like 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, anonymized, digital trace data from their customers to researchers interested in analyzing these data (Blumenstock 2012; Blumenstock, Cadamuro and On 2015; Blumenstock and Eagle 2010).

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 having to observe the same group at discrete time points. The ability to represent time as “continuous, rather than bundled” (Ruppert, Law, and Savage 2013: 36) opens a wealth of new opportunities to researchers, such as examining real-time trends in daily activities (Goulder and Macy 2014), mobility (Williams, Thomas, Dunbar, Eagle, Dobra 2015), attitudes (O'Connor, Balasubramanyan, Routledge, and Smith 2010), health behaviors (Heaivilin, Gerbert, Page and Gibbs 2011), and migration (Zagheni and Weber 2012; Zagheni, Garimela, Weber and State, 2014). Researchers may also examine these behaviors pre-, during and post- crisis events, such as natural disasters (Sutton, Spiro, Johnson, Fitzhugh, Gibbons and Butts, 2013; Reeder, McCormick, Spiro 2014) or terrorist attacks (Starbird, Maddock, Orand, Achterman and Mason 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 documents the difficulty of capturing attitudes and opinions related to controversial topics due to social desirability bias (Belli, Traugott, Young and McGonagle 1999; Holbrook and Krosnick 2010; Tourangeau and Yan 2007), or selective recall (Fadnes, Taube, and Tylleskär 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 contain information bots, 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 (Barbera 2016; Lewis, Zamith and Herida 2013). Recent data from the Pew Research Center (see Table 1) indicate that while non-Hispanic black and Hispanic internet roughly parallels non-

Hispanic white internet use overall, non-Hispanic black and Hispanic are actually more active on some social media sites than white users (Smith and Anderson, 2018). These numbers show that not only are racial/ethnic minorities highly present on sites such as Instagram and Twitter, in some cases they are proportionally over-represented.

Table 1: Social Media Use by Racial/Ethnic Category (from The Pew Internet and American Life Project: Mobile Messaging and Social Media 2015)

Race	Use Internet (%)*	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	28	78	49

* Internet penetration rates from Pew Research Center: Internet/Broadband Fact Sheet (2018)

Challenges 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 and cautioning researchers about the unique challenges that they present (Adams and Brueckner 2015; boyd and Crawford 2012; Couldry and Powell 2014; González-Bailón 2013; Goulder and Macy 2014; Felt 2016; Kitchin 2014; Lazer Kennedy, King and Vespigani 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. The following paragraphs 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, since digital 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, as 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 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 type 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 occurs 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 refer to the additional uncertainty that arises from this decision-making process as the “garden of forking paths.” The challenges we describe in this section remain even with solid theoretical grounding and should be addressed by the researcher.

Another limitation of using digital trace data is that it is not typically representative of populations to which demographic researchers often seek to generalize. While some platforms – such as cell phone technology – may be more pervasive than others, not everyone uses - and is therefore captured by - technologies that archive digital trace data (Graham, Hale and Stevens 2012). A keyword-based sample drawn from Facebook or Twitter would only contain individuals who have regular access to the internet and who 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 is neither a new problem nor is it unique to digital traces. Bias may arise, for instance, when utilizing standard survey procedures - such as phone-based sampling, which is known to only represent non-institutionalized populations (Pettit 2012). Moreover, as the percentage of households who have a landline number in the US is decreasing, traditional sampling methods that rely on phone-based interviews may become increasingly challenging. Due to 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 post-stratification techniques that allow them to draw inference about populations using non-representative data. Recent research related to surveying non-representative Xbox users about their intentions to vote offers promising results (Wang et al. 2015). Overall, demographers have developed approaches to correct for bias when ground truth is known in traditional, but imperfect data (Alkema et al. 2012). While 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.

