Bias and beyond in digital trace data

Momin M. Malik

CMU-ISR-18-105
August 2018

Institute for Software Research
School of Computer Science
5000 Forbes Avenue
Pittsburgh PA 15213

Thesis Committee:

Jürgen Pfeffer *(co-chair)*	Institute for Software Research
Anind K. Dey *(co-chair)*	Human-Computer Interaction Institute
Cosma Rohilla Shalizi	Department of Statistics & Data Science
David Lazer	Northeastern University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Societal Computing.

This research was sponsored by the the UPMC Center for Machine Learning and Health under award number 0055855, by the Charles Stark Draper Laboratories under award number SC001-0000000995, by the National Science Foundation under award numbers IIS-1404698 and 00361150115291, by the National Security Agency and Department of Defense under award number H9823014C0140, by the US Army Research Office under award number W911NF1310154, by the Office of Naval Research under award number N000141310835, by the US Army Contracting Agency Aberdeen Test Center under award number W911NF1410481, and by the ARCS Foundation. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Suggested citations

ACM Momin M. Malik. 2018. Bias and Beyond in Digital Trace Data. Ph.D. Dissertation. Carnegie Mellon University, Pittsburgh, PA. Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

APA Malik, M. M. (2018). *Bias and beyond in digital trace data* (Doctoral dissertation, Carnegie Mellon University). Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

Chicago Malik, Momin M. “Bias and Beyond in Digital Trace Data.” PhD dissertation, Carnegie Mellon University, 2018. http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

IEEE M. M. Malik, “Bias and beyond in digital trace data,” Ph.D. diss., Inst. Softw. Res., Sch. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, 2018. Available: http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

MLA Malik, Momin M. *Bias and Beyond in Digital Trace Data*. 2018. Carnegie Mellon U, PhD dissertation. SCS Technical Report Collection, http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html.

BibTeX

```
@phdthesis{malik2018,
    author = {Malik, Momin M.},
    title = {Bias and beyond in digital trace data},
    year = {2018},
    school = {Carnegie Mellon University},
    address = {Pittsburgh, PA},
    month = {08},
    url = {http://reports-archive.adm.cs.cmu.edu/anon/isr2018/abstracts/18-105.html}
}
```

Keywords
Computational social science; bias; generalizability; validity; digital trace data; measurement; machine learning; social media; social network analysis; mobile phone sensors; STS; data science; critical technical practice; critical social science; critical algorithm studies; critical data studies

Abstract

Momin M. Malik

Bias and beyond in digital trace data

Large-scale digital trace data from sources such as social media platforms, emails, purchase records, browsing behavior, and sensors in mobile phones are increasingly used for business decision-making, scientific research, and even public policy. However, these data do not give an unbiased picture of underlying phenomena. In this thesis, I demonstrate some of the ways in which large-scale digital trace data, despite its richness, has biases in who is represented, what sorts of actions are represented, and what sorts of behaviors are captured. I present three critiques, demonstrating respectively that geotagged tweets exhibit heavy geographic and demographic biases, that social media platforms' attempts to guide user behavior are successful and have implications for the behavior we think we observe, and that sensors built into mobile phones like Bluetooth and WiFi measure proximity and co-location but not necessarily interaction as has been claimed.

In response to these biases, I suggest shifting the scope of research done with digital trace data away from attempts at large-sample statistical generalizability and towards studies that situate knowledge in the contexts in which the data were collected. Specifically, I present two studies demonstrating alternatives to complement each of the critiques. In the first, I work with public health researchers to use Twitter as a means of public outreach and intervention. In the second, I design a study using mobile phone sensors in which I use sensor data and survey data to respectively measure proximity and sociometric choice, and model the relationship between the two.

Contents

Introduction

Summary. In the introduction, I lay out all my thoughts about the way I see data and modeling being framed. I go into detail about all the facets of modeling and of data that I see as relevant to rigorously treating digital trace data. I have a critique of the use of the term 'algorithmic' for discussing societal impacts of uses of data and modeling, as well as extensive discussion of differences between predictive and explanatory modeling and of fundamental limitations in modeling, and a systematic review of the ways in which social media and sensor data may not generalize.

Motivation

Unquestionably, large-scale digital trace data about human behavior have found widespread use. The primary use has been in business and the consumer and service tech industry (Savage and Burrows, 2007), but these data have begun to be used in research (Lazer and Radford, 2017) and public policy (Veale et al., 2018) in ways that are far more consequential for scientific understandings and human well-being. For example, social media data are being used to assign credit scores, and scientific studies have begun using social media and sensors to assess participants' physical and mental health. In many cases, the use of such data will have enormous benefits, but many writers have argued that if we use these data without recognizing their biases, we risk making inaccurate decisions, getting unreliable scientific findings, or even creating unjust public policy (Gayo-Avello, 2011; boyd and Crawford, 2012; Mahrt and Scharkow, 2013; Lazer, Kennedy, et al., 2014; Tufekci, 2014; Ruths and Pfeffer, 2014; Hargittai, 2015; Shah et al., 2015; O'Neil, 2016; Eubanks, 2018; Baeza-Yates, 2018; Hargittai, 2018). There is work that looks at the potential *consequences* of such bias; but what these biases are, what effect they may have on data usage, and how we might react to them has only begun to be studied. Thus, this thesis is motivated by the central question: when and how might results derived from digital trace data of human behavior fail?

By **digital trace data**, I mean the digitally collected records of activity, such as emails, browsing behavior, metadata from phone calls or other mobile phone usage, credit card purchase records, posts and activity logs from online social media platforms, and logs of sensor data from mobile phones (Lazer, Pentland, et al., 2009). By 'results', I mean both claims about *understandings* of human behavior, and *predictions* about future behavior made from statistical analysis and/or machine learning—two tasks that are surprisingly distinct (Shmueli, 2010; Breiman, 2001), which I discuss further below. And by 'fail', I mean that either predictions perform far worse on true out-of-sample data than on test data, and/or claims about associations and/or causal processes made from one set of data do not generalize to systems or time periods captured in other data (Tufekci, 2014; Lazer and Radford, 2017).

In some cases, the failure or success of results has nothing to do with the source of data, but rather with the nature of modeling. Modeling human and social phenomena turns out to be quite different from modeling physical phenomena (Gelman and Shalizi, 2012), which I discuss below. Far more than with physical phenomena, models applied to social phenomena need to be interpreted and applied provisionally (Box, 1979; Cox, 1990; Kass, 2011; Buja et al., 2016).

In other cases, possibly even the majority of cases, digital trace data will *not* fail; they will lead to good performance on out-of-sample data, and findings about associations will generalize.

I also do not mean to say that other (more traditional) types of data collection are superior. But to understand how we should treat digital trace data, it is instructive to compare it to another type of data, namely, survey data, which has traditionally been the most prominent example of quantitative social science data.

Survey research has many drawbacks. First, there are many forms of bias present in survey research such as recall bias, social desirability bias, and careless recall, which all add non-random error to responses. Second, there is the difficulty of getting a truly representative sample for the responses; *reachability* by one medium or another turns out to have an enormous impact on seemingly unrelated phenomena (Arceneaux et al., 2010) even before accounting for the self-selection bias in those who choose to respond to surveys at all, among many other challenges (Dillman et al., 2014). Third, survey data is chronically underpowered: while only about 1,100 people are required to make inferences about a population of any size to within a margin of error of 3 percentage points (or, to have a test powerful enough to detect something present among at least 3% of the population) at the .05 significance level, which is why this is a usual number for survey research (Salant and Dillman, 1994), the problem is that *subpopulations* in a sample of 1,100 may be too small to make inferences about. For example, a representatively sampled phone survey by Pew (Zickuhr, 2013) with $n = 2,252$ only captured 141 people who used 'geosocial services', and only 1% or $n = 1$ person in the sample used Twitter's geosocial service (geotagged tweets)! This is fine for estimating the percentage of the US population that uses geotagged tweets, but it is nowhere near enough to make statistical inferences about the demographics of geotag tweet users, which was my topic of interest in Malik, Lamba, et al. (2015). As Lazer and Radford (2017) write, "large samples contain enough unusual cases to robustly estimate heterogenous effects. Small data sets are blunt tools able only to detect large average effects. However, many associations of interest in sociology are contingent on individual and contextual factors."

However, the challenges of survey research are well-understood, and with this understanding comes sampling and weighting strategies, and ways of modifying survey design. Survey research is built around such challenges, as defining a sampling frame from which we can estimate selection, missing data, and coverage (Japec et al., 2015; Lazer and Radford, 2017) is a central part of rigorous survey research. Importantly, even when these problems cannot be overcome, conclusions made on the basis of survey data can take into account knowledge of these limitations and be appropriately circumspect with conclusions.

What I call for in this thesis is to build similar understandings of the limitations of digital trace data: the ways in which they can be biased and non-representative, in terms of demographics or contexts (e.g., generalizing from one social media platform to another), in order to either explicitly correct for when possible, and to know how to limit the scope of our claims when not.

There is no doubt that overblown claims have been made; assuming generalizability unless shown otherwise, which is an unfortunate trend in social media and sensor research (and popular coverage thereof), and is not likely to yield robust, rigorous findings (or long-term public confidence and continued support for research funding).

To take Gayo-Avello's (2012; 2013) example of predicting election results using Twitter, he claims that it is simply not possible, and provides modeling reasons why existing claims of prediction success should be treated as suspect. To this I would add that it also may be that the *informational content may simply not be present*. Of course, we can never *prove* that we cannot predict election results using Twitter; there is always the possibility that there is *some* feature extraction and *some* sophisticated and complex model that will extract a reliable signal from tweets. But the focus of this thesis is on theoretical reasons and empirical demonstrations to believe that some types of signals are not captured by digital trace data, and so any feature extraction or model that claims to extract or make use of such information (e.g., even with seemingly rigorous and strong demonstrations of external validity) is likely to be overfitting and not robust. If we understand what sort of processes we are and are not able to observe with digital trace data, we would be able to theoretically state that digital trace data would need to be paired with other forms of inquiry, whether qualitative or surveys, to have the informational content that would allow for robust findings.

With digital trace data, some challenges are new, but many are the same as with previous forms of data. As Lazer and Radford (2017) write, "In principle, big data archives offer measures of actual behaviors, as compared with self-reports of behaviors." However, actions are not necessarily unambiguously tied to underlying "sociocultural constructs, which are arguably more cognitive and normative than behavioral. How does one observe love, affection, or deceit from cell phone data?" I believe that these sociocultural constructs are what are causal for behavior, and so are what would in principle lead to both explanations and to predictions that are robust to changes in context. Of course, we were (and are) never able to physically measure psychological constructs; they may have no *a priori* physical reality (e.g., to know if some scanned brain activity is of 'happiness,' we would need to first have brain scans taken under already known experiences of happiness to serve as a reference). We always needed behavioral measurements or self-report to infer underlying psychological or sociocultural constructs, and the field of psychology in particular has developed systematic ways of doing this (DeVellis, 2017). These involve models such as factor analysis along with methodological approaches of testing for consistency and various forms of validity (e.g. face validity, external validity or generalizability, internal validity, and criterion-related or 'predictive' validity; Babbie, 2010), all with strong guiding theory.

Another, almost obscenely clichéd issue, is selection bias. Gayo-Avello (2011) warns of turning social media into another 'Literary Digest' poll, after a nonrepresentative 1936 poll that was disastrously wrong in predicting the presidential election. Similarly, Ruths and Pfeffer (2014) warn about a 'Dewey Defeats Truman' moment from social media data, another humiliatingly wrong prediction about a presidential election made in 1948, again on the basis of nonrandom sampling—as well as a rush to publicize findings (also a lesson in itself). Lazer, Kennedy, et al. (2014) echo this when identifying "Big Data Hubris," the idea that more data can solve any problem, and Lazer (2014) identifies the failure of Google Flu Trends (described in Lazer, Kennedy, et al., 2014, and discussed further below) precisely as big data's 'Dewey Defeats Truman' moment. Indeed, the hard-won lessons of the past about random sampling seem to be a lesson frequently forgotten in machine learning, as in the *mea culpa* of Cohen and Ruths (2013):

> “To the reader uninitiated in latent attribute inference [of the latent attribute of political affiliation]s, these performance claims can easily be taken to be an assertion about the performance of the system under general conditions. In fact, we suspect that most authors of these works had similar assumptions in mind (author’s note: we did!). Regardless of intentions, as we will show, past systems were not evaluated under general conditions and, therefore, the performance reported is not representative of the general use case for the systems.”

In fact, the methodological flaws that Cohen and Ruths (2013) admit to and call out in research are not just about the use of convenience samples, but selection on the dependent variable, as in the critique of Tufekci (2014), which Cohen and Ruths’s (2013) work shows leads to the appearance of results when there in fact is no modeling success. Specifically, people who explicitly expressed political affiliation on Twitter were an non-representative group, such that the performance of classifiers trained and tested on this group were much higher than performance in more general populations.

Like survey sampling, data trace data are not representative, but we can still find ways to make use of it. An American Association of Public Opinion Researchers (AAPOR) report on non-probability (nonrandom) sampling (Baker et al., 2013) recognizes that asking whether nonrandom sampling can be accurate (i.e., having statistical guarantees about generalizability, i.e., frequency guarantees about the size of errors of estimates) is a futile question given the infinite ways that sampling can be nonrandom, and instead it is most worthwhile to ask what nonrandom sampling can tell us. One proposal is outlined by Lazer and Radford (2017), who distinguish between seeing digital platforms as a microcosm of society as identified by Tufekci (2014), and seeing platforms as “distinctive realms in which much of the human experience now resides.” That is, Twitter is representative of Twitter and Facebook is representative of Facebook; Lazer and Radford (2017) identify the assumption that these platforms are general enough to be worthy of study, and therefore have scientific relevance, which I believe is correct.

Beyond selection bias, there are multiple contextual processes that have a causal impact on the observed data, processes that are missed if the data are taken at face value. To continue the analog with survey data, these are the equivalents for digital trace data of biases like social desirability bias, and forms of response bias. Many social media sites have emergent norms and conventions (Marwick and boyd, 2010; Honeycutt and Herring, 2009), variation in users and usage (Burke, Kraut, and Marlow, 2011), variation over time (Burke and Kraut, 2014; Liu et al., 2014; Efstathiades et al., 2016) and geography (Poblete et al., 2011), and different motivations for adoption that lead to different patterns of usage (Hargittai and Litt, 2011; Jacobs et al., 2015). Just because we observe records of actions rather than rely on self-report does not mean that forms of generalizability-dampening biases are no longer present; especially when trying to generalize from digital trace data to larger social processes, contextual factors present serious threats to validity.

Studying such contextual processes is the first task of my thesis including, in Malik, Lamba, et al. (2015) (Chapter 1) the first empirical demonstration of some concerns previously raised only theoretically in areas such as media studies and sociology (van Dijck, 2013; Gehl, 2014; Tufekci, 2014; Ruths and Pfeffer, 2014; Healy, 2015), and also discussed in Lazer and Radford (2017). However, the opportunities to study some of these processes are rare, and in general we will not be able to know about their presence or strength in a particular dataset. The second task, then, is to provide positive exemplars for how to proceed in scientific research and public policy. This leads to my thesis statement.

> **Thesis statement:** Social media and sensor data do not give unbiased, generalizable findings about human behavior: inferences about constructs are complicated by selection bias, medium-specific norms and culture, and algorithmic user manipulation, and raw measurements are of physical quantities rather than of causal underlying social constructs. But by studying these forms of bias and the data-generating processes of such data and understanding their limitations, we can establish proper scopes and study designs within which findings will be accurate, reliable, and fair for use in business decision-making, scientific research, and public policy.

Throughout, I pay particular (although not exclusive) attention to social network data. Social networks are representation of multiple types of relational phenomena (Borgatti, Mehra, et al., 2009), which have been shown to potentially capture causal processes of influence and group formation/dissolution that are not captured by any other kind of measurement (Moreno, 1934; Sampson, 1968; Zachary, 1977; Krackhardt, 1996). Social network principles are present throughout successful uses of digital trace data: these principles explain the effectiveness of the original PageRank algorithm, as well as the success of using person/product affiliation matrices to make recommendations as Amazon did. More direct applications of ideas of social network analysis are behind many link recommendation systems such as "People you may know" systems on both LinkedIn and Facebook, and the analogous "Who to follow" on Twitter (Su et al., 2016). There is also evidence that social network analysis techniques are popular in intelligence agencies, with the NSA using a "three degrees of separation" rule for extending investigations from metadata (Bump, 2013), and former director of the NSA and the CIA General Michael Hayden once declaring that "we kill people based on metadata" (Cole, 2014), pointing to deeply consequential usages. Law enforcement agencies in the US use networks in social media for surveillance (LexisNexis® Risk Solutions, 2014) including of activists not suspected of wrongdoing (Ozer, 2016). Facebook has recently patented a system for using Facebook friendships for determining credit scores (Meyer, 2015; Lunt, 2016), and the startup Lenddo used Facebook networks to make microloan decisions in the Philippines and Colombia as a proof-of-concept before turning to selling their system to financial companies (Morozov, 2013b; Hempel, 2015).

So far, I have identified problems only of generalizability and scientific rigor, not of ethics. Many uses of digital trace data will violate informed consent, but here I agree with Watts (2014), who notes that companies (who are the holders of the most digital trace data about us) have few practical constraints what they do with our trace data. Limiting the ability of researchers to publish about what such data can tell us only means that we will not know the extent to which or the accuracy with which companies can infer or predict things about us, not that such inferences or predictions are not being made or that they are not consequential. Informed consent is a tool for preserving autonomy and self-determination, but it is neither necessary nor sufficient to handle the threats from digital trace data to autonomy and self-determination.

I see two distinct regimes of dangers: one is that data and modeling *work* in achieving what modelers claim to do, and that this power is inequitably distributed such that it empowers those few who have access to data, computational resources, and modeling skills. Especially considering cases where repressive governments might have the ability to identify and more efficiently persecute dissidents, this is frightening. The other regime of danger is that these systems do *not* work, leading to unnecessary suspicion or even targeting, and to unjust distribution of resources such as employment opportunities and access to credit. In such cases, a lack of transparency and accountability means that whether or not these systems work in doing what they claim, they will successful maintain existing power structures (O'Neil, 2016; Pasquale, 2015; Eubanks, 2018; Bopp et al., 2017). In the first regime, the way to preserve autonomy and self-determination will be to either or democratize modeling knowledge and access to data or to restrict models that can be built or actions that can be taken from them. In the second regime, we would need to guarantee rights to audit and challenge data, models, systems built on them, and decisions made based on them. Practically speaking, some tasks will fall into one regime, while other tasks will fall into the other, so it important to investigate which is a work.

This thesis does not address the dangers of either of these regimes directly; but by investigating where and when digital trace data can 'fail', I contribute to knowing which regime we are in, and consequently which regulatory responses will ultimately be appropriate.

Background

How does modeling work? *Does* modeling work? Stinchcombe, in *When Formality Works* (2001), discusses how formal systems succeed in organizing the world. Many of his examples are mainly about organizations and the division of labor (e.g., in architectural/construction diagrams), rather than statistical models; divisions of labor must have some sort of 'real' relationship with the labor being divided, but it is fairly easy to imagine alternative divisions that would, if given enough attention, successfully accomplish the same task. However, mathematical formalism is distinct. It is hard to imagine a system other than Newtonian mechanics, or something equivalent to it, that could capture basic physical processes (of course, this intuition has been critiqued, including that how we even identify phenomenon is not inevitable or obvious (Hacking, 2000), and just because we cannot imagine alternatives does not mean that there could be alternatives; Kuhn (1962) famously pointed out that phlogiston theory explained phenomenon as well as oxygen theory in terms of making equally accurate predictions). Wigner (1960), in "The unreasonable effectiveness of the language of mathematics in the natural sciences", ultimately concludes that

> "The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in future research and that it will extend, for better or for worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning."

This has been taken up formally philosophically (Bangu, 2016; Sarukkai, 2005), pointing out that the seeming miracle may be an artifact of our framings rather than a mystery to be explained. But it is clear that even within the same framing, the same miracle is not present for social systems. Gelman and Shalizi (2012) write,

> "Social-scientific data analysis is especially salient for our purposes because there is general agreement that, in this domain, all models in use are wrong – not merely falsifiable, but actually false. With enough data – and often only a fairly moderate amount – any analyst could reject any model now in use to any desired level of confidence."