There are also research contexts where the “non-representativeness” 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 they are drivers of change. For instance, we may consider whether patterns of same-race connectedness within friendship networks seen offline perpetuate regardless of social context using data drawn from social media sites (Lee, Cesare, McCormick, Morris and Shojaie 2014). Analogously, we can study the impact 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, have the potential to introduce bias into digital trace data. For instance, those sampling directly from a particular website using the website’s

API do not necessarily receive a comprehensive set of the content they seek. Querying Twitter's streaming API, for example, provides 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 what data are included and what data are excluded. A common query strategy for those studying social processes using Twitter is to sample data based on hashtags (meta-data added by users 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 that lie outside its focus (Tufecki 2014). Furthermore, the resulting data may not represent the entirety of relevant content or stakeholders. It is important that researchers 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 that it is difficult to engage 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, Spiro and Starbird 2016). While convenient, these methods do raise concerns. Using automated text analysis techniques on large, textual data provides only a partial understanding of the meaning embedded within this content. While analyzing word counts or using machine learning techniques 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 (Tufecki 2014).

Barring the availability of a large team of coders, very large datasets of digital traces preclude a thorough qualitative analysis of their full content. However, this does not mean that researchers should dismiss qualitative analysis as a technique for understanding their data. While it may be tantalizing to utilize 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 (Tufekci, 2014; Andrews, Fichet, Ding, Spiro and Starbird 2016). 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. Just because digital trace

[2] See https://dev.twitter.com/streaming/overview for an overview of Twitter's public streaming APIs.

data are easily accessible, does not mean their use is always ethical (boyd and Crawford 2012). Indeed, most social media users value personal privacy online (Madden and Rainie 2015) and usually do not suspect that their information will be used for research purposes (Vitak 2015). While 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, extends 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, Floridi, and van der Sloot 2017).

Ethical management of digital trace data is complicated by a varying landscape of data ownership. For example, Facebook's data policy stakes claim on the data produced by those using their platform,[3] but Twitter upholds that users own the data they produce.[4] The premise of withholding data may be to protect the privacy of the individuals within it, but such non-disclosure also inhibits exploration from the scientific community. Project OPAL[3] (for 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 its size and nature, may leave users vulnerable to breaches of privacy but, due to its content, could serve to benefit researchers and policymakers. It is likely that the suite of tools developed by OPAL would be replicated by other organizations in the future.

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. While 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, recommendation 2.1, p. 40), it is still the case that 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 – particularly, researchers who 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 datasets can present formidable

[3] See http://www.opalproject.org/about-us/ for more information

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. While 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, and to favor learning of modern data science techniques for reproducible research. One example of this effort is the IUSSP scientific panel "Big Data and Population Processes," which currently offers training workshops at population and social media/informatics conferences.[4]

Overall, it is critical that social and demographic researchers engage in dialogue regarding the proper use and application of digital trace data. A survey of over 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 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 University of Washington's eScience Institute 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

There is growing interest in the use of digital trace data among demographers, as evidenced by multiple sessions in recent Population Association of America (PAA) meetings (Blumenstock and Toomet, 2014; Cesare, Spiro and Lee, 2015; Massey 2016; Mateos and Durand, 2014; Reeder et al. 2015; Rosello and Filgueira, 2016; Williams, Thomas, Dunbar, Eagle and Dobra, 2014; Kashyap, Billari, Cavalli, Quian, and Weber 2017; Zagheni, Weber and Gummati, 2017), recent publications (Blumenstock 2012; Blumenstock and Eagle 2012; Malik and Pfeffer, 2016; Mendieta et al. 2016; Palmer et al, 2013; Stevenson, 2014; Willekens, Massey, Raymer, and Beauchemin 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 (2006)'s analysis of interracial friendship using wedding photos posted online and Palmer et al. (2013)'s work on spatial mobility with data collected via a smartphone app. While work such as Barry (2006) stands out in

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

its innovation and novelty, little research has directly built upon this contribution. Overall, the technical capabilities to use digital traces for population studies by 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, like 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.

Researchers have begun to use digital trace data to examine topics traditionally discussed within the context of demography – such as migration, mortality and fertility – in new ways. 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, compare the volume of abortion-related searches in a particular area and the number of restrictions imposed upon abortions in the area. They find an inverse relationship between these measures, which leads to the conclusion 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, D’Amuri, and Marcucci (2013) find 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, Zagheni, Billari and Weber (2017) illustrate that combining data from Google Correlate/Google Trends and the American Community Survey allows researchers 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 fertility and reproductive health (De Chouhury, Counts, Horvitz and Hoff, 2014).

Some studies of mortality have leveraged digital traces. Tomlinson and colleagues (2009) state 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, Faye and Lishou (2013) show 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. There is potential to explore details related to health and mortality via other sources, such as search queries or social media data such as Twitter, Tumblr or Facebook (Eichstaedt et al. 2015), or other forms of archived digital data (Tourassi, Yoon and Xu, 2016).

Digital traces have been used extensively to examine migration as well. Blumenstock (2012) uses 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 and colleagues (2014) propose methods of calibrating cell phone data that produce information on intra- and inter-national mobility patterns that is as detailed or more detailed than traditionally collected survey data. Taking a different approach, Palmer and colleagues (2013) use cell phone surveys to study micro-interactions and examine social processes within activity spaces rather than residential census units. While cell phone data are valuable for studies of migration, other geo-tagged digital traces may also be used to examine human mobility. These include geo-referenced Yahoo! email data, which have been used to estimate profiles of international migration by age and sex (Zagheni and Weber 2012); geo-located 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, Rodriguez, Helbing and Zagheni 2014); and the network of Skype calls to track international migrations (Kikas, Dumas and Saabas 2015).

Some researchers advocate that combining digital trace data with systematically collected survey data can add much-needed dimensionality to data-rich but variable-poor digital traces. Snijders, Matzat and Reips (2012), for example, encourage 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 and colleagues (2014) argue that if Google flu trends data were combined with existing CDC data, researchers would be able to better calibrate their search to make more accurate predictions in health trends. Scholars have also combined digital data with other data sources to predict depression (De Choudhury, Gamon, Counts, Horvitz 2013) and food insecurity (De Choudhury, Sharma, and Kiciman 2016). 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 linked with social media data in one study can then be used to validate coding methods when no surveys are available. For example, Cesare and colleagues (2015) use survey data with respondent self-reported demographic information linked to Twitter data to examine trends in self-presentation. Similarly, Moreno and colleagues (2012) correlate 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.

Researchers can aggregate and de-aggregate information embedded within digital traces in unique and interesting ways. Profile photos contained within big datasets, 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 crowdsourced human intelligence can be used to accurately and reliably extract valuable information from these photos (McCormick et al. 2015). Similarly, Zagheni and colleagues (2014) use facial-recognition software to add demographic dimensionality 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 US Census to estimate users' demographic characteristics (e.g., Mislove, Lehmann and Ahn, 2011).

A number of scholars are currently developing methods to account for the bias created by the use of non-representative samples when baseline population data are both known and unknown. In regard to the former, Zagheni and Weber (2012) use email data to measure rates of international migration; in order work with these non-representative data they develop 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 non-representative polls, Wang and colleagues (2014) use multilevel regression and post-stratification based on respondents' demographic characteristics to predict election outcomes using Xbox as their survey tool. Their predictions are extremely similar to results from nationally representative data, both nationwide and state-by-state, thus illustrating that data drawn from "samples of convenience" – such as those drawn from digital traces – 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 traces sources, however, is an area in which there is significant room for methodological innovations from social scientists and statistical demographers.

Beyond the opportunity for methodological contributions, researchers can directly address the limitations of their data by appropriately bounding the conclusions they draw from them. When analyzing data from a particular social media site, for instance, a researcher should specify that their conclusions may be context-dependent and be in some way connected to the characteristics of the platform used. However, it is important to note that 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 (Shaw and McKay1942; Park and Burgess 1925). These cities are not representative of the entire US, but the authors make clear 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 adverts 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. We note that while some forms of digital data – such as Facebook adverts 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 Adverts Manager[5] 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, currently, provided free of charge, and can be accessed in a programmatic way via the Facebook API.[6] It is useful for advertisers when they plan

[5] http://www.facebook.com/business/

[6] https://developers.facebook.com/docs/marketing-api/audiences-api

their ad campaigns and need to develop an appropriate budget, or when they must decide whether to narrow or broaden their 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 two billion Facebook users, and freely obtain aggregate-level measures of demographic characteristics as well as topical interests.