While this talks of statistical significance rather than predictive power, even when it comes to prediction, statistical models of social systems have performances that are very far from the kind of performances achieved on physical systems. Still, statistical modeling is superior to subjective, heuristic and non-systematic approaches, as is most famously shown in the work of Meehl comparing clinical judgement to statistical prediction (Grove and Meehl, 1996; Grove, 2005; Meehl, 1954).

On the other hand, Meehl did not advocate for rejecting clinical approaches (Meehl, 1996), and argued for a mixture of methods. I think that any rejection of qualitative methods in favor of making use of quantities of data is dangerous, and note that the places where data and modeling are most limited (in determining *what* to measure, and in getting at the meaning-making that is the ultimate characterization of the human experience; Patton, 2015) are where qualitative methods excel. Wang (2013) calls for 'Thick Data' as a counterpart to big data; by pointing to some of the limitations of a data-and-modeling-only approach (albeit one that uses data and modeling itself to discover such limitations), this thesis supports such calls.

Much of what I implicitly critique, uses of social media or sensor data alone to make larger findings or conclusions, are uses of data and modeling that are justified by the impression that modeling 'works' for describing and controlling the world, rather than a careful case-by-case investigation of the internal validity, external validity, construct validity, etc., of a particular model. Note that one of the best places for counterpoints to this impression is from statisticians themselves. The statistical literature is filled with remarkably insightful, open and honest reflections about the tenuousness of the connection between models and reality, and pragmatic suggestions for modeling rather than dogmatic defenses in light of that (Buja et al., 2016; Box, 1979; Cox, 1990; Kass, 2011; Breiman, 2001; Chris, 2002; Freedman, 1997; Freedman, 1991; Berk and Freedman, 2003; Fisher, 1922; Shmueli, 2010; Gelman and Shalizi, 2012).

As for why such discussions are not more widely known, I believe they are overwhelmed by a competing narrative: the larger cultural ideas of what 'algorithms,' or 'big data,' or machine learning/artificial intelligence can accomplish, in what Jasanoff and Kim (2015) refer to as *sociotechnical imaginaries*. The existence of these terms also help distance machine learning from statistics, even while statistical machinery underlies its algorithms. Separated from statements of responsibility from statistics, imaginaries around algorithms and AI (and the market for expressions of that) have expanded. In addition to producing mistaken uses of data and modeling that could have been avoided by better heeding statistical advice, imaginaries like approaching the 'eye of god' with sensor data (Aharony et al., 2011) will have downstream effects in what scientists work towards accomplishing. As Morozov (2013a) charges, endemic among technology developers is a mindset of "technological solutionism": the view that every social problem can be reduced to a technical one, and having a corresponding technical solution. This both ignores causal social factors and thereby fails to solve problems, and in addition, contributes to disempowering individuals (making them dependent on technology to accomplish tasks). As an alternative, he points to the possibilities of technologies that enrich people's lives and deepen their experiences. Imagining digital trace data as approaching complete knowledge risks leading to research that ignores biases in data and statistical lessons of the limitations of modeling.

While I do not provide alternative imaginaries as does Morozov (2013a) with using technology for enrichment, or as does Gehl (2014) with his idea of social*ized* media, I strive to carry out rigorous, critical reflection about the nature and uses of digital trace data.

The closest existing parallel to what I attempt to accomplish is Agre's (1997) idea of *critical technical practice*. Agre discusses his own intellectual journey from artificial intelligence to social science; while his story is about the 'hard AI' rather than the 'soft AI' of machine learning, many of his critiques apply equally well to computer science using social data, particularly in how computer scientists "insist[s] on trying to read everything as a narration of the *workings of a mechanism*" (emphasis added). That is, unless social science theory can be read as describing some system that can be built, some action that can be taken, or some predictions that can be made (with a certain feature set), it is meaningless. He continues describing his process of breaking out of this way of thinking:

> "In broad outline, my central intuition was that AI's whole mentalist foundation is mistaken, and that the organizing metaphors of the field should begin with routine interaction with a familiar world, not problem-solving inside one's mind. In taking this approach, everything starts to change, including all of the field's most basic ideas about representation, action, perception, and learning. When I tried to explain these intuitions to other AI people, though, I quickly discovered that it is useless to speak nontechnical languages to people who are trying to translate these languages into specifications for technical mechanisms. This problem puzzled me for years, and I surely caused much bad will as I tried to force Heideggerian philosophy down the throats of people who did not want to hear it. Their stance was: if your alternative is so good then you will use it to write programs that solve problems better than anybody else's, and then everybody will believe you. Even though I believe that building things is an important way of learning about the world, nonetheless I knew that this stance was wrong, even if I did not understand how.
>
> "I now believe that it is wrong for several reasons. One reason is simply that AI, like any other field, ought to have a space for critical reflection on its methods and concepts. Critical analysis of others' work, if done responsibly, provides the field with a way to deepen its means of evaluating its research. It also legitimizes moral and ethical discussion and encourages connections with methods and concepts from other fields. Even if the value of critical reflection is proven only in its contribution to improved technical systems, many valuable criticisms will go unpublished if all research papers are required to present new working systems as their final result."

Agre labels this insistence on demonstrating critiques the "fallacy of alternatives". I follow Agre in arguing that I make a major contribution via my reflection, clarification, and theorization of what we can and cannot accomplish with modeling and digital trace data, apart from any technical contributions I make.

In the remainder of this section, I consider theoretical frameworks for how to think about digital trace data. I consider frameworks from critical and sociological work for thinking about the roles data and modeling play in society, but also seek to make technical clarifications. I then review the limitations of what statistical modeling, and by extension what machine learning, is able to accomplish with data. Lastly, I review current knowledge about sources of bias and threats to validity in digital trace data.

Critical and sociological work

Recent literature in social science and humanities has begun theorizing social media platforms. Two works in particular, van Dijck (2013) and Gehl (2014) I have found valuable. van Dijck (2013) theorizes that social media platforms are *techno-cultural constructs*, consisting of technology, users and usage, and content, but also *socioeconomic structures*, consisting of ownership, governance, and business models. Much of social media research within computer science looks only at the techno-cultural side, and looking at socioeconomic structures (as I do with governance in Chapter 2) has important implications both for understanding the platforms at large and for understanding the data produced from them. Gehl (2014) does not have as comprehensive a framework, but theorizes power relations found in digital trace data, saying that the databases of social media companies are “archives of affect, sites of decontextualized data that can be rearranged by the site owners to construct particular forms of knowledge about social media users.”

There is also a larger body of work about the societal impacts of machine learning systems. Much of this work focuses on the automated nature of control, and discusses under the umbrella of ‘algorithms’ (Gillespie and Seaver, 2016; Gillespie, 2014; Ziewitz, 2016). This literature provides some excellent theoretical ideas, although I believe that models are far more important than acknowledged in this literature, something which I attempt to clarify.

Specifically, there is work applying ‘algorithmic’ to governance (Gourarie, 2016), criticism (Ramsay, 2011), accountability (Diakopoulos, 2014; Diakopoulos, 2015), power (Bucher, 2012; Diakopoulos, 2014), discrimination (Miller, 2015), systems (Muñoz et al., 2016), culture (Dourish, 2016), paranoia (McQuillan, 2016), auditing (Sandvig et al., 2014; O’Neil, 2016), and harms (Tufekci, 2015). There are attempts to theorize ‘algorithms’ themselves (Gillespie, 2014; Ziewitz, 2016), in an area sometimes called ‘critical algorithm studies’ (Gillespie and Seaver, 2016). There is also the label of ‘critical data studies’ that deals with many of the same issues and references similar core literature (Iliadis and Russo, 2016; Dalton and Thatcher, 2014). The framing from this literature I find most helpful is ‘algorithmic governance’ (Ziewitz, 2016): this has a dual meaning in referring to both how models are governed, and the ways in which models are used to govern social processes and exert power.

However, I would argue that the core issue addressed in almost all of these works is actually that of the combination of *data and modeling*; not algorithms, and not data alone. I recognize the utility and rhetorical force of converging on the use of ‘algorithmic’ even if it is not the right term (e.g., Feldman et al., 2015, never mention ‘algorithmic fairness’ in their actual paper; it appears on a project website).[1]

Much of the power of data and models indeed come through the implementation, automation, and scalability of data pipelines through algorithms and software. And machine learning may refer to models as algorithms (especially if a model is synonymous with its implementing algorithm, like the perceptron). But in many cases, a statistical model can be abstracted away from the algorithms used to implement it. For example, whether a logistic regression is implemented using the algorithm of iteratively weighted least squares (or some other second-order method) or the algorithm of stochastic gradient descent (or some other first-order method) would be irrelevant when considering the relative weights the model gives to different predictor

[1] http://fairness.haverford.edu/

variables. Subsuming everything under ‘algorithms’ risks losing sight of the importance of considering the statistical *logic* of a model: that is, when, why, and how a model works.

Granted, some models are only computationally tractable given a particular algorithm or formalism (like kernel methods that use the ‘kernel trick’ of using inner products of observations to do infinite-dimensional regression), and certain models or approaches are chosen over others because of computational tractability (e.g., ℓ_1 regularization is convex and happens to result in variable selection, unlike ℓ_0 regularization which is actual variable selection but is not convex, so ℓ_1 regularization is used for variable selection), but this only means computational considerations affect the choice of model; after that, the logic of the model is ultimately what informs the consequences of its use. Applying this to ‘black box’ machine learning approaches that use models that pick terms and forms through automated procedures (rather than carefully chosen and tested model specifications), with the resulting models having superior predictive performance (Breiman, 2001), questions about if and when we can rely on these models need to be answered at least partially by investigating how modeling mechanics relate to the real world (e.g., in terms of the bias-variance tradeoff, covariate shift, etc.) including what it takes to rigorously validate these models (e.g., avoiding overfitting, accounting for dependencies between training and test sets, etc.).

In contrast, for talking about the limitations and harms of modeling, the machine learning literature has conversely coalesced around ‘fairness, accountability and transparency in machine learning’ (FATML). This I believe goes too far in the other direction of focusing solely on modeling, and neglecting to consider the larger systems for generating and collecting the data fed into models. There are intriguing proposals for eliminating disparate impact by building models that can handle protected or sensitive features, but I remain unconvinced that such technical solutions are flexible enough to cover the range of ways in which data and models may have negative consequences. For example, a technical solution that guarantees (some notion of) fairness based on *observed* covariates, and that has no way of identifying or correcting for correlations caused by historical inequities (e.g., how black populations in the United States were systematically excluded from home ownership and other means of accumulating intergenerational wealth) would be insufficient. Carr (2014), in a review and critique of Pentland (2014), writes,

> “A statistical model of society that ignores issues of class, that takes patterns of influence as givens rather than as historical contingencies, will tend to perpetuate existing social structures and dynamics. It will encourage us to optimize the status quo rather than challenge it.”

(Note that this quote, and especially the idea of “optimizing the status quo”, actually applies far more to machine learning approaches that automatically fit models to past data, rather than the more physics-style statistical models found in Pentland, 2014 that employ strong parametric assumptions—but even if the critique is not leveled at the most appropriate target, it is still a good critique. Also note that I discuss the models of Pentland, 2014, including their parametric assumptions and relationships to other statistical or machine learning models, in chapter 3.)

Statistical modeling

I focus exclusively on models based on probability, that is, those of statistics and machine learning. Alternatives to probability-based modeling are what Kolaczyk and Csárdi (2014) refer to as ‘mathematical’ models that can sometimes be analytically manipulated (such as classical mechanics models in physics), and simulation modeling (Gilbert and Troitzsch, 2005), which I discuss in Pfeffer and Malik (2017); for social science, both of these are primarily useful as a tool of *theory development*. For dealing with *data*, I believe both of these are insufficient, as they can only be initialized with some data and then the outputs qualitatively matched with other data. Neither can directly manipulate data to discover or estimate relationships. While both mathematical and simulation modeling can make 'predictions', quantitative predictions made from simulation tend to be extremely poor, and so this type of modeling is more commonly used to make qualitative statements about trends rather than quantitative statements about the magnitude of trends. Statistical methods, by using probability as a model for this variability (and, in statistics but less so in machine learning, also using probability to model the uncertainty of estimates; Cox, 1990), present a principled and systematic way of distinguishing underlying patterns from noise directly using data that is effective in making predictions. Both mathematical and simulation modeling have important use cases in the scientific process, but I do not consider them here.

It is also worth noting the relationship between statistics and machine learning, as machine learning is still poorly understood within social science. Machine learning is defined as a branch of artificial intelligence, devoted to machines that improve from 'experience' (Mitchell, 1997). It was originally an attempt to have machines improve by use of approaches like rule-based reasoning (for example, the 80s review of the field in Carbonell et al., 1983, made no mention of using statistical models); it was only decades later that, instead of mimicking what we theorize human learning mechanisms to be like, researchers discovered that a more effective approach to producing "intelligent" behavior was to use predictions of statistical models, fitted on large amounts of data (Halevy et al., 2009). The result is that today, machine learning is almost entirely based on statistics (Wasserman, 2014). The 'learning' is only a metaphor for mathematical optimization for fitting statistical parameters rather than the original (or 'strong') AI notion of machines with the ability to reason about the world, and is arguably an example of what McDermott (1976) calls the'wishful mnemonics' of AI: the tendency of artificial intelligence researchers to name programs or functions after what the researchers want the software to be or imagine it as being. McDermott charges that this confuses what functions and programs actually are and how they actually carry out their functions, which is arguably at work among social scientists. The applicability of machine learning to social science applications has been a topic of excitement, anxiety, and controversy (Gayo-Avello, 2011; Gayo-Avello, 2012; Junqué de Fortuny et al., 2013; Dhar, 2013; Athey, 2017; Cohen and Ruths, 2013; Hofman et al., 2017; Hindman, 2015; Lin, 2015; Kleinberg et al., 2015; Mullainathan and Spiess, 2017; Wallach, 2018), alongside similar emotions around 'big data' (Savage and Burrows, 2007; Savage and Burrows, 2009; Webber, 2009; boyd and Crawford, 2012; Tufekci, 2014; Hargittai, 2015; Ruths and Pfeffer, 2014; Lazer, Pentland, et al., 2009; Lazer and Radford, 2017)

There are four ways in which modeling may 'fail' because of misunderstandings of the nature of modeling.

Prediction The first is the interpretation of the word 'prediction' and a bias towards positive results. A 'prediction' in machine learning and statistics is not used in the colloquial sense of a statement about the future, but is a technical term synonymous with a 'fitted value,' an important distinction that is often rhetorically lost. Gayo-Avello (2012) makes the critique that what is called prediction "isn't prediction at all. I haven't found a single paper predicting a future result. They all claim that a prediction could have been made, but the analysis is post hoc." This is somewhat unfair to the technical definition, but he is correct in insisting we do not conflate the technical term with its colloquial meaning. A model that reports that it can 'predict *X*' has usually not actually shown that it can accurately foretell the future, only that it has found a model that *fits well* to existing data. A good fit is impressive, but not actually a demonstration of prophetic power. Of course, the purpose of held-out or test data is to provide an unbiased estimate of the true error, such that the performance of a model on test data should be a statement about its 'prophetic' ability; but in practice, test-data is re-used for testing multiple possible models, which creates a distribution over models and leads to an insidiously subtle form of overfitting (Dwork et al., 2015). Furthermore, dependencies in data, such as from temporal autocorrelation or network structure, can effectively share information across training and test splits and thereby inflate performance (Hammerla and Plötz, 2015; Bergmeir and Benítez, 2012; Chen and Lei, 2018; Racine, 2000; Dabbs and Junker, 2016). A pseudonymous series of blog posts (Lowly Worm, 2012b; Lowly Worm, 2012a) critiquing the paper "Twitter mood predicts the stock market" (Bollen et al., 2011) points out that including future information when training models is effectively "time traveling", which biases accuracy upwards. As the author notes with some satisfaction, a hedge fund that partnered with the paper authors to implement a trading strategy based on the paper's findings ended up shuttering after a single month (Lowly Worm, 2013).

These problems combine with the publication bias (both among author submissions and publishing venues' acceptances) away from publishing negative results, creating the same kind of positive-result bias present and critiqued in other areas of science (Gayo-Avello, 2012).

Prediction and explanation Second, correlation is not causation, but more subtly, prediction is not explanation. No matter how well a model fits (how well it 'predicts'), it is no guarantee that it is acting on the basis of real associations or relationships that can be interpreted rather than just correlations, many of which may be spurious. This is a fact recognized in statistics (or at least since Breiman, 2001) but still not widely appreciated outside of it, as it is both not obvious and indeed contradicts the hope that we can find parsimonious and predictive models (Forster and Sober, 1994). Shmueli (2010) goes further to note that an explanatory model might predict poorly, and that a successfully predictive model might not explain anything, giving an explicit example from Wu et al. (2007) of an underspecified model having a lower (true) expected prediction error than refitting the correctly specified model to the data it generated. The technical reasons for this get into issues of the bias-variance tradeoff, well-known issues discussed in terms of "Stein's paradox" (Efron and Morris, 1977), and how model selection techniques can easily be led astray by collinearity and coefficients close to zero (Zhao and Yu, 2006; Geer and Bühlmann, 2009). In a helpful demonstration, Mullainathan and Spiess (2017) use variable selection on different subsets of a data set and show similar predictive performance, but with very different sets of variables selected in.

Prediction and explanation are distinct tasks, and different modeling approaches are appropriate to each. For explanation, we should carefully separate out variables that are dependent on each other, understand nonlinear effects and interactions between variables, account for dependencies between observations, choose appropriate functional/parametric forms, and make sure that all causal processes are measured. In that case, the weights assigned to certain features will hopefully reflect the true contributions of different variables to a given process. For prediction, we can ignore face validity, internal validity, construct validity, collinearity, model specification, and all the usual concerns: all that matters is external validity, established via cross-validation (Arlot and Celisse, 2010). Overfitting is a large threat to external validity, but is also dealt with via further data splitting, for example to choose optimal levels of regularization.

Note also that in this sense, ‘interpretability’ or ‘explainability’ is a red herring; an ‘interpretable’ or ‘explainable’ model risks creating the impression that the *logic of our interpretations* is the same as the *logic of modeling*. For example, a decision tree is perfectly interpretable and explainable, but if an analyst were to look at a decision tree and disagree with one of the branches based on their substantive knowledge and want to change it, doing so would destroy the mathematical integrity of the fitted model. The model was built with correlations (specifically, searching over some space of possible decision trees to see which one best fits the training data) and not logic or the domain analyst’s experience; the interpretability is an illusion. This is a hard distinction to convey; as David Jensen noted in a talk,[2] when he discusses explainability in machine learning he is frequently met with people jumping to causality, despite his stern warnings to not do so.

When is prediction appropriate? Breiman (2001) famously charged that prediction is sufficient for many problems, and that statisticians were neglecting such problems in favor of talking about data-generating process. He gave the example of a *detection* task, detecting the amount of chemicals in water, where machine learning was perfectly appropriate and indeed more so than statistical approaches concerned with the data-generating process.

More systematically, Kleinberg et al. (2015) identify ‘umbrella’ problems where prediction is sufficient, versus ‘rain dance’ problems where we need to know about causality. Less fancifully, Mullainathan and Spiess (2017) call them $\widehat{Y}$ problems and $\widehat{\beta}$ problems. In any cases where we are *intervening* and using the results of predictions to make decisions that affect the system from which the data were drawn (rather than just reacting), we need to know true associations and causal processes to know what the result of interventions will be (and thereby make successful interventions). Different causal processes behind the same prediction will suggest very different interventions strategies: Aral et al. (2009) give the example of correlation in behavior among people connected in a digital social network. If this correlation is due to latent homophily, it suggests a blanket marketing strategy; but if the correlation is due to influence, it suggests identifying and targeting influencers.

The stock market may give an example of where prediction seemed sufficient but is not; when people’s actions have an effect on the system, effects which predictive models do not and cannot anticipate, it can lead to unforeseen negative consequences such as high-frequency trading (which make automated buying/selling decisions based partly on predictive models) contributing to ‘flash crashes’ (Kirilenko et al., 2017).

Alternatively, even if prediction is sufficient, lack of knowledge of true associations or of causality can still lead to failure. The key example of this is Google Flu Trends, the failure of which is documented by Lazer, Kennedy, et al. (2014). The system, built by finding correlations between Google search data and previous years’ CDC reports of flu incidence, failed to detect an off-season spike of flu, and overestimated the rate during the winter. Lazer, Kennedy, et al. (2014) charge that the system was half a ‘winter detector’, rather than solely a flu detector, something that would not happen with a proper understanding of causal pathways

[2] “Explainable artificial intelligence: Opportunities and challenges for public policy”. Heinz College, Carnegie Mellon University, February 19, 2018.