Facebook's Advertising Manager provides population estimates from large, non-representative 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 from Facebook Adverts Manager to estimate stocks of migrants in the US, and to understand biases in the population of Facebook users. For each combination of age, sex, country of origin and US state of destination, they examined the difference between the fraction of foreign-born individuals estimated by the American Community Survey 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 where the bias is regressed against a series of indicator variables for different demographic groups, countries of origin and US states of destination, Zagheni et al. (2017) found that there are important regularities in profiles of migrants by age and sex across US states of destination or countries of origin. These regularities were then leveraged to improve predictions. The approach relies on combining traditional data sources (e.g., the American Community Survey) and new emerging ones (e.g., Facebook data for advertisers) to generate estimates that are timely and geographically granular in developed countries. In the context of developing countries, existing 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 non-social media contexts. For example, Alkema et al. (2012) used a regression model with indicator variables for data quality in order 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 "re-purposed" 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 Adverts 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. In their study they sampled over 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 survey using non-representative samples that requires post-stratification techniques in order to weigh 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 for instance, Wang et al. 2015). Moreover, it is a challenge for which there is precedent, as it is related to issues that traditional phone-based surveys face nowadays: decreasing response rates that are non-random, and increasing numbers of households that do not have landlines and are excluded from the samples.

A tool like Facebook Adverts 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 bring methodological advances.

For a concrete example of how traditional demographic methods can be used to understand populations of social media users, consider the following illustrative case. A researcher may 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? And how much is it because LinkedIn users are really more educated? If each site is considered a ‘population,’ then a simple age decomposition analysis provides the answer.

To illustrate, we used Facebook's Advertisements Manager and LinkedIn's Campaign Manager[7] 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. campaigns designed to increase store visits, video views or clicks). Given that we wished to select as broad an audience as possible within our criteria, we selected a "reach" campaign. For Facebook, we requested population estimates for users located within the U.S. within specific age ranges, as well as subsets of these groups that Facebook identifies as "college grads." 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. B.A., B.S.,

[7] See: https://www.linkedin.com/ad/accounts

B.F.A, etc.). 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 use 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, Heuveline, and Guillot (2001)[8]. 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%).

Table 2: Age and Educational Attainment within Facebook and LinkedIn

Website	Location	Age interval	Total pop.	Pop. W/college degree	Degree rate	Prop. of users in age interval
LinkedIn	U.S.	18 to 24	9200000	6400000	0.696	0.190
LinkedIn	U.S	25 to 34	16000000	9600000	0.600	0.330
LinkedIn	U.S.	34 to 54	16000000	9800000	0.613	0.330
LinkedIn	U. S.	55+	7300000	3900000	0.534	0.151
Facebook	U. S.	18 to 24	39000000	8600000	0.221	0.171
Facebook	U. S.	25 to 34	60000000	23000000	0.383	0.263
Facebook	U.S.	34 to 54	81000000	28000000	0.346	0.355
Facebook	U.S.	55+	48000000	16000000	0.333	0.211
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						

Although most of the 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 normalized.

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 17th century by leveraging data that were collected for marketing purposes to estimate the age distribution of potential customers in the city of London. Today, new types of “re-purposed” digital traces provide opportunities for advances in the field. In this article, we reviewed the state-of-the-art uses of digital trace data in demography – the practices of a subfield sometimes referred to as “digital demography” - and highlight what we believe

[8] See pages 28-30 for details on conducting an age decomposition analysis.

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 about two thirds of all annual deaths and almost 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 like fertility, mortality and migration. These hopes are sustained by the observation that certain types of technology, like 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 "big 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. There is incredible potential in the use of digital trace data for social science and demographic research, and many points raised against these data sources are not insurmountable challenges but opportunities to methodologically advance these fields.

We believe that graduate and postdoctoral training and interdisciplinary collaboration is key to increasing the accessibility of "big data" to demographic researchers. As suggested by González-Bailón (2013), "…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" (158). 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 think of creative and effective ways in which digital trace data may 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 what sort of social factors impact 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, possibly even second-by-second basis. Indeed, the real-time generation of digital traces has been utilized already to examine time-sensitive trends such as how moods change over time (Goulder and Macy, 2011) or reactions to crisis events (Andrews, Fichet, Ding, Spiro and Starbird, 2016; Starbird, Spiro, Edwards, Zhou, Maddock, and Narasimhan,

2016) and there are opportunities to utilize 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 – a *discovery* stage, which uncovers unique patterns within population data, and 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 filled with biases reflective of the sources generating them. Managing these unique traits invites a dialogue 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 parallels, to a large extent, 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 (Beli and Smyth, 2017)

It is clear that demography as a field 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. 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 objects that are of central interest to disciplines like 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. This would be a straightforward application of demographic tools to estimate population growth from one Census (e.g., Keyfitz and Caswell 2005). Feehan and Cobb (2017) illustrate 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 impact 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, etc. 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 between demographers, computer scientists and researchers broadly involved in the area of social informatics.

This paper is a response to the growing use of digital traces, particularly social media data and cell phone data, in demography - as well as work that critiques these applications. It is intended to interpret how the field might evolve to accommodate the use of these data. We argue that demographers are well positioned to address two main challenges presented by digital trace data and to seize important methodological opportunities that these challenges open. First, in a data-driven world, demographers possess the skills needed to develop methods to extract useful information from large, but often noisy, messy and non-representative data. Second, demographers have the opportunity to use 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 reluctant to contribute to the study of digital trace data - could become primary innovators in this area.

References

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

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., & Pelletier, F. (2012). Estimating trends in the total fertility rate with uncertainty using imperfect data: Examples from West Africa. *Demographic research*, *26*(15), 332-361.

Andrews, C., Fichet, E., Ding, Y., Spiro, E. S., & Starbird, K. (2016). Keeping Up with the Tweet-dashians: The Impact of “Official” Accounts on Online Rumoring. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. ACM*, 452–465. http://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. Retrieved from http://dl.acm.org/citation.cfm?id=2427085

Araújo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using Facebook Ads Audiences for Global Lifestyle Disease Surveillance: Promises and Limitations. *Proceedings of Web Science 2017,* Troy, NY, 253-257.

Barberá, P. (2016). Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. *Working Paper for NYU*. Retrieved from: http://pablobarbera.com/static/less-is-more.pdf

Barry, Brent. (2006). Friends for Better or For Worse: Interracial friendships in the United States as Seen through Wedding Photos. *Demography, 43*(3), 491-510.

Belli, R. F., Traugott, M. W., Young, M., & McGonagle, K. A. (1999). Reducing Vote Overreperting in Surveys: Social Desirability, Memory Failure, and Source Monitoring. *The Public Opinion Quarterly*, *63*(1), 90–108. Retrieved from http://www.jstor.org/stable/10.2307/2991270

Berinsky, A. J. (1999). The Two Faces of Public Opinion. *American Journal of Political Science*, *43*(4), 1209–1230. http://doi.org/10.2307/2991824

Blei D. M. and Smyth P. 2017. Science and Data Science. *Science 114*(3): 8689-8692

Billari, F., D’Amuri, F., & Marcucci, J. (2013). Forecasting births using google. *Annual Meeting of the Population Association of America*. New Orleans, LA.

Billari, F. C., & Zagheni, E. (2017). Big Data and Population Processes: A Revolution? In Alessandra Petrucci, Rosanna Verde (edited by), *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, Firenze University Press, 2017, pp. 167–178, CC BY 4.0.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. http://doi.org/10.1126/science.aac4420

Blumenstock, J. E., & Eagle, N. (2012). Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technologies & International Development*, *8*(2), 1–16.

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*, London, UK. 6-10. http://doi.org/10.1145/2369220.2369225

Blumenstock, J. and Toomet. O. (2014). Segregation and 'Silent Separation': Using Large-Scale Network Data to Model the Determinants of Ethnic Segregation. *Annual Meeting of the Population Association of America*, Boston, MA.

boyd, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. http://doi.org/10.1080/1369118X.2012.678878

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.

Carberry, J. (2014). *Media Statement on Cornell University's Role in Facebook "Emotional Congation" Research*. *Cornell University Media Relations Office*. Retrieved from http://mediarelations.cornell.edu/2014/06/30/media-statement-on-cornell-universitys-role-in-facebook-emotional-contagion-research/

Cesare, N., Spiro, E., & Lee, H. (2015). Self-Presentation and Information Disclosure on Twitter: Understanding Patterns and Mechanisms Along Demographic Lines. *Annual Meeting of the Population Association of America*. San Diego, CA.