(or even with robust associations) rather than simply predictions. However, the problem might be fixed by simply getting data over a longer period of time that is more representative and updating the model (Lazer, Kennedy, et al., 2014), as prediction is sufficient.[3]

Prediction may also be used as an exploratory tool (Lin, 2015; Dhar, 2013), albeit with caution and many caveats (Yang and Yang, 2016; Mullainathan and Spiess, 2017). Specifically, machine learning employs a bevy of variable selection techniques, both from statistics (like the lasso) and developed within machine learning (like Correlation-based Feature Selection; Hall, 1999). However, neither the classic method of stepwise selection nor techniques used in machine learning techniques have theoretical guarantees that the 'true' variables will be selected; guarantees that exist are only for a set of variables that helps give the best predictive performance (Mullainathan and Spiess, 2017; Geer and Bühlmann, 2009). While it is logical that variables that predict well in held-out data are either themselves substantively important, or at least are correlated with variables of substantive importance, selected variables must be interpreted with caution.

Causality Third, we can never truly get causality from observational data. In the case of networks, the enormous difficulty of separating out homophily and contagion is demonstrated by Shalizi and Thomas (2011), but even in the case of non-network data, Arceneaux et al. (2010) show that omitted variable bias is insidious and enormous, and that even perfect applications of observational inference techniques cannot overcome it. On the other hand, experiments can give causality but not the ecological validity needed to generalize the identified relationships; Centola (2010) presents fascinating results, but generalizing from the processes on the clean structures of the study's artificially created networks to real-world networks is difficult. I do not mean to ignore recent advancements in causal discovery (Spirtes, Glymour, et al., 2001; Spirtes, 2010), although non-independent and identically distributed (non-IID) data such as that in networks are currently beyond the reach of such methods (Spirtes and Zhang, 2016), and the guarantees behind such techniques rely on extremely strong and untestable assumptions (Freedman, 2004) that are potentially behind limited adoption.

Representation Fourth, for understanding, there is a problem that the core assumption of statistical modeling is that processes and attributes can be separated out into distinct variables (or features, in machine learning). This, as Abbott (1988) famously pointed out, starkly contradicts the assumptions of major theoretical traditions of sociology. For the task of prediction, if the artifice of variables can be successfully used to predict meaningful future measurements, then this objection is sidestepped. However, as detailed above, being able to predict something does not mean we have understood it, both in a very practical sense of being able to successfully predict the result of interventions and in a humanistic sense of meaning-making being the ultimate target of understanding. Erikson (2013) makes a related argument specifically in the case of social networks; approaches that seek to explain and predict processes with the abstractions of social networks, without regard to the content of ties, made a set of theoretical assumptions about what is driving the world. And, as we establish above, even if 'formalist' approaches achieve (observational) predictive success, that at best be taken as a suggestion of explanatory power but is by no means a proof of it.

[3] Or at least it would be unless feedback loops (people being concerned that their flu status being known) lead people to hide or manipulate (Lazer, Kennedy, et al., 2014) the signal that is being used as a proxy to the behavior of interest.

Even if we accept the artifice of variables as the terms in which we understand the world, as discussed above, the actual variables of interest are usually posited underlying theoretical constructs that cannot be directly observed. Ways of getting at such constructs should be reliable (repeated measurements produce the same value, unless there is an actual change) and valid (they covary with other variables in the ways posited by theory) to be meaningful (DeVellis, 2017).

Altogether, we have that models that predict well will not necessarily generalize, and do not necessarily give insight into underlying processes. Conversely, arguments about (causal) processes must always be tentative, whereas experimental designs that can establish causality still must be tentative about ecological validity. All of these contribute to how modeling may 'fail' because of misunderstandings of the nature of modeling.

Then, there are challenges with the data that go into models.

Data

There have been multiple critiques of the nature of social media data, whereas sensor data has remained relatively unaddressed (something I rectify in Chapter 3). I would systematize possible concerns into three sets: those around norms and culture, those around variability in users, and those around socioeconomic structures. Part of this is taken from Malik and Pfeffer (2016a).

Norms and culture The first set of concerns are around *norms and culture*. Platforms can have specific cultures that pose a challenge to generalizability. Twitter, for example, has idiosyncratic conventions and cultural norms around the use of mentions and hashtags (Honeycutt and Herring, 2009; boyd, Golder, et al., 2010; Java et al., 2007; Kwak et al., 2010) that do not necessarily generalize (although other sites adopted Twitter's idea of hashtags), as well as having a culture of a many-to-many model of communication (Marwick and boyd, 2010) that again is not necessarily the same as other sites. Differences in norms definitely affect observed behavior, as Newell, Dimitrov, et al. (2016) demonstrate with reviews. Twitter also has an ugly cultural side of enormous hostility towards and harassment of women and those of marginalized identities (Matias et al., 2015) that sadly often does generalize across platforms, and that provides pressures that can limit participation in nonrandom ways (Stevenson, 2014).

Variability in users The second set of concerns are around *variability in users*, which come through motivations and adoption patterns, demographics, geography, and time. If adoption is nonrandom over the population, it has large consequences for how behavior generalizes and for how well the network structure captures and reflects an 'underlying' (or previously present, or global) network of social relations (Schoenebeck, 2013). Hargittai and Litt (2011) find that interest in following celebrity news is a major predictor for joining Twitter, which has consequences for the type of behavior observed there (such as following celebrities). Jacobs et al. (2015) show that there are differences in networking behavior between Thefacebook.com adopters who joined prior to the site opening to the public (i.e., those from a relatively small set of elite higher education institutions) and those who joined after. Studies which compare properties of online social networks to previously collected offline social networks (Wilson et al., 2012; Corten, 2012; Quercia et al., 2012; Ugander et al., 2011; Mislove, Marcon, et al., 2007) provide other evidence for differences. Furthermore, considering adoption alone is not sufficient, as platforms are the location of interactions and flows that potentially change pre-existing social relations and create new ones (Burke and Kraut, 2014).

One main theory for explaining behavior on online platforms relates to the presence of an imagined or "invisible audience" (boyd, 2008; Marwick and boyd, 2010; Litt, 2012; Bernstein et al., 2013); users' behavior on online platforms is a performance for such audiences. Users engage in such performance from social needs (Raacke and Bonds-Raacke, 2008) such as a need for self-presentation as well as to conform to peer pressure (Krasnova et al., 2008), such that their behavior will reflect the expectations or demands of peer groups or whoever is imagined to be their audience. For example, a link on either Twitter or Facebook (in different ways) is not just the opening of a pathway of communication (both symbolically and, under certain privacy settings, literally), but is a symbolic act in itself to signal (Donath and boyd, 2004; Lampe et al., 2007; Donath, 2007) approval, validation, popularity, or other attributes, depending on the context of the (total, cross-context) dyadic interaction (Kooti et al., 2012; Huang et al., 2010; boyd, Golder, et al., 2010). This also implies that the things that are easiest to measure may not be the things that are most important to measure. The basic construct of a tie on a platform is the easiest for gathering network data, but is not necessarily the most useful. One line of work shows that far more informative are other types of signals such as communication and interaction (Jones et al., 2013; Burke, Kraut, and Marlow, 2011; Romero et al., 2011; Viswanath et al., 2009), browsing (Schneider et al., 2009; Benevenuto et al., 2012), and positive words in Facebook walls and inboxes (Gilbert and Karahalios, 2009).

On the demographics side, again taking Twitter, work shows that its users are not demographically representative using both representative surveys (Duggan et al., 2015) and comparisons of Twitter data to Census data (Hecht and Stephens, 2014; Mislove, Lehmann, et al., 2011). Furthermore, social media platforms are neither globally uniform (Poblete et al., 2011) nor a static, stable environment across years (Liu et al., 2014).

Socioeconomic structures The third set of concerns relate *to socioeconomic structures* (van Dijck, 2013). The possibility of making money from link farming (Ghosh, Viswanath, et al., 2012) or from selling bots to inflate metrics (Donath, 2007) has attracted spammers, and as anybody who analyzes Twitter data quickly finds, spam is widespread (Thomas, McCoy, et al., 2013) despite Twitter's attempts to filter it out (Thomas, Grier, et al., 2011), and this can distort research findings (Ghosh, Viswanath, et al., 2012). Data access also falls into this category: Facebook data is nearly impossible to publicly access, and for Twitter, the most accessible channel of data (allowing for specific queries), the free Streaming Application Programming Interface (API) (Gaffney and Puschmann, 2014), has strict rate limits within which sampling is not necessarily random (Morstatter, Pfeffer, and Liu, 2014; Morstatter, Pfeffer, Liu, and Carley, 2013). This nonrandom sampling distorts not only absolute frequency (how often something appears on Twitter) but even relative frequencies (whether one thing or another is more frequent). An alternative, the Sample API, is a random sample so frequencies are proportional to incidence Twitter overall; but at 1% of Twitter, and no ability to request data about specific users, hashtags, languages, etc., there is not enough statistical power to detect small phenomena.

It is easy to understand how the actions that are technologically possible on social media platforms (Tufekci, 2014), would make a difference in behavior. For example, Facebook ties are symmetric, requiring mutual consent, whereas ties on Twitter can be directed, not requiring mutual consent, which means it is not meaningful to directly compare networks on the two platforms (some form of symmetrizing would be required). Then, there is how users react to affordances, for example how people either limit their thoughts to what can fit into 140 (now 280) characters on Twitter, or else find ways to encode that a narrative stretches across multiple tweets (e.g., the convention of starting tweets with "1/14, 2/14,. . . ,14/14"). But important to note is that affordances and platform designs are not incidental to socioeconomic structures. Relating to van Dijck's (2013) governance angle, Gehl (2014, p. 43) writes about how social media platforms are designed to have user labor act as "affective processors" to produce data about those users, data which are then stored and used, but that are inaccessible to users. In Chapter 2, I theorize as 'platform effects' (Malik and Pfeffer, 2016b) the ways in which social media sites are successful in encouraging certain types of user behavior and labor. The presence and pressure of platform features encourage activity (Burke, Marlow, et al., 2009) and networking, especially through recommender systems, but also through commercial incentives (such as promotions) and public pages on which unconnected users can interact (and potentially go on to connect). Beyond constraints, norms and engineering manipulation, users may change behavior in reaction to the site, for example in documented increases in the level of Facebook privacy settings (Dey, Jelveh, et al., 2012).

Again drawing from van Dijck (2013), social media platforms are not independent of one another but, through competition, shared users, and corporate and political links, form an "ecosystem of connected media" (van Dijck, 2013, pp. 18–23). Very little work has studied the 'shared users' side of this, with one rare example being Newell, Jurgens, et al. (2016) examining how users migrated between platforms in protest of Reddit policies. Such work is especially difficult given how gathering data from one platform is hard enough, let alone having to manage multiple platforms, and linking users across those platforms. Research findings are also part of this ecosystem: they have the potential to immediately be incorporated back into the platform design to do tasks like "designing feed algorithms, [and] promot[ing] content with topics that viewers are more likely to respond to" (Wang, Burke, et al., 2013).

Mobile phones and sensors Sensors and mobile phones have not yet been similarly critiqued, but their challenges are similar. There is the question of representativeness of who uses smartphones (or, whether the amount of data we get from different people is randomly distributed over smartphone users). There are systematic differences introduced by certain types of people using phones in certain ways (e.g., turning off location services to save battery), systematic differences introduced by different mobile phone manufacturers (including different protocols for saving energy in ways that affect different sensors differently across phone models, and different sensing hardware), differences in the data collected from contextual sources like detected cell towers and WiFi hotspots based on the density of those sources in different areas (e.g., rural vs. urban) and, potentially, changes in behavior introduced by services built on sensors (e.g., people's locations being influenced by the directions they are told to take from their map apps, or their choice of establishments to patronize being driven by geographic recommendations).

There is also not necessarily a 1-to-1 relationship between phones and people: Margolin et al. (2014)[4] found in national surveys that in the US 14% of respondents reported having more than one cell phone and 21% reported sharing their phone, numbers that were respectively 27% and 32% for respondents in Spain. Phones may not be reliable proxies for people in other ways: while monitoring one mobile phone study,

[4] This is unpublished work, provided by Drew Margolin via personal communication, June 2018.

researchers observed two phones charging side-by-side with the owner of neither phone nearby![5] Eagle (2005) discusses the related ‘forgotten phone’ problem, where study participants would sometimes forget their phones at home; Eagle addressed this by building a classifier based on accelerometer readings that could detect a non-moving phone as a proxy for being forgotten. More systematically, Patel et al. (2006) investigated the proximity of mobile phones to their owners, using a Bluetooth beacon on lanyards that participants wore continuously for detecting phones. The beacon detecting a strong Bluetooth signal from the phone indicated the phone being “within reach”, a weak signal indicated the phone being in the “same room”, and an absent signal indicated a phone further away. The 16 participants ranged from having their phone within reach 17% of the time to 85% of the time, with an average of 58%. Phones were within the same room on average an additional 20% of the time. Many episodes of phones being further away were when participants were at home, and were because of reasons like avoiding disruptions to self and others, regulation of phone use, and protection, in addition to forgetting phones. Dey, Wac, et al. (2011) replicated this study with smartphones some years later, showing that the previous finding about when phones were within reach still held, with 28 participants having their phones at hand only 50% of the time. However, the percentage of time that phones were within the same room was much higher, at an additional 40%.

While the concerns around variability are not new to statistics, and accounting for norms and culture within modeling is a recognized challenge, the effects of the *socioeconomic structures* in which data are generated present new modeling concerns that requires far more work that currently exists. Chapter (2) in particular focuses on this.

Responses to bias

What value do large-scale digital trace data hold in light of these problems? One solution is to mimic techniques for survey data, and devise sampling and reweighting strategies to get representative estimates from nonrepresentative samples. However, I believe this will not work in the case of digital trace data. The effects of socioeconomic structures within which data are generated change the nature of the data. In some cases, there is no neutral frame to try and recover with weightings and such (e.g., it is strange to imagine a notion of a ‘natural’ microblogging service, free from Twitter’s idiosyncratic capabilities, constraints, and platform design; after all, the environment of Twitter did much to establish the very idea of microblogging). If there is a theoretical neutral frame (some notion of people’s ‘actual’ social networks rather than the networks of their online ties, as is considered by Schoenebeck, 2013) then data to know how to recover it may be infeasible to access through limitations on access to data and users.

Abbott (2004) splits social science research across three levels at which questions are posed: case study analysis, which is “studying a unique example in great deal detail”, small-*N* analysis, which is “seeking similarities and contrasts in a small number of cases”, and big-*N* analysis, which is “emphasizing generalizability by studying large numbers of cases, usually randomly selected.” He elaborates on the intellectual case for small-*N* analysis: “By making these detailed comparisons, [small-*N* analysis] tries to avoid a standard criticism of single-case analysis—that one can’t generalize from a single case—as well as the standard criticism of multicase analysis—that it oversimplifies and changes the meaning of variables by removing

[5] David Lazer, personal communication, May 19, 2017.

them from their context.” Elsewhere, he suggests that a powerful heuristic in social science research is to shift the question: what I suggest doing is such a shift, disaggregating the idea of large-*N* analysis from generalizability, and instead imagining ‘large-*N* small-*N* analysis’ that uses a large number of cases to make the studies and conclusions with the level of claimed generalizability currently used for small-*N* cases. This, I argue, is the most effective response to digital trace data, rather than thinking it can replace representative sampling (Savage and Burrows, 2007; Savage and Burrows, 2009), the type of thinking that Lazer, Kennedy, et al. (2014) call “Big Data hubris”, or not dealing with the question of generalizability (Baker et al., 2013).

Small-*N* analysis, as set out by Abbott (2004), cannot statistically generalize; but as put by Luker (2010), small-*N* analysis permits ‘logical’ generalization where the researcher makes an argument about the larger relevance of findings of a specific group. Relatedly, Foucault Welles (2014) argues for the importance of marginal and small groups, conclusions about which by definition do not generalize because the groups do not represent the mean of society: for this purpose, small-*N* analysis is superior to large representative samples.

The approach I take in responding to the biases I identify is, for the biases I outline in a chapter, to present a companion study that suggests a study design and scope for the generalizability of findings that avoids the problems of bias for generalizability. While not as satisfying as a one-time or general-purpose technical fix, I believe that this approach is ultimately more robust, feasible, and (given all the unknowns of the data-generating processes of digital trace data) scientifically responsible.

Outline

This thesis is split into two parts.

In Part I, I present critiques of social media and sensor data, demonstrating ways in which they are not generalizable. Drawing on the above discussion, I show limitations of demographic representation in Chapter 1. In Chapter 2, I take up the idea of platform effects, looking at how behavior observed on social media platforms is causally influenced by the decisions of platform designers. Lastly, in Chapter 3, I provide the first empirical critique of uses of sensor data in studying social networks, focusing on tying assumptions made by existing work to existing theory in social network analysis.

In Part II, I suggest ways that we can overcome limitations in generalizability, paralleling the critiques of the first three chapters. In Chapter 4, I present work done in collaboration with public health researchers on Twitter; we propose answering the question of how to best make use of the massive and valuable amounts of public health information on Twitter as one of intervening and public outreach, rather than primary epidemological research. Lastly, in Chapter 5, I present the design and results of a study that I argue is a meaningful way to make use of mobile phone sensor data, supported by theory and appropriate statistical methods.

Taken together, I provide an important empirical contribution to complicating the nature of digital trace data, and follow up critiques with methodology-based responses to the limitations I identify.

Contribution

The contribution of this thesis is first to rigorously model certain subtle biases in social media data: Chapter (1) is the first nation-wide multivariate spatial analysis of the demographics represented in geotagged tweets, showing the ways in which geotagged tweets are non-representative. Chapter (2) is the first work to carry out an empirical demonstration of previously theorized “platform effects” (van Dijck, 2013; Gehl, 2014; Tufekci, 2014; Ruths and Pfeffer, 2014; Healy, 2015), and in its use of natural experiments with data artifacts, anticipates the calls of Lazer and Radford (2017). The methodology of this work has subsequently been applied in the context of online discussions and civic engagement (Pablo et al., 2017), and the theory has contributed to other frameworks (Jacobs, 2017).

In Chapter (3), I provide the first extension of critiques around social media data to sensor data, such as RFID tags and Bluetooth capabilities of mobile phones, that also serves as the first theorization of the nature of sensor data from a social science perspective. I consider what sort of constructs are captured in digital trace data, and what other constructs of interest are involved with sensor measurements. Using this, I identify promising future avenues for research that have not yet been pursued. This is also poised to make a major contribution, serving as the bridge between technical knowledge and sociological theory that can bring sensor research into the social science mainstream.

Chapter (4) responds to attempts to use Twitter for public health, many of which have focused on its potential for public health monitoring. However, given the ways in which we know Twitter is biased, this is not likely to be robust approach. Instead, we shift the scope of the goals; we use Twitter for what it was intended to do, monitor and interact with activity on Twitter. I also take special care to demonstrate rigor with machine learning, using cross-validation in ways that best inform us about out-of-sample performance.

Lastly, Chapter (5) acts on some of the recommendations I make in Chapter (3), carrying out a study with mobile phone sensors to compare proximity data to self-reported friendships. In contrast to previous research using mobile phone data to model friendships, I take a machine learning approach, using cross-validation to simulate an application setting for testing model performance. This is the first work to systematically consider co-location features, and has the potential to set benchmarks for friendship detection tasks.

Overall, this thesis is a major contribution towards the reliable, accurate, and responsible use of social media and sensor data across business, science, and policy. It engages in critical reflection to the use of digital trace data, and thereby helps “provides the field with a way to deepen its means of evaluating its research” (Agre, 1997). Furthermore, as pointed out in ‘critical algorithm studies’ (Gillespie and Seaver, 2016) and elsewhere (Gehl, 2014; Agre, 1997), there is a deep need to rigorously bring together critical social theory and computer science in order to ensure that large-scale computational systems are effectively serving humanity. By demonstrating theoretical critiques in modeling terms, this thesis is one of the first works to address this need.

Part I

Critiques

Chapter 1

Demographic biases[1]

Summary. Geotagged tweets are an exciting and tremendously popular data source. But, like all social media data, they potentially have biases in who are represented. Motivated by this, I investigated the question, 'are users of geotagged tweets randomly distributed over the US population'? I carry out a statistical test by which I answer this question strongly in the negative, by linking approximately 144 million geotagged tweets within the US, representing 2.6m unique users, to high-resolution Census population data. Utilizing spatial models and integrating further Census data to investigate the factors associated with this nonrandom distribution, I find that, controlling for other factors, population has no effect on the number of geotag users, and instead it is predicted by a number of factors including higher median income, being in an urban area, being further east or on a coast, having more young people, and having high Asian, Black or Hispanic/Latino populations.

Compared to the previously published version, I have an updated literature review, a correction to the main model (previously, the reference category of a categorical variable was incorrectly chosen, it has now been changed to the majority category), and updated figures (plotting skewed distributions as complementary cumulative density functions as is recommended in Clauset, Shalizi, et al., 2009, rather than as log-log scatterplots or CDFs).

1.1 Geotagged tweets

'Geotagged' or 'geocoded' tweets, where users elect to automatically include their exact latitude/longitude geocoordinates in tweet metadata, provide data that are:

[1] This is an updated version of a paper previously published as: Momin M. Malik, Hemank Lamba, et al. (2015). "Population bias in geotagged tweets". In: *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*. ICWSM-15 SPSM, pp. 18–27. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662.

- High-quality: geotagging is automated, so there are fewer chances of data error such as from user specification (Graham et al., 2014; Hecht, Hong, et al., 2011);[2]
- Precise: geotags are to a ten thousandth of a degree in latitude and longitude;
- Richly contextual: geotags are connected to tweets with all their temporal, semantic, and social content;
- Easily available, through the Streaming API;
- Large: using the Streaming API, a researcher can build a collection of tens of millions of tweets.

Unsurprisingly, this makes them an enormously attractive source for studying a wide range of human phenomena (Hong et al., 2012). Previous to the publication of Malik, Lamba, et al. (2015), works had used geotagged tweets to study

- mobility patterns (Hawelka et al., 2014; Yuan et al., 2013; Cho et al., 2011),
- urban life (Doran et al., 2013; Frias-Martinez et al., 2012),
- transportation (Wang, Al-Rubaie, et al., 2014),
- natural disasters, crises, and disaster response (Morstatter, Lubold, et al., 2014; Lin and Margolin, 2014; Shelton et al., 2014; Sylvester et al., 2014; Kumar et al., 2014), and
- public health (Sylvester et al., 2014; Nagar et al., 2014; Ghosh and Guha, 2013)

as well as the interplay between geography and

- language (Hong et al., 2012; Eisenstein et al., 2010; Kinsella et al., 2011),
- discourse (Leetaru et al., 2013),
- information diffusion and flows (Kamath et al., 2013; Liere, 2010),
- emotion (Mitchell et al., 2013), and
- social ties (Stephens and Poorthuis, 2014; Takhteyev et al., 2012; Cho et al., 2011).

Furthermore, maps of geotagged tweets tend to look remarkably similar to maps of population density (figs. 1.1 and 1.2; see also Leetaru et al., 2013), even if there are differences at a finer scale (figs. 1.3 and 1.4). This naturally leads to the question: are Twitter users who send geotagged tweets (henceforth, ‘geotag users’) randomly distributed over the population? This is a critical question because, if users who elect to geotag are systematically different from people in general, the results of studying geotagged tweets will not have external validity.

[2] Note that through the use of the API, users and services can tag their tweets with arbitrary geocoordinates. We found some evidence of this being used for generating high visibility in a spam-like manner, but only in a few cases. Still, what is most important is that the precise and numerical nature of geotags do not allow users to specify (linguistically) whimsical or ambiguous locations as they can do in the ‘location’ field (and users who whimsically locate their tweets in Antarctica or the middle of the ocean would not be picked up by a geobox around the contiguous United States, see below).

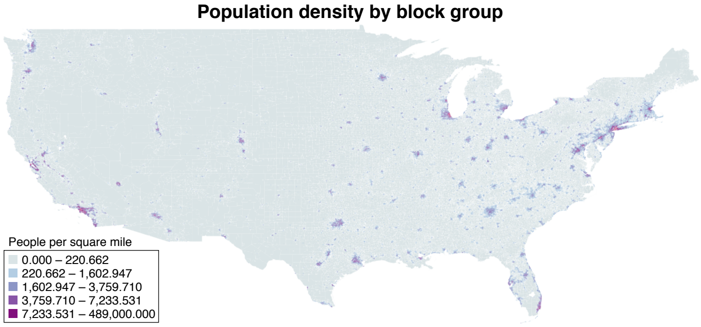


FIGURE 1.1: Quintiles of population per square mile by 'block group' (see below) in the 2010 Decennial Census.

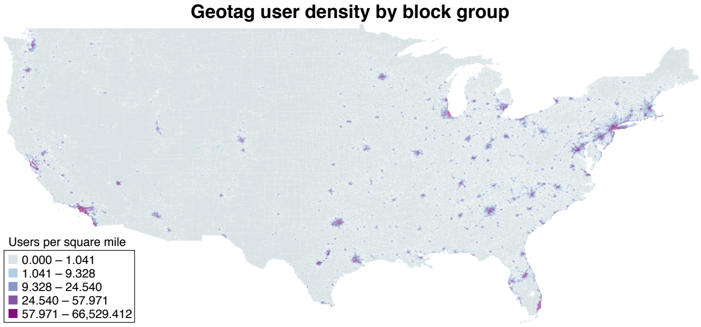


FIGURE 1.2: Quintiles of geotag users, uniquely assigned (see 'mobile users' below) per block group, divided by block group area.

Since this study's publication in 2015, geotagged tweets continue to be used for a variety of substantive purposes, such as studying home alcohol consumption (Hossain, Hu, et al., 2016), finding vectors of food poisoning (Sadilek, Kautz, DiPrete, et al., 2016), further looking at mobility (Fiorio et al., 2017), making its calls for considering the impact of biases as relevant as ever.

Conversely, this study has contributed to a growing area that seeks to study, understand, and correct for the biases discussed here. Citing my results, Brogueira et al. (2016) did not assume generalizability, and were careful to state results as first and foremost about Twitter users. Brent Hecht, whose 2014 article with Monica Stephens (Hecht and Stephens, 2014) greatly informed the theory of this study, published further work building on this result (Johnson, Sengupta, et al., 2016; Thebault-Spieker et al., 2017), including a work looking at how biases affect ultimate results (specifically, how geolocation inference performs more poorly for rural users). Montasser and Kifer (2017) took up weighting schemes to correct for population biases. Citing my result as one motivation, Mowery (2016) looks at the effect of misdiagnoses on attempts to estimate flu prevalence using Twitter. This work is even cited in further survey research (Mellon and Prosser, 2017), showing how new top-down approaches work with survey estimates to illuminate phenomena. In one case, McNeill et al. (2016) found that demographic biases did not significantly affect estimates of local commuting patterns.

In an independent angle of study about the meaning of geotagged tweets, coming some years after the

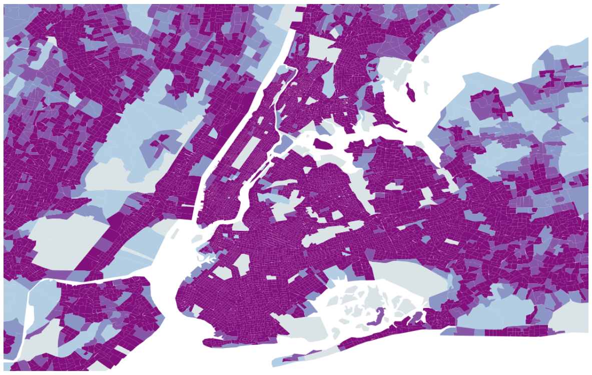


FIGURE 1.3: Detail of fig. (1.1) for New York.

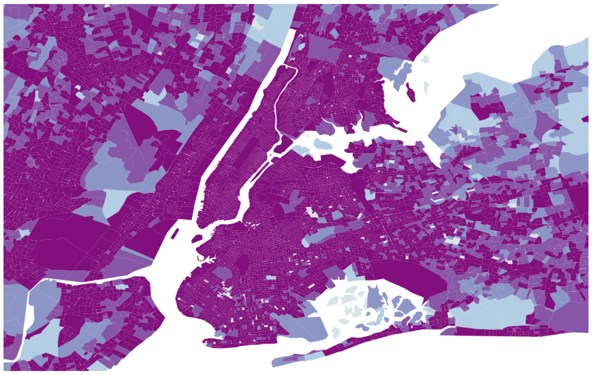


FIGURE 1.4: Detail of fig. (1.2) for New York.

publication of this study but complementary to it, Tasse et al. (2017) conducted surveys in which they found that geotag tweet users use the tags “consciously and turn geotagging on and off frequently.” They suggest thinking of geotagged tweets as “postcards, not ticket stubs”: that we should study them not as though they are a record of people’s behaviors, but as conscious and selective declarations of having been in a certain place at a certain time. This study, by looking from the perspective of user motivations, provides theoretical reasons that back my finding that geotagged tweets are not representative. This explanation of behavior also explains some of the qualitative results we observed, namely that airports were the heaviest outliers for their ratio of population to geotag tweet users: many people tweet to declare their travels, rather than to identify where they live or spend time.

Hemank Lamba, Constantine Nakos, Jürgen Pfeffer and I used the Twitter API to get a collection of 144,877,685 geotagged tweets from the contiguous US, from which we extracted 2,612,876 unique twitter handles. We uniquely assigned each handle to a *block group*, a geographic designation of the US Census Bureau that is the smallest geographic unit for which Census data is publicly available. We then linked the counts of unique geotag users per block group to the 2010 Decennial Census population counts per block group. I created a statistical test for the null hypothesis that geotag users are randomly distributed over the US population, and found sufficient evidence to reject this null. Using other Census data, I then use a Simultaneous Autoregressive (SAR) model (also known as a ‘spatial errors’ model) to test some candidate explanatory factors and investigate what is nonrandom about this distribution. This, to my knowledge, was the first paper to use statistical testing to establish population bias along multiple dimensions in geotagged tweets across the entire United States.

1.2 Background and related work

This study followed an increasing body of work about biases in who and what is represented in social media data. The first work with Twitter data was by Mislove, Lehmann, et al. (2011), who found an overrepresentation of populous counties and an underrepresentation specifically of the Midwest, an undersampling in counties in the southwest with large Hispanic populations, an undersampling in counties in the south and midwest with large Black populations, and an oversampling of counties associated with major cities with large White populations. However, these findings come from interpretations of distributions and county-level cartograms, rather than from statistical testing, and they rely on the user-defined 'location' field, which has been shown to have many inconsistencies (Graham et al., 2014; Hecht, Hong, et al., 2011). The present study is on the one hand deeper because I use the far higher resolution of block groups and carry out statistical tests, but on the other hand not as general because my findings apply only to characteristics of *geotag users* within the US population rather than to geotag users within the Twitter population, or to Twitter users within the US population. Also worth noting is that Twitter has undergone large changes since the data used by Mislove, Lehmann, et al. (2011), both in the governance and management of the platform itself (van Dijck, 2013) and in patterns of user behavior (Liu et al., 2014). Sloan et al. (2013) followed up the work by Mislove, Lehmann, et al. (2011) by building a large-scale system for demographics inference in order to make social media data more usable for further sociological research, although they did not look specifically at biases.

More recently, Hecht and Stephens (2014) investigated urban biases across the US, a topic previously investigated on Foursquare by Ishida (2012). Following Goodchild (2007), Hecht and Stephens (2014) adopt the term *Volunteered Geographic Information* (VGI) for this type of data. Collecting 56.7m tweets from 1.6m users over a 25-day period in August and September 2013 and comparing it to Census data, they use a method of calculating a reduced effective sample size in order to correct for spatial dependencies. From this they calculate ratios of users per capita and find a bias towards urban areas, with 5.3 times more geotagged tweets per capita in urban regions as in rural ones, a magnitude even more pronounced in Foursquare data. Longley et al. (2015) investigate biases across a number of factors, focusing on the Greater London area. Using work on forename-surname pairs identifying gender, age and ethnicity, they parse usernames and other profile information to get a collection of estimated names, which they then compare to the 2011 UK Census and find an overrepresentation of young males, an underrepresentation of middle-aged and older females, an overrepresentation of White British users, and underrepresentation of South Asian, West Indian, and Chinese users, although tests of significance are not applied. Theoretically, Blank (2016) makes a similar point, that uneven demographics has implications for what signals are present in Twitter data.

Shelton et al. (2014) carry out a smaller-area case study of geotagged tweets, and do not use statistical modeling, but dramatically illustrate potential harms from relying on biased geotagged tweets. Looking at tweets about Hurricane Sandy in the New York area, they showed that the areas with the most severe disaster relief needs were not necessarily the areas that had the most tweets. Thus, they conclude, a naïve approach of using tweet frequency for directing relief efforts would have disadvantaged people in certain outlying areas, and focused on areas potentially with many complaints but with less dire needs.

Coming from another methodological direction, a nationally representative survey study of smartphone owners ($n = 1,178$) by Pew (Zickuhr, 2013) looks at the demographics of location service users. Overall, 12% of those surveyed reported using what Pew terms 'geosocial' services (which includes geotagged tweets, and excludes informational services like Google Maps). Interestingly, the survey finds the most frequent users of geosocial services are those of low*est* income and middle income; those of low*er* income use it less, and those of upper income use it least. More 18-26 year olds use geosocial services than older users, and almost double the proportion of hispanic smartphone owners (both English- and Spanish-speaking) use geosocial services as compared to non-hispanic white and non-hispanic black smartphone owners. However, out of the respondents who specified which geosocial services they use (n=141), most reported using Facebook (39%), Foursquare (18%) or Google Plus (14%); only 1%, or 1 respondent, used Twitter's geosocial services (i.e., geotagged tweets), such that it is not possible to make inferences about geotag users from the results of this study.

Our paper answered the general call for stronger methodological investigations about the nature of population representation in social media data (Ruths and Pfeffer, 2014; Tufekci, 2014), as well as the specific call for combining geographic data from user-generated sources with non-user-generated sources, such as Twitter data with the Census (Crampton et al., 2013).

1.2.1 Ecological inference

One major limitation of this work that I realized only after publication is the problem of *ecological inference*, inferring individual behavior from group-level data. A canonical illustration given by King et al. (2004) is if we have the number of blacks and whites in voting districts, and we have the number of people in each district who voted, given enough districts can we determine the conditional probabilities of whites voting and blacks voting and not voting? Surprisingly, the answer in general is no; the marginals clearly give bounds on the conditional probabilities, but this turns out to in general not be enough to get the desired point estimates. As O'Loughlin (2000) points out, geographers have tended to skirt the problem of ecological inference by talking about properties of *areas* rather than of individuals within those areas. Since over- or under-representing areas associated respectively with dominant or marginalized demographics may effectively produce the same outcomes as over or under-representing individuals of those demographics (i.e., misrepresentation happening through a mediator of geography), and since ecological inference is far from a solved problem (Freedman, Klein, et al., 2009a; Freedman, Klein, et al., 2009b), I take this approach: my results are about properties of *areas* we can predict to be over/underrepresented from using geotagged tweets.

1.3 Method

1.3.1 Data collection

Geo-Coded Twitter Data. From Twitter's Streaming API, we collected 144,877,685 tweets from April 1 to July 1, 2013 using the geographic boundary box $[124.7625, 66.9326]W \times [24.5210, 49.3845]N$. This covers the contiguous US (i.e., the 48 adjoining US states and Washington DC but not Alaska, Hawaii, or offshore US territories and possessions). Consequently, all our tweets are geo-coded with lat/long GPS coordinates. As Morstatter, Pfeffer, Liu, and Carley (2013) report from the Twitter Firehose, about 1.4% of tweets are geotagged; and elsewhere (Morstatter, Pfeffer, and Liu, 2014) they report the Streaming API is more likely to be biased when the response to a query exceeds 1% of the total volume of tweets. Given also that North America accounted for only 22.32% of geotagged tweets in their collection, a fraction consistent with what Liu et al. (2014) report finding in a collection of decahose data covering the time period I consider, it is reasonable to assume that the use of the Twitter API to collect tweets geotagged in the US covers all or nearly all of geotagged tweets within the given time frame and geographic bounds. Similarly, in the 1% sample, Sloan et al. (2013) found 0.85% of the tweets worldwide being geotagged, also less than 1%.

Since the distribution of geotagged tweets over geotag users is characteristically long-tailed (fig. 1.5), with a minority of users sending out the majority of tweets, I decided that the relevant quantity was the number of geotag users rather than the number of tweets. I identified 2,612,876 unique user accounts in our data, which is the basis of my analysis.

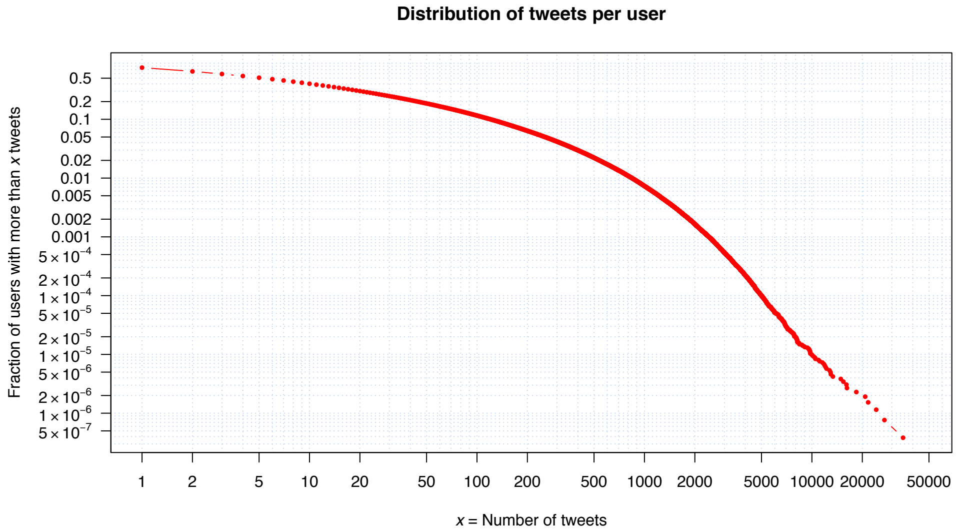


FIGURE 1.5: A long-tailed distribution of the number of users who have tweeted a certain number of tweets, plotted as a survival function (complementary cumulative distribution). Because of this skew, I focus on unique users alone, and ignore the volume of tweets.

Geospatial Data. Each block group has a unique identifier, the 12 digit *FIPS Code*, consisting of identifiers for state (first two digits), county (next three digits), tract (next six digits), and block group (last digit).[3]

The contiguous US plus Washington DC include 215,798 block groups[4] (2010 specification) which range in size from .002 square miles to 7503.21 square miles. Block groups are designed by the Census Bureau to have roughly comparable population sizes. I verified this by noting that, in log scale, the distribution of populations per block group has a symmetric distribution and stable variance (fig. 1.6).

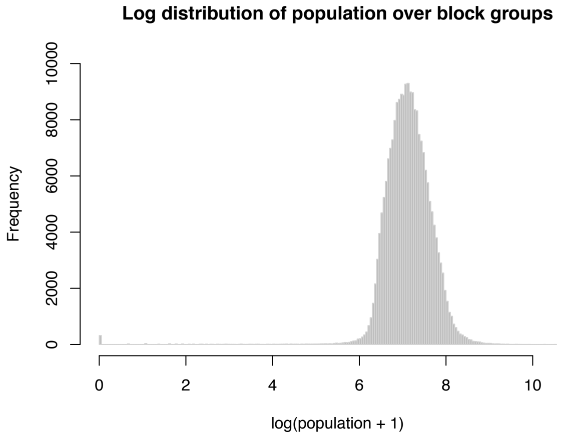


FIGURE 1.6: The Census Bureau designs block groups to enclose population sizes that are comparable. However, it does also allow for block groups with zero population, which is the zero-inflation visible after the add-one smoothing of log(population+1).

For every state, the US Census Bureau provides geographic boundary files ('shapefiles') that includes the GPS coordinates of the borders of every block group within the state. I combined the shapefiles of the 48 contiguous states and the District of Columbia, deleting 364 block groups representing bodies of water (identifiable by being coded as having zero area, and having a FIPS code ending in zero[5]). With Python code (utilizing the `shapely` package) we identified the Census block group into which each tweet fell.

I found 364 block groups with zero area; these also had zero population, and their FIPS codes all ended with 0. These turn out to correspond to bodies of water. While not all have zero geotag users tweeting from within them (for example there are 1,821 users who tweeted from an area of the East River bounded on

[3] Specifically: I use FIPS state codes 01 (Alabama) through 56 (Wyoming), excluding 02 (Alaska) and 15 (Hawaii). The FIPS specification skips 03, 07, 14, 43 and 52 (codes previously allocated for American territories, now depreciated). The District of Columbia is included in the sequence, with FIPS code 11.