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

De Choudhury, M. S. Sharma, E, Kiciman. (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. San Francisco, CA, 1157-1170. http://doi.org/10.1145/2818048.2819956

De Choudhury, M., & Gamon, M. (2013). Predicting Depression via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA, 128–137. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351

De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013. “Predicting Postpartum Changes in Emotion and Behavior via Social Media.” Pp. 3267–3276 in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 3267-3276.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, a. E., … Tatem, a. J. (2014). Dynamic Population Mapping using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, *111*(45). http://doi.org/10.1073/pnas.1408439111

Duggan, M. (2015). Mobile Messaging and Social Media 2015. *Pew Research Center: Internet and American Life Project*. Retrieved from http://www.pewinternet.org/files/2015/08/Social-Media-Update-2015-FINAL2.pdfhttp://www.pewinternet.org/files/2015/08/Social-Media-Update-2015-FINAL2.pdf

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., J, Agrawal, M., Dziurzynski, L.A., Sap, M. and Weeg (2015). “Psychological Language on Twitter Predicts County-Level Heart Disease Mortality.” *Psychological Science 26*(2): 159-169.

Fadnes, L., & Taube, A. (2009). How to identify information bias due to self-reporting in epidemiological research. *The Internet Journal of Epidemiology*, *7*(2), 1–8. http://doi.org/10.5580/1818

Feehan, D. and Cobb, C. 2017. How Many People Have Access to the Internet?: Estimating internet adoption around the world using Facebook. Presented at *IC2S2*. Oxford, United Kingdom.

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

Gelman, Andrew and Eric Loken. (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. *Department of Statistics, Columbia University*. Retrieved September 4, 2017 (http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf).

Golder, S. A., & Macy, M. W. (2014). Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*, *40*(May), 129–52.

http://doi.org/10.1146/annurev-soc-071913-043145

González-Bailón, S. (2013). Social Science in the Era of Big Data. *Policy and the Internet*, *5*(147), 160. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/1944-2866.POI328/full

Graham, M., Hale, S., & Stephens, M. (2012). Featured graphic: Digital divide: the geography of Internet access. *Environment and Planning*, *44*(5), 1009–1010. http://doi.org/10.1068/a44497

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 1–31. http://doi.org/10.1093/pan/mps028

Heaivilin, N., Gerbert, B., Page, J. E., & Gibbs, J. L. (2011). Public health surveillance of dental pain via Twitter. *Journal of Dental Research, 90*(9), 1047–1051. http://doi.org/10.1177/0022034511415273

Hogan, B. (2010). The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386. http://doi.org/10.1177/0270467610385893

Holbrook, A. L., & Krosnick, J. A. (2010). Social Desirability Bias in Voter Turnout Reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67. http://doi.org/10.1093/poq/nfp065

Karpf, D. (2012). Social Science Research Methods in Internet Time. *Information, Communication & Society*, *15*(5), 639–661. http://doi.org/10.1080/1369118X.2012.665468

Kashyap, R. Billari, F. C., Cavalli, N., Quian, E. and Weber, I. 2017. "Ultrasound Technology and "Missing Women" in India: Analyses and Now-casts Based on Google Searches. *Annual Meeting of the Population Association of America,* Chicago, IL.

Keyfitz, N., & Caswell, H. (2005). C3 - The Matrix Model Framework. In Nathan Keyfitz and Hal Caswell (Ed.), *Applied Mathematical Demography, Third Edition* New York: Springer, 47-70.

Kikas, R., Dumas, M., & Saabas, A. (2015). Explaining International Migration in the Skype Network. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors - SIdEWayS '15*, Cyprus, Greece, 17-22. http://doi.org/10.1145/2806655.2806658

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

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 10779–10779. http://doi.org/10.1073/pnas.1412469111

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

Lazer, David and Jason Radford. 2017. "Data ex Machina: Introduction to Big Data." *Annual Review of Sociology 49*: 19-39.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–5. http://doi.org/10.1126/science.1248506

Lee, H., Cesare, N., McCormick, T. H., Morris, J., & Shojaie, A. (2014). Redrawing the “Color Line”: Examining Racial Homophily of Associative Networks in Social Media. *Annual Meeting of the Population Association of America.* Boston, MA.