[4] Probably due to a rounding error in geographic calculations, I lost three small island block groups (2 in Florida, 1 in New York), such that my n = 215,795.

[5] "Geographic Terms and Concepts - Block Groups", 6 December 2012, United States Census Bureau, `https://www.census.gov/geo/reference/gtc/gtc_bg.html`, accessed 3/2015. Note that this page references block groups "beginning with zero", but since the 'block group' part of a FIPS code is only the last digit, this should be interpreted as, "FIPS codes ending in zero.'

one side by the Brooklyn Bridge), for comparison with (potentially) populated areas I removed these water block groups (tab. 1.1).

FIPS code	Users	Description
06 083 990000 0	2,526	Channel Islands, CA
36 061 002500 0	1,821	Brooklyn Bridge, NY
51 810 990100 0	1,643	Coast off Virginia Beach, VA
36 061 009900 0	1,629	Chelsea Piers, NY
24 003 990000 0	1,373	Coast off Annapolis, MD

Table 1.1: Most popular bodies of water for tweeting from.

Socioeconomic Data. While the ideal would be to have rich and timely demographic data about the users who sent the tweets in our data (as attempted in Sloan et al., 2013), this was not realistic to collect for 2.6m users. But by aggregating data at the level of block groups, I can link Twitter data to the enormously rich demographic data the Census Bureau makes available at this level. I primarily use data from the 2010 Decennial Census, which I supplement with median income (not available in the Decennial Census) estimates from the 2009-2013 American Community Survey. For this ACS data, there were 1,224 block groups with missing values for median income, few enough that I filled these out as zeros rather than using imputation or smoothing. I also set 21 block groups with the value "2,500-" to 2,500, and 2,651 block groups with the value "250,000+" to 250,000. The 2009-2013 ACS had 54 block groups in the contiguous US whose boundaries (and FIPS) codes were from the 2000 Census, for which I found equivalent block groups in the 2010 Decennial Census to which to map. While the ACS 1-year estimates are more timely, they are more sparse and only at the county level (U.S. Census Bureau, 2008), and I decided to prioritize the accuracy and completeness of values in the Decennial Census for this analysis. I similarly decided to not use the ACS 2009-2013 estimates for population quantities as there was more missing data, and there was high correlation between the 5-year estimates and 2010 Decennial Census figures across variables (generally around .95). Still, prioritizing timeliness over completeness, and looking at the county level with 2013 ACS 1-year estimates, may be the focus in future analysis.

The Census Bureau also makes estimates of the same quantities at 1-, 3-, and 5-year intervals through the American Community Survey, and there are estimates from 2013; however, 1-year ACS estimates only cover areas with populations over 65,000 and only at the level of counties (U.S. Census Bureau, 2008), only the 5-year estimates cover all population sizes and go to the block group level. The 5-year estimates were only slightly more contemporaneous and I found them to include far more missing data. I thus decided to prioritize resolution and completeness[6] over timeliness for the greater power, and because I assume that shifts in population would not be enough to change the basic dynamic between population and tweets. However, this is a testable assumption, and future work may wish to look at the county level in order to study geotagged tweets with more timely demographic estimates.

[6] "American Community Survey: When to use 1-year, 3-year, or 5-year estimates", 23 March 2015, United States Census Bureau, http://www.census.gov/acs/www/guidance_for_data_users/estimates/, accessed 3/2015.

Mobile users. My construct of interest is the *number of potential geotag users*, for which population is the available proxy; there are cases where there are more geotag users than population, which points to tourists or, more generally, mobile users, as a complicating factor (Hecht and Stephens, 2014). I counted 18,835,284 distinct user-block group instances (i.e., if I were to use the number of unique users appearing within each block group, I would have inflated the user count by six times).

Hecht and Stephens (2014) provide a useful review of techniques to uniquely assign users to a single geographic region. They identify two candidate techniques: temporal, where a user must send at least two tweets a set number of days apart in a region for the user to be located uniquely in that region, and 'plurality rules,' where the most frequently tweeted-from region is taken as the unique location of the user. Checking the 'location' field fails because of the low quality of the information there (Hecht, Hong, et al., 2011). As one other option, Wang, Chen, et al. (2014) use the location of the first geotagged tweet sent by a user as the location of the user. This is the simplest, but also has no motivation beyond convenience.

Despite the drawbacks of plurality not accounting for people local to two regions, my comparison is with the US Census which also does not account for this possibility. However, another problem is that foreign tourists are not counted in the US Census (unlike domestic tourists, who reside in some US block group), and of which there were 70m in the US in 2013[7]. This is substantial when compared to the total 2013 US population of 316m[8] (of which 307m are counted in the block groups I use). If many foreign tourists send geotagged tweets, it would introduce unaddressed bias; since our data collection only had geotagged tweets in the US, short of massive additional data collection I am unable to identify foreign tourists (such as by looking at the proportion of geotagged tweets outside of the US). This is a potential problem in my analysis that may be a topic for clarification in future work.

Additionally, I filter users by the number of tweets, considering only those with a certain number of tweets.[9] As the distribution of tweets per user (fig. 1.5) is smooth and has no natural break point, I arbitrarily pick 5 and 10 as cutoffs to use alongside all users.

1.3.2 Statistical models

Random distribution over population. The basic relationship in which I am interested is between population and geotag users. In order to make a concrete test for random distribution, I suggest a model where there is a linear relationship between the population count and the number of users, i.e., users are drawn from the population at a constant rate subject to some noise. We can imagine the noise is heteroskedastic, which suggests the following data-generating process over population P, users U, and mean-zero noise term ε:

$$U = \alpha P + \varepsilon P \tag{1.1}$$

[7] "2013 Monthly Tourism Statistics: Table C - Section 1: Total Visitation, Canada, Mexico, Total Overseas, Western Europe Non-Resident Visitation to the U.S. By world region/country of residence 2013", n.d., http://travel.trade.gov/view/m-2013-I-001/table1.html, accessed 3/2015.

[8] "Population, total", 2015, The World Bank, http://data.worldbank.org/indicator/SP.POP.TOTL, accessed 3/2015.

[9] I thank an anonymous reviewer for this fruitful suggestion.

I transform both users and population to stabilize their variances, so this then becomes

$$\log U = \log \alpha + \log P + \log \left(1 + \frac{\varepsilon}{\alpha}\right) \tag{1.2}$$

Then, consider the linear model

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon' \tag{1.3}$$

If eqn. (1.1) described the true data-generating process, from eqn. (1.3) we should get that $\hat{\beta}_1 = 1$, and then $\exp(\hat{\beta}_0)$ would estimate the value of the proportion α. That is, the $\log \alpha$ term is the intercept of the regression of $\log P$ onto $\log U$, and $\log\left(1 + \frac{\varepsilon}{\alpha}\right)$ is a mean zero error term now independent of P, and we have a null hypothesis $H_0 : \beta_1 = 0$. While this may seem unrealistic as a null model, other quantities that we would believe are randomly distributed proportional to population indeed match this. For example, I regressed log population onto log males and found it to be meaningful (presented below under results). With this validation, I argue that the model of eqn. (1.1) is a reasonable way of representing a quantity being randomly distributed over the population. Note that my interest is not in fitting this specific model and interpreting the parameters, but just having a way to test the null hypothesis of random distribution. Note also that I originally sought to compare log population density to log geotag user density as a way of treating measures on different block groups as equivalent (given that block groups are already designed to somewhat control for the variance in population density), but found that it produced excellent fits that did not disappear when the data was shuffled, suggesting that the dividing by area created artifactual relationships.

Model specification For comparison with analyses of race and Hispanic populations (Mislove, Lehmann, et al., 2011; Zickuhr, 2013), I use Census variables[10] P0030001 through P0030008 and P0040001 through P0040003. For comparison with analyses by age (Longley et al., 2015; Zickuhr, 2013), I use P0120003 through P0120049 and aggregate across gender into the same age bins as in Zickuhr (2013). Existing analyses by sex (Longley et al., 2015; Zickuhr, 2013; Mislove, Lehmann, et al., 2011) is based on name-based inference or survey data; I decided that, while the Census does have sex data, the even distribution of sex across the US means that the sex ratio of a block group is not a meaningful proxy for geotag users who live there. For comparison with analyses of urban and rural populations (Hecht and Stephens, 2014; Zickuhr, 2013), I use P0020002 through P0020005.[11]

Thus, in total, I include terms for populations, the black population, the Asian population, the Hispanic/Latino population, the rural population, and respective populations of people ages 10-17, 18-29, 30-49, 50-64, and 65+. For all of these, I stabilize variance with a log transformation with add-one smoothing. I include median income (Zickuhr, 2013), and test for a northern/eastern effect by including the (demeaned) latitudes and longitudes of block group centroids, and for a coastal effect by including terms for latitude and longitude squared.

[10] "Census Data API: Variables in /data/2010/sf1/variables", 2010, http://api.census.gov/data/2010/sf1/variables.html, accessed 3/205.

[11] The Census API returned zero values for these, so I manually downloaded the variables of "P2. URBAN AND RURAL" for each state individually from factfinder.census.gov.

Spatial autocorrelation. Discretization into uneven geographic units (as block groups certainly are) can cause statistical artifacts. Specifically, if the divisions do not correspond to the contours of the underlying spatial process (and there is little reason to believe they would), there will be dependencies between proximate geographic areas, and not accounting for this can inflate the R^2 statistic, shrink standard errors, and give misleadingly significant results. I use the standard statistic for measuring spatial autocorrelation, Moran's I,

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \overline{X})(X_j - \overline{X})}{\sum_i (X_i - \overline{X})^2} \tag{1.4}$$

This is the empirical covariance, appropriately normalized, of the values of variable X between geographic units i and j. $W = [w_{ij}]$ is an $n \times n$ matrix of weights, discussed below. Rather than exploring autocorrelation in individual variables, I look for spatial autocorrelation in the residuals of a linear model (Anselin and Rey, 1991). For management of spatial data and implementation of computation and estimation for spatial models, I used the R package `spdep` (Bivand and Piras, 2015; Bivand, Hauke, et al., 2013).

Moran's I is well-investigated in terms of its asymptotic and theoretical properties (Gaetan and Guyon, 2012). It is tested under a null hypothesis of zero autocorrelation, either using assumed normality along with analytic forms of the higher moments of the statistics under normality or else permutation testing, which requires no distributional assumptions and which may be approximated by MCMC methods (Gaetan and Guyon, 2012). As I found that my variables and residuals were approximately normally distributed, I used tests based on asymptotic normality, for which the higher moments have analytics forms, rather than MCMC methods that make no distributional assumptions (Gaetan and Guyon, 2012) as the number of block groups made permutation testing computationally expensive. Fortunately, most of my variables had symmetric distributions with stable variance in the log scale.

Spatial autocorrelation is not inevitable, and indeed evidence of spatial autocorrelation may be due to model specification that can be eliminated by adding additional controls (Bivand, Pebesma, et al., 2013); alternatively, if spatial autocorrelation is not a quantity of interest, including it in a regression is itself a control. While we may test for spatial autocorrelation in the variable of interest if spatial dependencies are of explicit interest, a way more appropriate to my bivariate model is to look for spatial autocorrelation in the residuals of a linear model (Anselin and Rey, 1991). For management of spatial data and implementation of computation and estimation for spatial models, I used the R package `spdep` (Bivand and Piras, 2015; Bivand, Hauke, et al., 2013). I have found little work applying spatial models developed in econometrics, epidemiology and ecology to geographically dispersed social media data (an exception is Sylvester et al., 2014)), and hoped to bring such models to wider attention as thematically well-suited for analyzing issues of bias and representation (although, since the publication of this article, I have not seen this happen).

Weights matrix. Measuring spatial autocorrelation requires a 'weights matrix' of adjacencies between geographic units. There are multiple ways to generate this, and the choice of how to do so represents a substantive decision based on the problem at hand (Gaetan and Guyon, 2012). However, given that we do not know in advance the form of the spatial autocorrelation, in practice we can test for autocorrelation over different choices of weights matrices to see which is most appropriate (Anselin, Sridharan, et al., 2007). Thus, I consider the following weights matrices:

- Queen contiguity (regions sharing a corner or edge are adjacent, equivalent to 8-connectivity in image processing);
- Rook contiguity (regions sharing an edge are adjacent, equivalent to 4-connectivity in image processing)
- k-nearest-neighbors for $k = \{2, 3, 4, 5, 6, 7, 8\}$, calculated from the midpoints of block groups.

For the contiguity cases, I consider both row-normalized (which normalizes the 'effect' of each neighboring unit such that they sum to one) and binary (which gives greater possibility for autocorrelation between a unit and its neighbors for units with more neighbors). In the row-normalized case, I also employ Lagrange Multiplier tests developed in that contest (Anselin, 2002).

Spatial errors model. I model the relationship between population and geotag users using a Simultaneous Autoregressive (SAR) model, which is where one or more terms in the regression are correlated with itself. The main autoregressive model assumes that the residuals of unit i are correlated with the residuals of those units j adjacent to i, which is known in econometrics literature as a spatial errors model. The adjacencies are indexed exactly by the terms of the weights matrix. This gives the following two equations,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{1.5}$$

$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \varepsilon \tag{1.6}$$

where u are the correlated residuals, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ are the uncorrelated error terms, and the coefficient λ is the 'spatial multiplier' that captures the strength of the spatial autocorrelation (Anselin, 2002). We can rewrite these in a single form as either

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \lambda\mathbf{W})^{-1}\varepsilon \tag{1.7}$$

or, substituting eqn. (1.5) back into eqn. (1.6),

$$\mathbf{Y} - \lambda\mathbf{W}\mathbf{Y} = \mathbf{X}\beta - \lambda\mathbf{W}\mathbf{X}\beta + \epsilon \tag{1.8}$$

The terms $\lambda\mathbf{W}\mathbf{Y}$ and $\lambda\mathbf{W}\mathbf{X}\beta$ are known as spatial lags. While there are other SAR models, I use spatial errors as the simplest to interpret and the most appropriate for my purpose.

A spatial errors model lags the explanatory and response variables by the same multiplier. Other SAR models use lags differently; a different coefficient on the spatial lag for $\mathbf{Y}$ and the spatial lag for $\mathbf{X}\beta$ yields a spatial Durbin model, $Y = \rho\mathbf{W}Y + \mathbf{X}\beta + \mathbf{X}\beta\gamma + \mathbf{u}$, and if we only include a spatial lag on $\mathbf{Y}$, it becomes a spatial lag model, $Y = \rho\mathbf{W}Y + \mathbf{X}\beta + \mathbf{u}$. Estimation of the models results in different numerical issues, with the spatial errors model being the most straightforward to compute and to interpret (Bivand, Pebesma, et al., 2013), and the most appropriate as I only seek to account for spatial autocorrelation and not necessarily to measure it.

I originally sought to compare log population density to log geotag user density as a way of treating measures on different block groups as equivalent (given that block groups are already designed to somewhat control for the variance in population density), which I found generated extremely good fits and extremely significant coefficients. However, when I shuffled the data to break the relationship (I both tried shuffling densities, and shuffling counts and dividing them by shuffled areas), the estimated coefficients had the same values, and the R^2 remained the same, suggesting that the model fit in the case of density is an artifact of how transformation combines the underlying densities. In contrast, for my current model, I found a shuffle test broke the significance of the slope term, which is what should happen (in which case, the estimated intercept becomes the logarithm of the mean of the response).

Zero values. Zero values frequently cause problems, especially when transforming to log scale. I considered removing all block groups with zero population, and all block groups with zero geotag users, as these required padding that caused some data artifacts (visible in plots below). However, I found that excluding them only improved measures of model fit, such that including them (via add-one smoothing) leads to a more conservative estimate.

1.4 Results and discussion

1.4.1 Observational results

The block groups with the highest number of distinct users (before users are assigned uniquely) are major international airports and major tourist attractions (table 1.2).[12] The inclusion of several international airports on the list suggests that geotagging tweets during the process of travel is a common user behavior. There were some areas with zero population but nonzero users; out of these, the ones with the highest counts of distinct users are mostly the same: major airports and parks.[13]

Conversely, there were only 67 block groups from which nobody sent geotagged tweets; only 30 of these also had no population (these were national forests, minor airports, areas off highways, etc.). Of those that did have a population, the most populous was a block group with a population of 4,854 within San Quentin State Prison in California. The second-most populous block group is also a Corrections Department building in Texas, and third is a state prison in California (although not all prisons lack geotag tweet users; the block group of Rikers Island in New York has geotagged tweets from 22 users).

Out of the 2,612,876 unique users I identified, 2,216,219 (84.82%) had a single block group from which they tweeted most frequently. The others had ties for which block group was the highest; for these users, I uniquely assigned them to one of their block groups by randomization. I tried analyses on just the 84.82% as well, but found it made little substantive difference in the results.

[12] Block groups may be looked up by their FIPS code at `http://www.policymap.com/maps`.

[13] Interestingly, Central Park has a nonzero population (of 25), as do some airports. Some other tourist attractions (e.g., Universal Studios) also appear.

TABLE 1.2: Block groups from which the most users have sent geotagged tweets.

FIPS code	Users	Description
32 003 006700 1	28,280	Las Vegas Strip
06 037 980028 1	23,100	Los Angeles Int'l Airport
32 003 006800 4	16,748	McCarran Int'l Airport
13 063 980000 1	15,481	Atlanta Int'l Airport
12 095 017103 2	15,392	Walt Disney World
36 081 071600 1	15,067	JFK Int'l Airport
11 001 006202 1	14,906	National Mall
36 061 014300 1	14,605	Central Park
06 059 980000 1	14,576	Disneyland
17 031 980000 1	13,610	Chicago Int'l Airport

In the terminology of Guo and Chen (2014), the most active accounts belong to 'non-personal users.'[14] In this case, the most active tweeter (44,624 tweets) seems to be a commercial service for travel, the second-most active (35,025) is an automatic news updater in Florida, etc. Starting from the 13th most active tweeter, with 12,922 tweets, there were accounts that appeared on inspection to be personal ones. As for number of block groups traversed, the top 'traveler' (23,547 block groups) is the same as the top tweeter, and others are similarly non-personal users. Across block groups, it is not until the 18th most mobile user, traversing 1,209 block groups, that there is a personal user.

How much mobility is there between units? Figures 1.7 and 1.8 show respectively that while there is minimal mobility between states, with only 22.39% of users sending geotagged tweets from more than one state and only 7.83% send from more than 2. However, there is a great deal of mobility between (possibly neighboring) block groups, with 65.24% of users sending geotagged tweets from more than one block group.

How well does unique assignment do? As one check, I consider the ratio of geotag users to population; there are 509 block groups where this ratio is greater than 1 (for users with 5 or more tweets only, there are 353, and for users with 10 or more tweets only, there are 290), indicating either the failure of population as proxy for potential geotag users or of the method of assigning mobile users. As I found the block groups with the largest ratios to be airports, it seems to be a case of the latter.

The largest ratio is in the block group containing Los Angeles International Airport, 1365.5 to 1 (558.25 to 1 and 287.25 to 1 for the two respective filter levels). The second-highest ratio is the block group in Manhattan containing Bryant Park, and the remainder of the top five are more major airports. This points to the method of unique assignment unsuccessfully handling tourist destinations even with filtering for a minimum number of tweets.

[14] They find that only 2.6% of geotag users are non-personal. This should be small enough to have no effect on results, so I did not employ filtering. However, this may be considered in a future work.

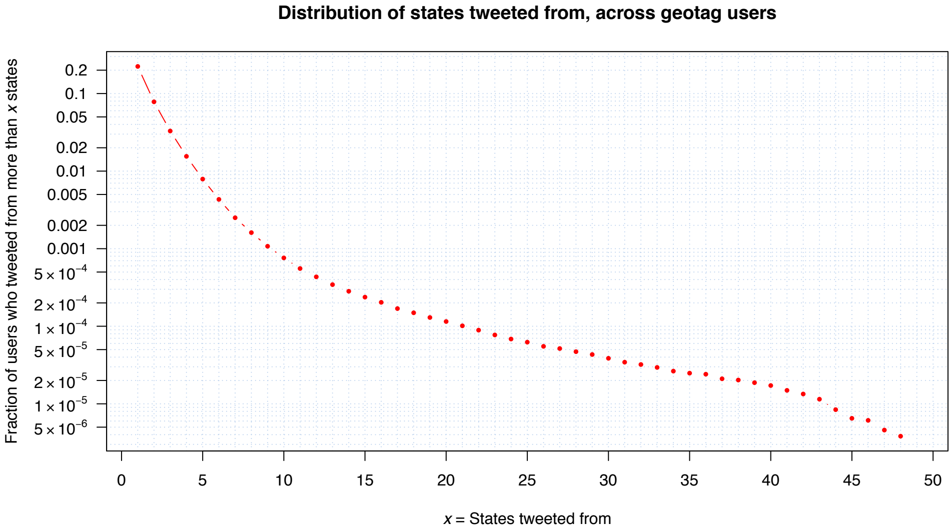


FIGURE 1.7: A full 77.61% of geotag users in our set tweeted only from one state, and having tweeted from 5 or fewer states accounts for 99.21% of users.

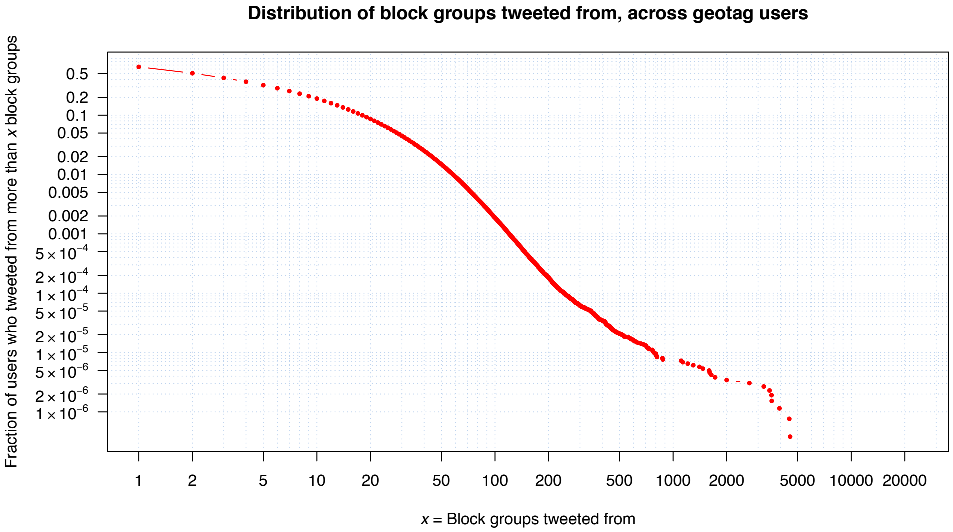


FIGURE 1.8: 34.76% of geotag users tweeted only from one block group. 27 or fewer block groups were 95%, 50 or fewer block groups were 99%. One outlier at 23,547 excluded.

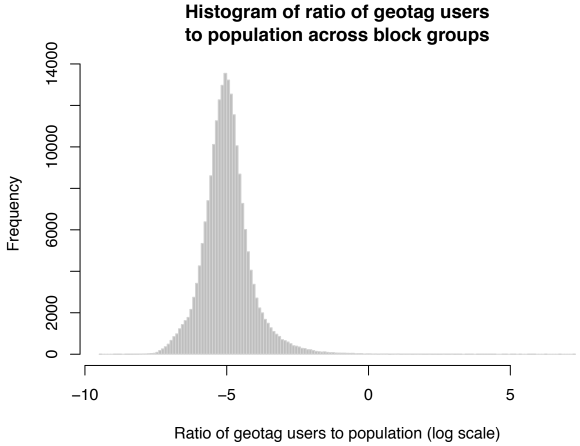


FIGURE 1.9: Ratios above zero are obvious failures of the metric, as the number of uniquely assigned geotag users should not exceed the population, but the distribution is smooth and symmetric.

1.4.2 Bivariate regression model

I first test my null hypothesis of a linear regression yielding a coefficient of 1 to the logarithm of the population. Looking at the plot of the relationship of the logarithm of the two (fig. 1.11), there is a faint linear relationship, although the slope does not appear to be 1. An OLS regression fits slope $\hat{\beta}_1$ = .4916 (.002996) and intercept $\hat{\beta}_0$ = -1.219 (.02143),[15] although recall that the standard errors are not reliable under spatial autocorrelation.

Compare this plot to the plot of the test case mentioned earlier, the distribution of males over the population, pictured in fig. (1.10). The true ratio of males to total population across the block groups we consider is .4915; according to my model, the exponential of the intercept should be this, and the coefficient of the log population term should be 1. Indeed, log(.4915) is within the 95% confidence interval (log(.4914), log(.4962)), and 1 is just outside the 95% confidence interval (.9980, .9994), but this is without accounting for how spatial autocorrelation shrinks estimated standard errors. The R^2 value of this model is also impressive at .975, although under spatial autocorrelation R^2 is inflated thereby not interpretable. Overall, my model fits the relationship of males to population exactly as we would expect it to fit to something randomly distributed over the population.

Using this as a validation of my statistical test, we can strongly reject the null hypothesis that $\hat{\beta}_1$ = 1 even without correcting for spatial autocorrelation. And the R^2 value for this regression is a paltry .109, too small to worry about being inflated. Thus, we can conclude that geotag users are not randomly distributed over the US population, and indeed that the population count is not very informative about the number of geotag users.

[15] Filtering for only those users who have 5 or more tweets and for those users with 10 or more tweets, the respective fitted slopes are .5192 (.002932) and .5136 (.002786).

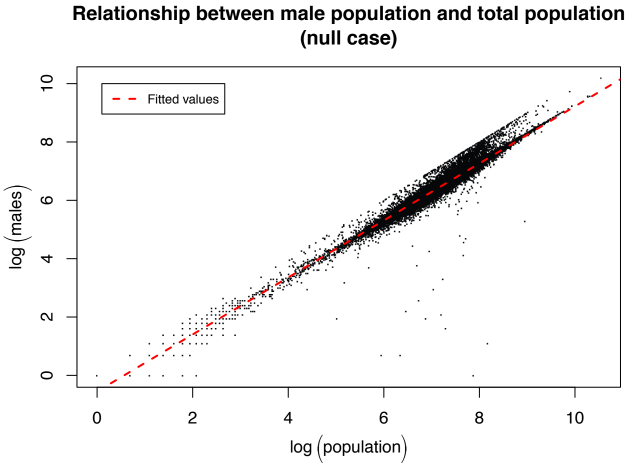


FIGURE 1.10: The relationship between males and total population behaves exactly as we expected of a quantity randomly distributed over the population, making it an effective null model against which to compare the observed distribution of geotag users.

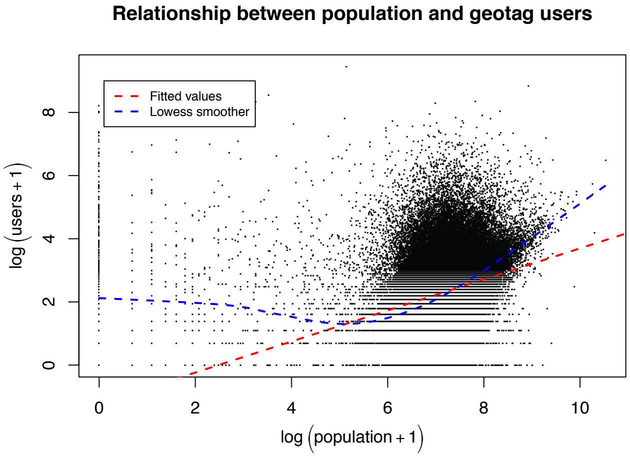


FIGURE 1.11: Eliminating zero-count observations reduces the artifacts visible at $x = 0$ and $y = 0$ but does not substantially change the fit.

1.4.3 Weights matrix and spatial autocorrelation

Testing the residuals in my basic model for spatial autocorrelation using Moran's I against all weights matrices considered above, I find the results reported in table (1.3).

I found identical results of Moran's I for binary weights matrices and row-normalized weights matrices in the k-nearest neighbor case. For the two contiguity cases, row normalization made a difference, and I list both values. In all cases, an asymptotic test against the expected value of 0 was significant at $p < .0001$. The

TABLE 1.3: Selected Values of Moran's I in residuals

	Population vs Users	Population vs Male
2nn	.3699	.2336
4nn	.3550	.2142
6nn	.3398	.1996
8nn	.3270	.1883
Rook	.4166 (b) .3992 (rn)	.2125 (b) .2201 (rn)
Queen	.4151 (b) .3919 (rn)	.2097 (b) .2154 (rn)

For the Rook contiguity case and the Queen contiguity case, binary (b) and row-normalized (rw) weights gave different values.

autocorrelation in the population-user model is stronger than in the 'null' population-male model. It appears, then, that the spatial autocorrelation is strong enough that the choice of weights matrix is not critical. For the population to user model fit on counts of users with 5 or more tweets, or 10 or more tweets, the spatial autocorrelation was similar (generally lower, but still higher than the autocorrelation of population vs. male).

1.4.4 Spatial errors model

The maximum likelihood method of fitting a SAR model involves computing the log determinant of the $n \times n$ matrix $|I - \lambda W|$, which is infeasible at my n of over 200,000. An alternative method finds the log determinant of a Cholesky decomposition of $(I - \lambda W)$, although this then requires W to be a symmetric matrix (Bivand, Pebesma, et al., 2013). Since all of the candidate weights matrices picked up spatial autocorrelation at a significant level, I use a binary contiguity weights matrix. I tried both Rook and Queen, and they gave comparable fits, so I report only for Rook (1.4).

The spatial multiplier term is significant, although neither the coefficients nor the standard errors are substantively different than the previous model. However, calculating Moran's I on the residuals of this model gives a value of -.02367, with a p-value of 1, meaning we have successfully controlled for spatial autocorrelation.

I then investigate the full model specified above. I interpret this model in the standard way: for a log transformed explanatory variables X_i, a 1 percent change is associated with a β_i percent change in $\mathbf{Y}$. I present the results of the regression on counts of only those users with 5 or more tweets. This is shown in table (1.5).

As before, testing for spatial autocorrelation finds no significant amount, with a p-value of 1.

I considered using the youngest ages (ages 0-9) as the omitted category in order to accord with how the Pew study Zickuhr, 2013 does not cover usage by children. However, it is more appropriate to exclude ages 18-29, as it theoretically may be considered the baseline category. Furthermore, Pew does have an earlier

TABLE 1.4: Spatial errors basic model, binary Rook contiguity

	Dependent variable:
	log(user + 1)
log(population + 1)	.4401*** (.002655)
Intercept	$-$1.138*** (.01890)
$\hat{\lambda}$:	.1107***
LR test value:	73,375
Numerical Hessian $\widehat{\mathsf{se}}(\hat{\lambda})$:	8.4241e$-$06
Log likelihood:	$-$222,020.8
ML residual variance (σ^2):	.4206
Observations:	215,795
Parameters:	4
AIC:	444,050
Note:	***p<.0001

study on geosocial service usage by children ages 12-17 Zickuhr, 2012, finding that teenagers and adults used geosocial services at the same rates, about 18% in 2012. Ideally we would aggregate the Census data into age bins of 0-11 and 12-17 to correspond to those of Pew; unfortunately this is impossible from Census data, as the Census provides counts for ages 0-4, 5-9, 10-14, and 15-17. Coding 0-9 and 10-17 is the closest we can get.

Controlling for other factors, population still has a significant, positive, and large effect.[16] The hypothesis test I built, by which I rejected a random distribution over the population, is still valid; the revision is about population having an effect versus not having an effect, but either way it is not the only effect. Formally, adding in other factors indeed improve the model fit: running a bivariate spatial errors model with > 5 users against only population, I get an AIC of 444,000 (versus 423,530), and a likelihood ratio test of the bivariate (i.e., restricted) model against the full model rejects at the p<.0001 level the null that the restricted model is correct.

The term for area included as a control is significant, with a one percent rise in block group area associated with a 16.56% rise in geotag users. It seems here that size overcomes the effects of population density (as mentioned above, block group population has stable variance only in log scale even though block groups are designed to enclose populations of roughly comparable size). Consistent with survey findings (Zickuhr, 2013), a 1% larger Hispanic/Latino population is associated with 3.78% more geotag users. However, the

[16] In the original paper, which used ages 0-9 as the reference category, I had found that population only lost its significance for users with 5 or more geotagged tweets; I theorized that this was an appropriate cutoff (to exclude users who only tried geotags and to not include only power users) and thus privileged the model outputs for this dependent variable. However, with this result, and generally how population is significant and large across different variations of the model, I revise my previous conclusion. Also notable is that now the effect of median income is no longer significant. Given that its effect size was weak before, this is not too much of a change, but to see no significant effect (not even a weak effect) is still surprising.

TABLE 1.5: Spatial errors full model with ages 18-29 as the omitted category, binary Rook contiguity, users with >5 tweets only. Revised from published version.

	Dependent variable:	
	log(user + 1)	s.e.
log(population + 1)	.4277***	(.006479)
log(area)	.1656***	(.001809)
log(asian + 1)	.1249***	(.001603)
log(black + 1)	.06130***	(.001483)
log(hispanic + 1)	.03787***	(.002112)
latitude (demeaned)	.02522**	(.007352)
longitude (demeaned)	.01962***	(.002864)
latitude^2	-.0003490**	(.00009910)
longitude^2	.00006872***	(.00001475)
median income ($10K)	.001234	(.00068035)
log(rural + 1)	-.05791***	(.001119)
log(ages 00-09 + 1)	-.01104	(.005509)
log(ages 10-17 + 1)	-.05442***	(.005389)
log(ages 30-49 + 1)	-.05466***	(.007479)
log(ages 50-64 + 1)	-.1793***	(.007126)
log(ages 65 and up + 1)	-.2585	(.003857)
Intercept	.1497	(.1998)
$\hat{\lambda}$:	.1039***	
LR test value:	39,934	
Num. Hessian $\widehat{\text{se}}(\hat{\lambda})$:	.0003735	
Log likelihood:	-211,745	
ML resid. var. (σ^2):	.3871	
Observations:	215,795	
Parameters:	19	
AIC:	423,530	
Note:	**p<.001; ***p<.0001	

effect size is smaller than either that of the Asian population (a 1% rise is associated with a 12.49% rise in geotag users) and, in contrast to survey findings, that of the Black population (a 1% rise is associated with 6.13% rise in geotag users). This might point to the Pew sample not including enough Twitter users, as there is an active Black community on Twitter that had gained scholarly attention even when this article was published (Clark, 2014; Florini, 2014; Sharma, 2013).

The latitude, both in linear and quadratic terms, is significant at the p<.001 level. Thus, after controlling for population size and longitude, being further north or towards the mean latitude of the US is associated with more geotag users. While the effect size of latitude is larger than those of longitude, so are the standard errors (hence being significant at a lower level), hence the true effect of latitude is not necessarily stronger.

While I tried to test for nonlinearity in income, inclusion of a squared term for median income made the matrix computationally singular; however, inspecting the bivariate relationship did not yield any evidence for a nonlinear effect, and the linear effect is weak and nonsignificant (a $10,000 rise in the median income is associated with a 0.12% rise in the number of geotag users).

Consistent with findings about urban biases (Hecht and Stephens, 2014), I find that a 1% higher rural population is associated with a 5.79% decrease in the number of geotag users.

There is a negative effect from having a higher population of any other age group except for ages 30-49, which surprisingly is associated with slightly more geotag users.[17] Under this choice of omitted category, the size of the population of ages 65 and up and of the population of ages 0-9 are no longer significant. In contrast to Zickuhr (2012), I find that there is a significant difference between the number of teenage users and the number of adult users.[18] However, the different age bins make these results not exactly comparable, since it is certainly possible that children of ages 10-11 use geotags at a far lower rate than those of ages 12-17, dragging down the mean of a category that combines the two groups. It is surprising that the negative effect from population of ages 65 and up is not significant; as I speculated before, this might be due to mixed populations, for example places that are popular for retirement also being popular for tourism.

As is usual with logarithmic dependent variables, the intercept is not particularly interpretable as it would be a prediction for a block group at the center of the US with a population of 1.

Running the SAR model using all users, instead of just those with 5 or more tweets, produces similar results, except that log population is significant with coefficient -.04196 (.007858); this suggests a nonlinear effect, and indeed, an added squared term for the log population came out as significant and positive at .06329 (.0008394). This points to some noise for those people who only ‘try out’ geotagged tweets but do not adopt their use that disappears if we maintain a minimum tweet threshold. When running the model on only those users with 10 or more tweets, results are again similar except the longitude squared term is no longer significant (p = 0.1870), and the latitude term becomes significant (p = 0.02017). This might be from the coasts having more users who try out geotagged tweets for a longer period of time before choosing not to continue. These subtle differences point to opportunities for modeling the demographics of different types of users (as determined by number of geotagged tweets or other factors), although I do not explore them more here.

[17] This is inconsistent with how using ages 0-9 as the omitted category in the original paper gave a larger coefficient for 18-29 year olds than for 30-49 year olds, pointing to possible issues with model misspecification.

[18] Zickuhr (2012) compares 12-17 with 18 and up. To mimic this, we-coded ages into only three categories of ages 0-9, ages 10-17, and ages 18+, and re-ran the model using 18+ as the omitted category. The significance and direction of the coefficients for ages 0-9 and 10-17 were identical to the full model.

1.5 Conclusion

Geotag users are not representative of the US population. Despite the volume of geotagged tweets and their impressive coverage (there were only 67 block groups out of 215,795 with no geotagged tweets), the users who send geotagged tweets are nonrandomly distributed over the population in subtle ways. These include predicable and already established biases towards younger users, users of higher income, and users in urbanized areas, as well as surprising biases towards Hispanic/Latino users and Black users that, in the latter case, have not been seen in large-scale survey research. I also demonstrate an unsurprising but previously unreported coastal effect, where being located on the east or west coast of the US is associated with more geotag users. Geotag users may not be a random sample of the population of any given block group, but given the fine level of detail and large-scale demographic variability, the demographics of a block group is a reasonable proxy for the demographics of geotag users located in that block group. Certainly, even with complications of uniquely assigning mobile users, it is enough to establish the nonrandom distribution of geotag users, and some candidate biases.

While from this study, I am unable to say whether or not geotag users are representative of the *Twitter* population; they are a self-selecting group, and my analysis is further not able to say anything about why certain demographic profiles would be more likely to select in (or what other causal features there may be behind the decision of some people of a given demographic to use geotagged tweets but not others). But the interesting question that can be addressed with the given data is whether geotagged tweets can be a useful proxy for the *general* population within the US. This is a critical question because geotagged Tweets are an enormously popular source of data for studying a wide variety of social and human phenomena. For future work, I emphasize that findings using geotagged tweets should not be assumed to generalize, and conclusions should be restricted only to geotag users with their population biases.

Future Work There are a number of directions for future work. The most obvious is to update the data and models with more recent ACS estimates, and geotagged tweets collected in the same year. In terms of model terms, in cases where it is possible to measure differences in usage by gender, there are strong reasons to hypothesize that fewer women than men use geotagged tweets, based on the larger and more severe harassment received by women (Matias et al., 2015; Hess, 2014; Meyer and Cukier, 2006) and how abusers use knowledge of physical location to make explicit or implicit threats (Matias et al., 2015; Megarry, 2014). Indeed, one recommendation for targets of abuse is to turn off geolocation.[19] Furthermore, there are strong theoretical reasons to consider interaction effects between race and gender (Clark, 2014; Dixon, 2014).[20]

Other directions are to see the effect of filtering out non-personal users, and to build ways to filter out foreign tourists and better uniquely place geotag users in the block group that is likely to be their residence. Modeling demographic differences between users of different levels of use is also possible with this data. I have applied one spatial model, but spatial modeling is a rich area with many other available techniques. For example, there are also relevant disease mapping models that break down incidence by various demographic

[19] Recommendations for 'Social Media Safety' from the Rape, Abuse & Incest National Network, https://rainn.org/sexual-assault-prevention/social-media-safety.

[20] I thank Amanda Jean Stevenson (2014) for pointing this out to me.

strata (Bivand, Pebesma, et al., 2013) that would be appropriate here, as well as nonparametric models that might better capture irregular effects. Furthermore, I elected to not consider the temporal aspect; there is work on spatio-temporal modeling (Longley et al., 2015; Sylvester et al., 2014; Nagar et al., 2014; Kamath et al., 2013) but it tends to be in the short-term window of a day or week. With reliable spatio-temporal models of how the prevalence of geotagged tweets per block group changes over longer periods of time and a better understanding of the demographic characteristics towards which geotag users are biased, I may be able to create models to provide a rapid and high-resolution proxy for demographic changes such as processes of gentrification, or urbanization, or urban decay; that is, utilize the very biases of social media data to make inferences about larger phenomena. This was already done on a smaller scale, within the city of London, using a combination of Twitter and Foursquare data, by Hristova et al. (2016); they find correlations between properties of networks on those sites and measures of gentrification via the UK's Index of Multiple Deprivation. It may be possible to scale this up, making use of the geographic span of Twitter usage.