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

Lewis, S. C., 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), 1–33. http://doi.org/10.1080/08838151.2012.76170

Lohr, S. (2012, February 11). The Age of Big Data. *New York Times*. http://doi.org/10.1126/science.1243089

Madden, M., & Rainie, L. (2015). Americans’ attitudes about privacy, security and surveillance. *Pew Research Center*, 47. Retrieved from http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

Malik, M. M. (2016). Social media data and computational models of mobility: A review for demography. In *Proceedings of 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 M. K. Gold (Ed.), *Debates in the Digital Humanities*, Minneapolis: University of Minnesota Press, 460-476.

Marwick, A., & boyd, d. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. http://doi.org/10.1177/1461444810365313

Massey, D. (2016). Measuring Racial Prejudice Using Google Trends. *Annual Meeting of the Population Association of America.* Washington D.C.

Mateos, P., & Durand J. (2014). Netography and Demography: Mining Internet Forums on Migration and Citizenship. *Annual Meeting of the Population Association of America.* Boston, MA.

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods & Research*. http://doi.org/10.1177/0049124115605339

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)*, Quito, Equador, 126-132.

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). Who is doing computational social science? Trends in big data research (White paper). London, UK: SAGE Publishing. doi: 10.4135/wp160926. Retrieved from https://us.sagepub.com/sites/default/ files/CompSocSci.pdf

Mislove, A., Lehmann, S., & Ahn, Y. (2011). Understanding the Demographics of Twitter Users. In *Proceedings of the 5th Internatoinal Conference on Weblogs and Social Media (ICWSM 11)*. Barcelona, Spain. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Mohr, J., & Boganov, P. (2013). Topic models: What they are and why they matter. *Poetics*, *41*(6), 545–569.

Moreno, M. A., Christakis, D. A., Egan, K. G., Brockman, L. N., & Becker, T. (2012). Associations between displayed alcohol references on Facebook and problem drinking among college students. *Archives of Pediatrics*, *166*(2), 157–163. http://doi.org/10.1001/archpediatrics.2011.180.Associations

National Research Council (NRC). (2014). Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences. Washington, DC: Natl. Acad. Press

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. a. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media,* Washington, DC, *122*–129. Retrieved from: http://doi.org/citeulike-article-id:7044833

Ojala, J, Zagheni E., Billari, F., and Weber I. (2017). Fertility and Its Meaning: Evidence from Search Behavior. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. Montreal, Quebec, Canada.

Palmer, J. R. B., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2013). New approaches to human mobility: using mobile phones for demographic research. *Demography*, *50*(3), 1105–28. http://doi.org/10.1007/s13524-012-0175-z

Park, R. E., & Burgess, E. W. (1925). *The City*. University of Chicago Press.

Pettit, B. (2012). *Invisible Men: Mass Incarceration and the Myth of Black Progress*. Russel Sage.

Pötzschke, S., & Braun, M. (2016). Migrant sampling using Facebook advertisements: a case study of Polish migrants in four European countries. *Social Science Computer Review*, 0894439316666262.

Preston, S.H., P. Heuveline, and M. Guillot. 2001. *Demography: Measuring and Modeling Population Processes*. Blackwell.

Pew Research Center. (2018). Internet/Broadband Fact Sheet. Retrieved from: http://www.pewinternet.org/fact-sheet/internet-broadband/

Reeder, H., McCormick, T. H., & Spiro, E. (2014). Online Information Behaviors During Disaster Events: Roles, Routines, and Reactions. *Center for Satistics and the Social Sciences Working Paper*, (144).

Reis, B. Y., & Brownstein, J. S. (2010). Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC Public Health*, *10*(1), 514. http://doi.org/10.1186/1471-2458-10-514

Rosello, J. L. D., & Filgueira, F. (2016). Big Data in a Small Country: Integrating Birth, Maternal and Child Statistics in Uruguay. *Annual Meeting of the Population Association of America*. Washington D.C.

Rosenfeld, M. J., & Thomas, Reuben J. (2012). Searching for a Mate : The Rise of the Internet as a Social Intermediary. *American Sociological*, *77*(4), 523–547.

Ruggles, S. (2014). Big Microdata for Population Research. *Demography*, *51*(1), 287–297. http://doi.org/10.1007/s13524-013-0240-2

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling Social Science Methods: The Challenge of Digital Devices. *Theory, Culture & Society*, *30*(4), 22–46.

http://doi.org/10.1177/0263276413484941

Sagiroglu, S., & Sinanc, D. (2013). Big Data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. http://doi.org/10.1109/CTS.2013.6567202

Ševčíková, H., Raftery, A. E., & Waddell, P. A. (2007). Assessing uncertainty in urban simulations using Bayesian melding. *Transportation Research Part B: Methodological*, *41*(6), 652-669.