Chapter 2

Platform effects[1]

Summary. In this Chapter, I use the rapid introduction of Facebook's "People You May Know" as a natural experiment by which to observe the causal effect of a recommender system on user behavior. I theoretically frame this as an example of decisions of platform governance having a causal effect on user behavior, which has larger implications for how we think about the data we get from social media platforms.

Compared to the published version, I update all fits to nonparametric quantile fits (previously, I had used one high-order polynomial regression), and have added an extensive additional discussion of literature, including of examples of causal observational inference with social media data that I had previously missed (as well as one nearly simultaneous publication), as well as two important theoretical works I had missed.

2.1 Introduction

In social media data, the design and technical features of a given platform constrain, distort, and shape user behavior on that platform, which I call the *platform effects*. For those inside companies, knowing the effect a particular feature has on user behavior is as simple as conducting an A/B test (i.e., a randomized experiment), and indeed such testing is central to creating platforms that shape user behavior in desirable ways. But external researchers have no access to the propriety knowledge of these tests and their outcomes. This is a serious methodological concern when trying to generalize human behavior from social media data: in addition to multiple other concerns, observed behavior could be artifacts of platform design. This concern has thus far only been raised theoretically (Tufekci, 2014; Ruths and Pfeffer, 2014), and not yet addressed empirically. Even theoretically, the problem is deeper and more subtle than has been appreciated; it is not just a matter of non-embedded researchers having access to the data (Savage and Burrows, 2007; Lazer, Pentland, et al., 2009; Huberman, 2012; boyd and Crawford, 2012), but also that even when researchers have access, without full knowledge of the platform engineering and the decisions and internal research that

[1] This is an updated version of a paper previously published as: Momin M. Malik and Jürgen Pfeffer (2016b). "Identifying platform effects in social media data". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 241–249.

went into design decisions, the data can be systematically misleading. This topic relates also to discussions in the humanities about the nature of social media platforms as governance and management entities (van Dijck, 2013; Gehl, 2014), and how models have a self-reinforcing property of creating the very reality they purport to describe or explain (Healy, 2015).

One way to study and quantify platform effects as an external researcher is to look for available data that include a significant platform change. Making the assumption that, in absence of the exogenous shock (the change) the previous 'trend' would have remained the same, we can apply the observational inference method of *regression discontinuity design* (Imbens and Lemieux, 2008; Lee and Lemieux, 2010; Li, 2013). While not as certain as experimental design, observational inference methods are the best available way for outside researchers to understand the effects of platform design.

As another theoretical contribution which directly anticipates the call of Lazer and Radford (2017), I argue that *data artifacts*, rather than being incidental or annoyances to be corrected (Roggero, 2012), are a rare place where usual order breaks down, which can provide a glimpse into otherwise inaccessible underlying mechanisms. Here, data artifacts are providing important insights into inner working of platform engineering and management.

I select two data sets: the Facebook New Orleans data collected by Viswanath et al. (2009), and the Netflix Prize data, described by Koren (2009a). The latter is no longer publicly available since the close of the Netflix prize, although the terms of use do not mention any expiration on use for those who have already downloaded it.

In the Netflix Prize data set, Koren (2009a), a member of the team that ultimately won the prize (Koren, 2009b), points out a curious spike in the average ratings in early 2004. As such a change has modeling implications (previous data should be comparable in order to properly use for training purposes), he explores the possible reasons for this, ultimately identifying an undocumented platform effect as the most likely driver. Then, the Facebook New Orleans data contain an identified, and ideal, example of a platform effect: a clear exogenous shock and a dramatic difference after, through the introduction of the "People You May Know" (PYMK) feature on March 26, 2008. This discontinuity is only mentioned in Zignani et al. (2014); the original paper of the data collectors (Viswanath et al., 2009) does not mention it (although, in another example of a platform effect in collected data, they do note that on July 20, 2008, Facebook launched a new site design that allowed users to "more easily view wall posts through friend feeds" which they use to explain a spike in wall posts towards the end of the collected data).

In sum, I re-analyze the Netflix Prize and Facebook New Orleans data to study possible platform effects in the data. The contributions of this paper are:

- To empirically verify previously expressed theoretical concerns about the possible effects of platform design on the generalizability and external validity of substantive (social scientific) conclusions;
- To import into the social media research community a statistical model that allows quantitative estimation of platform effects;
- To quantify two specific cases of common platform effects, the effect on a social network of a triadic closure-based recommender system and the effect of response item wordings on user ratings.

2.2 Background and related work

Authors from multiple disciplines (Tufekci, 2014; Ruths and Pfeffer, 2014) have expressed methodological concerns that the processes found in data derived from social networking sites cannot be generalized beyond their specific platform. Most troublingly, the same things that would cause results to not generalize, such as nonrepresentative samples, idiosyncratic technical constraints on behavior, and partial or uneven data access, are generally unknown and undetectable to an outside researcher (and potentially even to engineers and embedded researchers). Some innovative methods of data comparison have been used to derive demographic information in social media data (Chang et al., 2010; Mislove, Lehmann, et al., 2011; Sloan et al., 2013; Hecht and Stephens, 2014; Longley et al., 2015; Malik, Lamba, et al., 2015) and to identify biases in public APIs (Morstatter, Pfeffer, Liu, and Carley, 2013; Morstatter, Pfeffer, and Liu, 2014), but platform effects remain empirically unaddressed. Part of the problem is that social media platforms are private companies that seek to shape user behavior towards desirable ends, and do so in competition with one another (van Dijck, 2013; Gehl, 2014); thus, the details of features and functionality which successfully guide user behavior are understandably proprietary in ways that representation and data filtering need not be. The results of research experiments, most notably Kramer et al. (2014), deal only indirectly with platform design and engineering. Outside accounting via testing inputs (Diakopoulos, 2014) is an important way of identifying overall effective outcomes, but such cross-sectional audits lack a baseline to know how much a given platform design successfully shapes behavior.

Instead, one way to study the problem is the econometrics approach of finding cases that can be treated as ‘natural experiments’ (Angrist and Pischke, 2008; Gelman, 2009). I have located two such instances, the Facebook New Orleans data and the Netflix Prize data, where known or suspected change in the platform led to a shift, documented in publicly available data.

Zignani et al. (2014) used the data of the Facebook New Orleans network (Viswanath et al., 2009), along with data from the Chinese social networking site Renren, to investigate the delay between when it is possible for an edge or triangle to form (respectively, when a node enters the network, and when two nodes are unconnected but share a neighbor) and when it actually forms, which they respectively term *link delay* and *triadic closure delay*. They note that on March 26, 2008, there is a drastic increase in the number of links and triangles (my version of those plots given in figs. 2.1 and 2.2), corresponding to the introduction of Facebook’s “People You May Know” (PYMK) functionality. While this was not the central investigation of their paper, they used it as an opportunity to see how an external feature changed their proposed metrics. They find that this increase consists primarily (60%) of links delayed by over 6 months, and also includes many (20%) links delayed by more than a year. They continue to note, “Although the link delay [metric] reveals interesting characteristic in edge creation process, it is not able to capture the reason behind it, i.e., which process causes the observed effects or which algorithms were active in the early rollout of the PYMK feature.” However, from their finding that far more triangles were created than edges (based on their fig. 2b, the ratio of new triangles to new edges rose from about 2 before the introduction to about 4 afterwards), it suggests that the created edges were based heavily on triadic closure. They conclude that the external introduction of PYMK manipulated a parameter or parameters of the underlying dynamic network formation process, and furthermore, it did not increase the link creation or triadic closure uniformly, but with bias towards more delayed links and triads. While they say they were able to quantify the effects and impact of the PYMK feature, this did not include estimating the local average treatment effect, which is my specific interest.

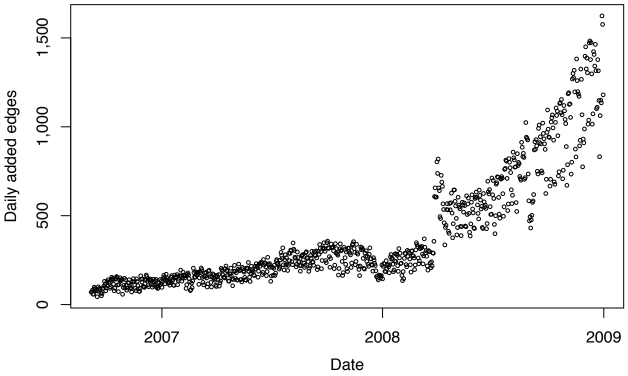


FIGURE 2.1: Observed edges added (friendship ties made) in Facebook New Orleans data.

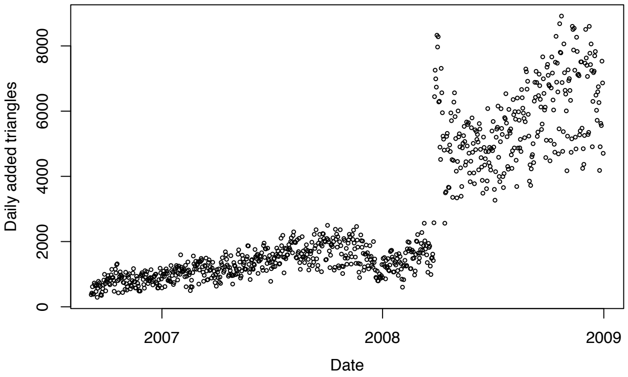


FIGURE 2.2: Triangles created with the added edges in Facebook New Orleans data.

The goal of the Netflix Prize competition was prediction and not explanation (Shmueli, 2010; Breiman, 2001), for which it is not necessary to understand the spike (only to account for it in a model, in order to effectively use past data for training). However, checking for data artifacts is fundamental for any type of data model, and Koren (2009a) devotes some time to investigating an odd spike observed in average ratings in early 2004, about 1500 days into the data set (this plot is recreated in my fig. 2.3). He proposes and explores three hypotheses:

1. Ongoing improvements in Netflix's 'Cinematch' recommendation technology and/or in the GUI led people to watch movies they liked more;

2. A change in the wordings associated with numerical ratings elicited different ratings (e.g., perhaps a rating of 5 was originally explained as "superb movie" and then was changed to "loved it");

3. There was an influx of new users who on average gave higher ratings.

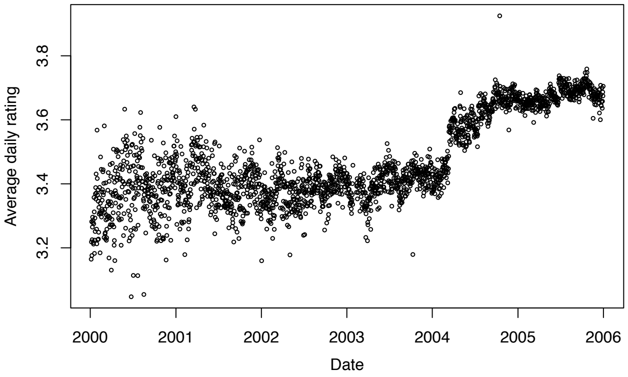


FIGURE 2.3: Observed daily averages for the Netflix Prize data.

By noting that the shift also occurs among users who were present both before and after the observed increase, he rejects the third possibility. He finds some support for the first possibility from a model that decomposes ratings into a baseline effect and a user-movie interaction effect (which corresponds to the extent to which users rate movies "suitable for their own tastes"); the interaction effect shows a smooth increase and the baseline has less variability, but there is is still clearly a sudden jump in the baseline. He writes, "This hints that beyond a constant improvement in matching people to movies they like, something else happened in early 2004 causing an overall shift in rating scale." Note that the change in wordings associated with numerical ratings is Koren's (2009a) guess to what the change was; he specifies that uncovering exactly what the "something else" was "may require extra information on the related circumstances." That such a change in wording *could* produce a shift in ratings is supported by decades of research in survey research into response options (Dillman et al., 2014), but otherwise no further evidence is given.

2.2.1 Causal modeling

Other works have been seeking out cases of natural experiments in social media data. Oktay et al. (2010) appears to be the first, discussing quasi-experimental designs and using Stack Overflow as an example setting. They demonstrate the use of interrupted time series by looking at whether users receiving an 'epic' badge, which is determined by hitting a daily reputation cap 50 times, decreases their daily posts; while only having 54 such examples to consider, it is enough to determine that getting this badge reduces the number of posts. In a case of applying regression discontinuity design, Li (2013) identified Yelp ratings being rounded to the nearest star as appropriate for RD design.

Sharma et al. (2015) provide a systematic statement of the problem of recommender systems: "little is known about how much activity [recommender] systems actually *cause* over and above activity that would have occurred via other means (e.g., search) if recommendations were absent. Although the ideal way to estimate the causal impact of recommendations is via randomized experiments, such experiments are costly and may inconvenience users." Surprisingly, much of the large body of work on recommender systems has not necessarily prioritized this causal aspect, with Sharma et al. (2015) identifying only three papers using experimental designs and only seven using observational data to study the causal effect of recommendations.

They use access to browsing history from opt-in users of the Bing toolbar in Internet Explorers to see whether users going to Amazon pages, whose URLs contain flags for the origin of a click, did so through Amazon's search engine, another product page, other Amazon pages (e.g., wishlists), or from an external website. This by itself is not sufficient, they explain because the causal question is a counterfactual one: users might have still found products through other channels had recommendations not existed (in which case the recommendation merely provide convenient access to item pages). They used browsing data to reconstruct the overall Amazon recommendation graphs, and then used the presence a shock to only one of a pair of co-recommended objects as an instrumental variable. They found 4,774 shocks to 4,126 objects from September 1, 2013 to May 31, 2014. Their final estimate of the causal impact of recommendations is 3%, in line with the experimental findings they reference.

Contemporaneously with the submission of the article version of this chapter (Malik and Pfeffer, 2016b), and presented one month prior, is Su et al. (2016), who looked at the effect of recommendations on network structure via the introduction of Twitter's "Who To Follow" feature in 2010. While the discontinuity design with the introduction of the recommendation system is nearly identical to this work, the expected results are different, because Facebook's intended usage (with connections requiring reciprocation to exist) is different than that of Twitter (whose network is built on asymmetric follower relationships). We would expect Facebook to base recommendation systems on processes like (undirected) triadic closure, whereas we would expect Twitter to use cumulative advantage, and possibly *transitive closure* (Ripley et al., 2017), which is if i sends a tie to j and j sends a tie to k, then i will send a tie to k (e.g., k is higher in a hierarchy than j). Note that this is distinct from *two-out-star closure*, which is which i sends a tie to k, and j sends a tie to k, then i will send a tie to j (i.e., i connects with j because they are similar in their following of k).

Similar to some of my results, Su et al. (2016) find an initial increase in daily numbers of new edges, although they find a decrease afterwards (whereas I find no decrease in edges, but a decrease in the number of triangles created with each added edge). And, they find the fraction of new edges that are reciprocated decrease after the recommendation system's introduction. Overall, they find popular users benefitting disproportionately after the introduction of Who To Follow. They also find an increase in triadic closure, but by looking only at undirected triangles; it would be worthwhile to look at directed triangles, to see if the evidence is for a transitivity closure effect of a two-out-star closure effect which could shed light onto the likely mechanisms of the recommender system and the effect on the flows that are possible within the network. Also interesting is their discussion, where they conclude that a "mismatch between the recommender and the natural network dynamics thus alters the structural evolution of the network." Considering the implications of this, they note that cumulative advantage is often an undesirable property in terms of homogenization of ideas, although they also note that alternatively, there may be latent preferences towards following popular users such that the recommendation system only optimized natural dynamics, not caused unnatural dynamics. While their theoretical consideration of counterfactuals in terms of 'natural dynamics' is interesting, I would note that these systems do not necessarily have analogs to other networks (is it 'natural' to form networks based on communications limited to 140 or 280 characters? or to join a platform out of an interest in following celebrities, as Hargittai and Litt, 2011, find?) such that it may not be meaningful to talk about what is 'natural': we can only talk about comparison to dynamics within another regime of platform design, regulation, and engineered affordances. But the overall lesson is that different recommendation systems, designed to fulfill the goals and purposes of different platforms (respectively for Facebook and Twitter, to maximize connections between pre-existing acquaintances and to encourage more information consumption), will have different effect on platform networks. By identifying goals, we can anticipate what kinds of mechanisms recommendation systems might be designed with, and both the expected, direct effects (more following on Twitter, more friends on Facebook) and potentially indirect, unanticipated, or undesirable effects (rich-get-richer on Twitter, high local clustering on Facebook).

One other recent work is that of Cottica et al. (2017), who consider that online communities are indeed 'managed', and look at how to find the effects of such management. For example, they consider whether onboarding policies have an effect on degree distributions.

2.2.2 Social media networks

This work also relates to the theorization of social media networks. The most comprehensive theorizing is Kane et al. (2014), to which I connect my current investigation. They review Borgatti and Cross's (2003) grouping of social network research into four canonical types: how the network environment exerts influence on members, how resources spread through networks, how network structures benefit and/or constrain individuals, and how nodes use a network to access and benefit from resources. They then note,

> "In a social media context, network content is the digital content contributed by users, which may provide information, influence, or social support. . . Digital content flows through networks differently than other types of content; a physical object moving through a network occupies only one place at a time, whereas digital resources can be copied, manipulated, aggregated, and searched. While digital content is consistent with [social network analysis], its distinctive characteristics might mean that research on social media networks needs a specialized subset of measures and theories, with adaptations from traditional social network research.
>
> "Social media platforms quantify or formalize relationships or interactions between nodes by explicitly representing them in a formal data structure, operating on a computerized platform. This formalization provides relational capabilities in social media networks that are not present in offline social networks, including the ease of visualizing and analyzing the connections. However, the relational formalism of social media platforms also limits relational capabilities, such as by limiting the amount of nuance people can attribute to labels such as 'friend' or 'follower' (Gilbert and Karahalios, 2009). If people are limited to establishing similar formal connections with diverse sets of others including trusted confidants, casual acquaintances, and family members in their social networks, the platform homogenizes all of these relational connections as being equivalent (e.g., friends, contacts). Thus, while traditional SNA knows what ties mean but has difficulty eliciting these social data and measuring them objectively, social media can objectively measure ties through their digital traces but has trouble articulating the nuanced meanings of ties in a social context."

They note, however, that

> "Traditional SNA proceeds from a natural science paradigm, observing and describing the fundamental components of social networks in ways that best reflect how these networks are observed in the offline world. Social media, however, introduces questions of *design science*, how to implement the fundamental components of the network (i.e., nodes and ties) to achieve particular types of network behaviors (Ren et al., 2007)...In social media networks, tie features are not exclusively a reflection of the underlying social relationships that occur in the network but instead determine in part the nature and characteristics of the relationships that will occur on the platform. These design decisions regarding how relational ties are implemented will enable and constrain users' interactions."

That is, they bring up the causal aspect that platform design has on user behavior:

> "On the one hand, the features of an information system enable and constrain its users in particular ways, resulting in similar behavior among users of the same system. For social media, these features may be technical (e.g., capabilities provided by the platform), normative (e.g., policies and rules of the platform), or economic (e.g., incentives for certain types of use behaviors)."

This is not to say that users have no agency; Kane et al., 2014 note that

> "users may employ systems in ways that were unintended or unanticipated by designers (Boudreau and Robey, 2005)". But in the case of PYMK, it seems as though the platform succeeded in manipulating user behavior. In itself, this is not negative, as it manipulated it towards potentially desirable ends. Increasing network connectivity potentially increases access to resources: "The ability of users to *articulate their relational connections* and *view and navigate those connections* involves a capacity to visualize and manipulate the *network structure*—that is, how people establish and manage the connections between others in a network. Similarly, the ability to *establish a digital profile* and *access and protect content* contributed through the platform primarily involves *network content*, or how digital resources are shared and accessed through a network."

Based on this, they suggest adapted versions of the a 2 × 2 set of canonical types of social network research for social media, crossing structure and content with homogeneity and heterogeneity. This gives the research topics:

- structural homogeneity induced by the platform, e.g., how different types of ties (e.g., friendship ties, messages, tags) affect behavior and network formation
- content homogeneity induced by the platform, e.g., how available profile features affect behavior
- performance variation in structure from user behavior, e.g., how users or third parties utilize network structure to develop structural capital
- performance variation in content from user behavior, e.g., how people use content access mechanisms to access different resources.