Shaw, C. R., & McKay, H. D. (1942). *Juvenile Delinquency and Urban Areas.* University of Chicago Press.

Smith, A. and Monica A. (2018). "Social Media Use in 2018." *Pew Research Center: Internet and American Life Project.* 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. (2012). "Big Data": Big Gaps of Knowledge in the Field of Internet Science. *International Journal of Internet Science*, *7*(1), 1–5. http://doi.org/10.3923/ijds.2012.1.10

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. In *iConference 2014 Proceedings*, Berlin, Germany, 654–662.

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*, San Jose, CA, 360-371.

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of professionals to the U.S.: Evidence from linkedin data. In *6th International Conference on Social Informatics (SocInfo) 2014*, Barcelona, Spain, 531–543

Stevenson, A. J. (2014). Finding the Twitter users who stood with Wendy. *Contraception*, *90*(5), 502–507. http://doi.org/10.1016/j.contraception.2014.07.007

Sutton, J., Spiro, E. S., Johnson, B., Fitzhugh, S., Gibson, B., & Butts, C. T. (2014). Warning tweets: serial transmission of messages during the warning phase of a disaster event. *Information, Communication & Society*, *17*(6), 765–787. DOI: http://doi.org/10.1080/1369118X.2013.862561

Tamgno, J. K., Faye, R. M., & Lishou, C. (2013). Verbal autopsies, mobile data collection for monitoring and warning causes of deaths. In *International Conference* *on Advanced Communication Technology, ICACT*. PyeongChang, Republic of Korea, 495–501. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84876241864&partnerID=tZOtx3y1

Taylor, L. Floridi, L, van der Sloot, L. 2017. Group Privacy: New Challenges of Data Technologies. *Springer*.

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property Volume*, *11*(5), 240–273. Retrieved from http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20

Tourangeau, R., & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin*, *133*(5), 859–83. http://doi.org/10.1037/0033-2909.133.5.859

Tourassi G., Yoon, H. J., and Xu S. (2016). A Novel Web Informatics Approach for Automated Surveillance of Cancer Mortality Trends. *Journal of Biomedical Informatics 61*, 110-118.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (pp. 505–514). Oxford, UK. Retrieved from http://arxiv.org/abs/1403.7400

Vitak, J. (2015). I like it….whatever that means: The evolving relationship between disclosure, audience, and privacy in networked spaces. Lecture, University of Washington, Seattle, WA.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2014). Forecasting elections with non-representative polls. *International Journal of Forecasting*. http://doi.org/10.1016/j.ijforecast.2014.06.001

Willekens, F., Massey, D., Raymer, J., Beauchemin, C. (2016). International migration under the microscope. *Science*, *352*(6288), 897–9. http://doi.org/10.1126/science.aaf6545

Williams, N. E., Thomas, T. A., 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. http://doi.org/10.1371/journal.pone.0133630

Zagheni, E., Garimella, V. R. K., 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*, Seoul, Republic of Korea: ACM Press. DOI: http://doi.org/10.1145/2567948.2576930

Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, *36*(1), 13–25. http://doi.org/10.1108/IJM-12-2014-0261

Zagheni, E., & Weber, I. (2012). You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 3rd Annual ACM Web Science Conference*. Evanston, IL. DOI http://doi.org/10.1145/2380718.2380764

Zagheni, E., Weber, I. and Gummadi, K. 2017. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Annual Meeting of the Population Association of America,* Chicago, IL. (forthcoming in Population and Development Review).

Zeng, L., Starbird, K., & Spiro, E. S. (2016). Rumors at the Speed of Light? Modeling the Rate of Rumor Transmission During Crisis. *49th Hawaii International Conference on System Sciences (HICSS)*, 1969–1978. DOI: http://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*(4), 313–325. http://doi.org/10.1007/s10676-010-9227-5

Zwitter, A. (2014). Big Data ethics. *Big Data & Society, 1*(2), 1-6, http://doi.org/10.1177/2053951714559253