This chapter fits into the first category: looking at how the platform features induce structural homogeneity, in my case how user behavior becomes homogenized in one particular direction by platform design.

Lastly and most important is Healy (2015), a preprint of which was published as early as 2012.[2] Healy discusses the 'performativity thesis', "the claim that parts of contemporary economics and finance, when carried out into the world by professionals and popularizers, reformat and reorganize the phenomena they purport to describe, in ways that bring the world into line with theory". He extends this to argue "that social network analysis is performative in the same sense as the cases studied in this literature."

The performativity thesis has two versions, the more interesting but empirically more difficult to demonstrate strong version, where models *create* or *determine* the reality they purport to describe ("the performative process brings the empirical phenomena into line with the original model. . . the model helps make itself true, in the sense that before its public appearance the system did not behave in accordance with the model's predictions, whereas subsequently it does"), and a more circumspect and empirically clear weak version, where models merely shape things. With this background, Healy asks,

> "Is there a parallel in the network case? In the previous section we saw a range of web services that put calculative devices in the hands of users in interesting ways. These devices act as 'cognitive prostheses,' in Callon's phrase—they allow users to do things they were unable to do before, such as easily see three or four degrees out of their social network, or discover which of thousands of strangers is most similar to them in their taste in books or music, or quickly locate people with similar financial goals, and so on. It is a relatively short step from here to taking advantage of these tools in ways that bear on actors' conformity to some aspect of network theory. To take a simple but significant example, Facebook uses its data on the structure of social relations to routinely suggest lists of 'people you may know' to users, with the goal of encouraging users to add those people to their network. In this way, the application works automatically to encourage the closure of forbidden triads in people's social networks—something which, in theory, should be the case anyway. This is likely also to increase the degree of measurable homophily in the network. Were a complacent analyst subsequently to acquire some Facebook data and run some standard tests on the network's structure, they would find—to their satisfaction—some confirmatory results about the structure of 'people's social networks.' Moreover, they would be able to claim that these results were plausible partly because of the scale of the data used for the analysis."

In other words, a paper from critical sociology independently proposed the exact same example as I use here, Facebook's People You May Know feature, as a window into how applications of models have a causal effect on human systems. Beyond this shared example, if van Dijck (2013) and Gehl (2014) provide the critiques that I operationalize around social media, on the side of networks, Healy's (2015) critique is effectively what I operationalize. Of course, the idea that the way we discuss the world (and how we act based on framings) can influence the world is a general constructivist insight (see also Hacking's 'dynamic nominalism'; Hacking, 2007), the core idea from which we both drew.

Subsequent work has also followed up on the sociological angle of manipulation, power, and control. Bucher (2017) takes up the critical angle of how Facebook users feel about being managed by Facebook's features, where they are even aware of this. From an STS perspective, Yeung (2017) further theorizes about 'regulation by design', which is how recommender systems are a form of regulatory governance.

[2] I thank Abigail Jacobs for bringing this paper to my attention.

2.3 Data and methods

2.3.1 Facebook New Orleans

Viswanath et al. (2009) detail how they collected the Facebook New Orleans data through a manual crawl of the New Orleans network, starting from a single user and using breadth-first search. Considering that Facebook started as a college-based network, the boundary specification (Laumann, 1973) of users who added themselves to the "New Orleans" network primarily (or those who chose to add it secondarily, perhaps after a college network) may not meaningfully match the college-centric boundaries within which links actually formed (especially since, as the authors point out, regional networks have more lax security than university networks, which require a valid email address from the university's domain). Second, only visible profiles could be accessed: the authors estimate, by comparison with statistics from Facebook, that they collected 52% of the users in the New Orleans network.

The Facebook data come in the form of timestamps of added edges between 63,731 unique nodes. About 41.41% of edges do not have a timestamp. On the data download page, Viswanath et al. (2009) write that "the third column is a UNIX timestamp with the time of link establishment (if it could be determined, otherwise it is [blank])" without elaborating on the reasons for missing labels; I make the assumption that these were the edges already present at the start of data collection. However, I find a great deal of repeated edges. Of the 1,545,686 rows of data, there are only 817,090 unique edges (i.e., 52.86% row are unique, 47.14% are redundant). Breaking it down, of the 640,122 rows that have no timestamp, only 481,368 represent unique edges, and of the 905,564 rows that have a timestamp, only 614,796 represent unique edges. 88,494 edges are repeated twice, 728,596 edges are repeated three times, and no edge is repeated more than three times. I make the decision to drop these repeated edges, assuming that repetition was the result of a repeat visit from multiple crawls (and assuming that timestamps were gathered by the time of detection via BFS, rather than extracted from profiles).

To the unlabeled edges I assign the minimum time present among the remaining edges, and for repeated edges I take their first instance only. Using the igraph library (Csárdi and Nepusz, 2006) I take the initial graph and calculate the number of edges, the number of nodes (i.e., non-isolates), the number of triangles, and the transitivity. Since the inter-arrival times are not particularly relevant for my question, I care only about the change in the relative rate over time, I aggregate my analyses by day to create time series: for each day, I add the edges that appeared on that day and recalculate the graph metrics. After, I also calculate the daily density using $2M/(N^2 - N)$ for the number of nodes N and number of edges M. I then difference each of these series, and for each day get the number of edges added, the number of nodes added, the number of new triangles, the change in transitivity, and the change in graph density. (Note that daily aggregation followed by differencing is equivalent to a histogram with day-wide bins, as Zignani et al. (2014) do for the number of triangles and edges.)

2.3.2 Netflix Prize

The Netflix data come in the form of text files for individual movies, with each line being the rating that a given user gave along with the date from 1999-11-11 to 2005-12-31. Following Koren's (2009a) plot, I take the daily average in order to see the sudden jump. Examining the number of ratings (i.e., the number of binned observations) per day, I find that they increase linearly in log scale. However, until 1999-12-31, ratings are not daily and even when present are small, whereas from 2000-01-05 (the next day for which there is data) there are daily ratings in the thousands. I take only the data on and after 2000-01-05.

My own investigation pinpointed the discontinuity as occurring on or around March 12, 2004. I could not find any public record of a platform change at that time nor any clues in press releases around then, and Netflix did not respond to a request for further information.

Statistically, the Netflix data are more straightforward as there is no social network.[3] However, the independence assumptions are more complicated; with a single dynamic network as in the Facebook New Orleans data, I can assume that the network-level rate metrics like the number of added triangles are independent observations across days. If we only consider the average daily rating, we do not take into account multiple ratings by the same individual (and, as Koren (2009a) notes, it is important to correct for different baseline average ratings across users, e.g. making sure an overall ‘stingy’ user’s ratings are comparable to those of an overall ‘generous’ user). But my interest is not in a full model of user ratings (predictive or explanatory), only a model of the average change to user behavior from a suspected platform effect. That is, we are interested in the marginal effect for which such dependencies are not relevant, and for which we can invoke the random sampling on ratings as a guarantee that my estimate will not have biases in representation.

2.3.3 Causal estimation with discontinuities

Regression discontinuity (RD) design is used to estimate causal effects in cases where there is an arbitrary (and preferably strict) cutoff along one covariate. As shown in Hahn et al. (2001), when the appropriate conditions are met, the treatment is effectively random in the left and right neighborhoods of the cutoff c. Causal effects are defined in terms of counterfactuals Y_{0i} (the value of the response were observation i to not be treated) and Y_{1i} (the value of the response were i to be treated); the point difference between the two at the time of intervention for treated populations is called the *local average treatment effect* (Imbens and Angrist, 1994), α. Given an observed Y_i, this is given by

$$\alpha \equiv E(Y_{1i} - Y_{0i} | X_i = c) = \lim_{x \downarrow c} E(Y_i | X_i = x) - \lim_{x \uparrow c} E(Y_i | X_i = c) \tag{2.1}$$

In the linear univariate case, the model is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{1}(x_i > c) + \beta_3 x_i \mathbf{1}(x_i > c) + \varepsilon_i \tag{2.2}$$

which effectively fits two separate lines, one for each ‘population’ before and after the cutoff, with the estimated $\hat{\alpha}$ being the difference between the two fitted lines at the cutoff. The interest is generally in estimating the causal impact, but as a specification test (Imbens and Lemieux, 2008), the joint test for $H_0 : \beta_2 = \beta_3 = 0$ corresponds to a null hypothesis that there is no discontinuity. This model and the corresponding test may be generalized with higher-order polynomial terms. The model also has a natural

[3] Netflix did briefly attempt to add social networking features in late 2004. However these were discontinued in 2010, with part of the justification being that fewer than 2% of subscribers used the service.

nonparametric extension: separately fit the same smoother on either side of the discontinuity to estimate the effect, or, test for the discontinuity by seeing if confidence intervals overlap.

Note that the exemplars of RD design are not temporal, and many standard parts of time series modeling are incompatible with RD design. For example, a discontinuity is necessarily nonstationary, and differencing will destroy it (I fitted ARIMA models, and found that differencing was indeed necessary), and similarly, a one-sided moving average smoother applied to both sides of the discontinuity will leave a gap. I found two alternative methodologies created specifically around time series, 'interrupted time series analysis' (McDowall et al., 1980; Wagner et al., 2002; Taljaard et al., 2014) and 'event studies' (MacKinlay, 1997), but both are essentially less formal versions of RD design and still neither account for temporal features (namely, autocorrelation). I also tried Gaussian Process (GP) regression (Rasmussen and Williams, 2005; MacDonald et al., 2015), as it is able to capture temporal dependencies (Roberts et al., 2012). A squared exponential covariance function gave largely similar results, including posterior intervals about as wide as confidence intervals from other methods (and thus perhaps still not capturing autocorrelation) when fitting separately to either side of the discontinuity. I note that it may be possible in future work to adapt covariance functions that account for 'changepoints' (Garnett et al., 2010) not just to make predictions in the presence of discontinuities, but to do causal inference within the RD framework.

As we are interested in the central tendency rather than on features of the time series, I prioritize the use of the RD framework over time series modeling. To apply RD design, I make the assumption that the respective times at which People You May Know and whatever change took place in Netflix were introduced were effectively random. I use time as the covariate, with the respective cutoffs for the two data sets of 2008-03-26 and 2004-03-12 (i.e., I code for the potential discontinuities starting on those days). I apply nonparametric models, and specifically, local linear regression as is standard in regression discontinuity design (Imbens and Lemieux, 2008) and is also appropriate for time series (Shumway and Stoffer, 2011).

While a nonparametric smoother has the advantages of being able to fit cyclic behavior without including specific cyclic terms, confidence intervals still fail to capture the extent of cyclic variance and so are too optimistic even beyond not accounting for temporal autocorrelation (Hyndman et al., 2002). Prediction intervals are an alternative as they include the overall variance, but are not straightforward to calculate for smoothers. Another alternative, which we use for the Netflix data and for edge counts in the Facebook data, is to use local linear quantile regression (Koenker, 2005) to get tolerance (empirical coverage) intervals, and specifically, using the interval between a fit to 5% and to 95% to get a 90% tolerance interval (I found too much noise for fits at 97.5% and 2.5% to use a 95% tolerance interval). This is analogous to an idea in Taylor and Bunn (1999), who produce forecast errors for an exponential smoother using quantile regression.

For consistency, when I do this I also use quantile regression for the central tendency (i.e., using the median instead of the mean), which is also known as or "robust regression" and has the advantage of being more robust to outliers.

2.4 Results and discussion

2.4.1 Netflix Prize data

First, I note that the number of daily ratings increases over time (fig. 2.4), which corresponds to decreasing variance in the time series plot, suggesting use of weighted least squares. Weighting by the number of daily ratings (so that the days with more ratings are counted more heavily) improved diagnostics across the parametric models I considered; however, I found that the addition of polynomial terms up to and even past 7th order continued to be significant, leading me to prefer the nonparametric approach that can capture the cycles without becoming cumbersome. In fig. (2.5), I show the results of the local linear quantile regression. As we can see, at the cutoff the two 90% tolerance intervals do not overlap, allowing us to reject the null hypothesis that there is no discontinuity at the 0.10 level.

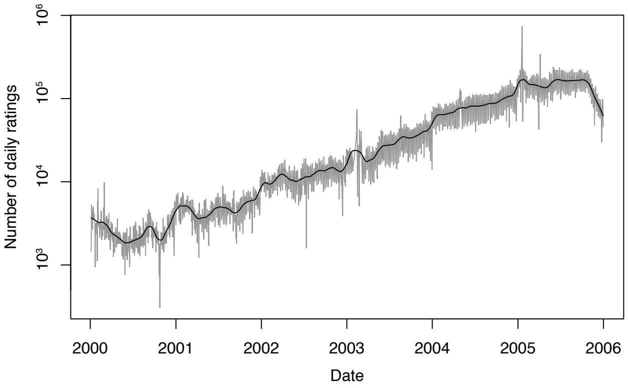


FIGURE 2.4: The number of Netflix ratings increases over time (y-axis shown in log scale); and, we can observe from fig. (2.3), the variance decreases over time, suggesting using the counts as weights. The fitted local linear smoother, which I used for weights, is shown in black. The bandwidth of .026 was selected via 5-fold cross validation.

To test if the model detects jumps at non-discontinuity points, I tried each day as a cutoff. Other than the actual discontinuity, the only points where the tolerance intervals did not overlap were two points before the cutoff I used (March 10th and 11th) and one day after (March 13th). Since I had initially located this date through manual (graphical) investigation, and the choice was not unambiguous within several days, it is unsurprising that the model picks this up as well. While this ambiguity is likely a matter of noise, platform engineers commonly deploy new features gradually to decrease risk, so it is also possible that the ambiguity is a gradual rollout that the model is also detecting.

Sensitivity to the smoothing bandwidth (the tuning parameter which controls the size of the neighborhood used in local fitting) is a concern for estimating the causal effect, so as is recommended, I report the estimates across multiple bandwidths. From 5-fold cross-validation, the optimal bandwidth of 6 (i.e., using kernel $K(x^*, x_i) = \exp\{-.5((x^* - x_i)/6)^2\}$), performed poorly under specification testing, identifying many discontinuities. Larger bandwidths (where the estimator tends towards linear) performed better, but at large bandwidths, again many discontinuities were identified. This is not ideal but unsurprising given the loss function used in quantile regression; quantiles are less swayed by extreme values, such that the non-overlap of tolerance intervals properly capture that there is a discontinuity even far from the actual discontinuity. The estimate of the causal effect may still be good, but with the failure of the specification testing at both low and high bandwidths, I report only within the range that performed well.

I estimate the local average treatment effect, the average amount by which the platform change resulted in a change in user ratings, as 0.118 from a bandwidth of 25 (pictured in fig. 2.5), 0.126 from a bandwidth of 50, 0.124 for a bandwidth of 75, and 0.119 for a bandwidth of 100. Considering the ratings prior to the cutoff had a mean of around 3.44, these amounts are a substantial increase, and are about 3% of the total possible range of ratings (from 1 to 5). This is a less involved case than Facebook, since movie preferences are a relatively low stakes phenomenon, but it shows the application of regression discontinuity. If the cause of the discontinuity is indeed a change in wordings, it shows that, just as in survey research, a change to the format changes the distribution of answers; but unlike in surveys, with large-scale online (streaming) systems, changes become visible as discontinuities in time.

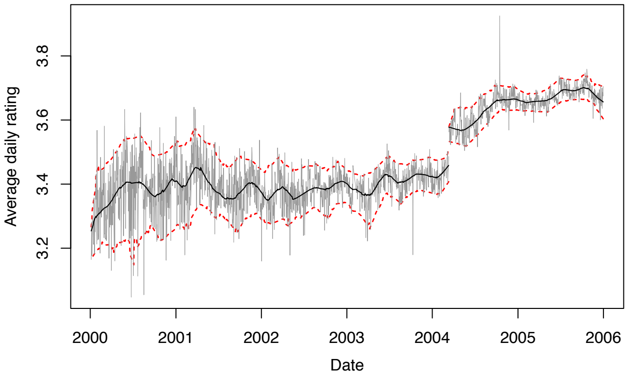


FIGURE 2.5: The solid line shows the local linear fit for the median Netflix ratings. The dashed lines give a fitted 90% tolerance interval, from local linear quantile fits to 5% and 95%. The intervals on both sides of cutoff do not overlap.

2.4.2 Facebook New Orleans data

Fig. (2.6) shows the discontinuity in the Facebook New Orleans data across four graph metrics. In addition to the daily counts of the number of added edges and added triangles as examined by Zignani et al. (2014), the discontinuity is pronounced in the transitivity and the density as well (although the units of these are so small as to not be particularly interpretable, so I do not estimate a local average treatment effect).

For the number of edges, I first used a fifth-order polynomial Poisson regression (not pictured), which had excellent regression diagnostics, from which I estimated a local average treatment effect of 356. This is more than a doubling of the pre-cutoff daily average of 314. However, the confidence intervals from the Poisson regression were very narrow and performed poorly under specification testing (as did bootstrap prediction intervals, which were very wide), in addition again to the problem of relying on higher-order polynomial terms rather than just relying on a nonparametric approach, so I also made fitted tolerance

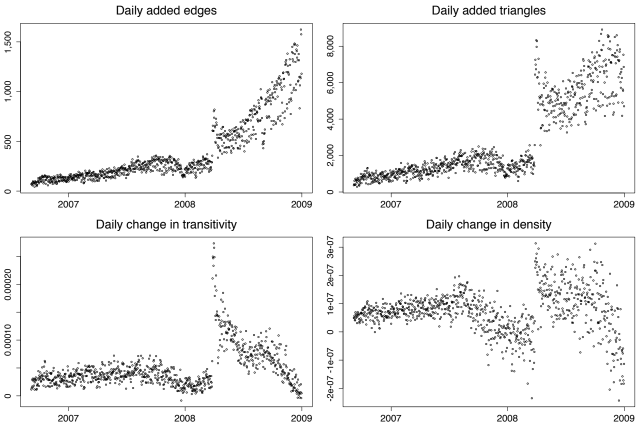


FIGURE 2.6: For the Facebook New Orleans data, the daily added edges and triangles created (top left and right, respectively), and the daily change in transitivity and graph density (bottom left and right, respectively).

intervals using local linear quantile regression as with the Netflix data, shown in fig. (2.7). Again, the optimal bandwidth found from 5-fold cross-validation was small and performed poorly under specification testing, as did large bandwidths (tending towards linear). Reporting within the range that performed well under testing, I estimate the local average treatment effect as 319 from a bandwidth of 25 (pictured), 278 for a bandwidth of 50, 228 for a bandwidth of 75, and 201 for a bandwidth of 100.

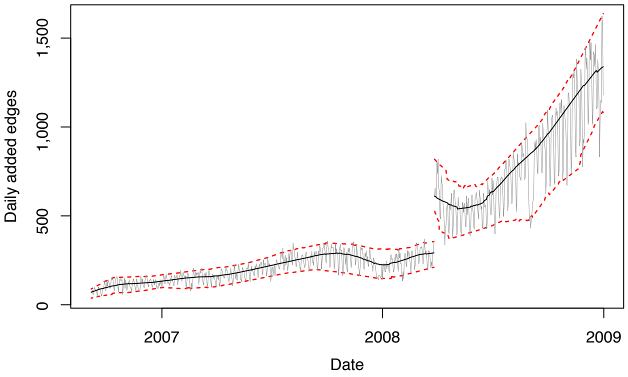


FIGURE 2.7: A local linear fit for the median number of edges added daily in Facebook New Orleans. The dashed lines give a fitted 90% tolerance interval, from local linear quantile fits to 5% and 95%. The intervals on both sides of cutoff do not overlap.

As the number of edges and triangles are closely related (fig. 2.8) and there are enough observations for a ratio to not be a noisy estimation target, I follow Zignani et al. (2014) in taking the ratio of triangles to edges. This represents the average number of triangles created by each added edge, and captures the extent of triadic closure on a scale more interpretable scale than that of changes in transitivity (which are in the ten thousandths). For a parametric model with an indicator for the discontinuity as described in eqn. (2.2), up to fourth-order polynomial terms were significant additions to the model in partial F tests, which is implausible and not parsimonious, so I again prefer a nonparametric fit, shown in fig. (2.9), which estimates a local average treatment effect of 3.86. This is even more dramatic than the effect in Netflix; given that the mean ratio was estimated at 6.25 before the jump, this is an increase of 61.8%.

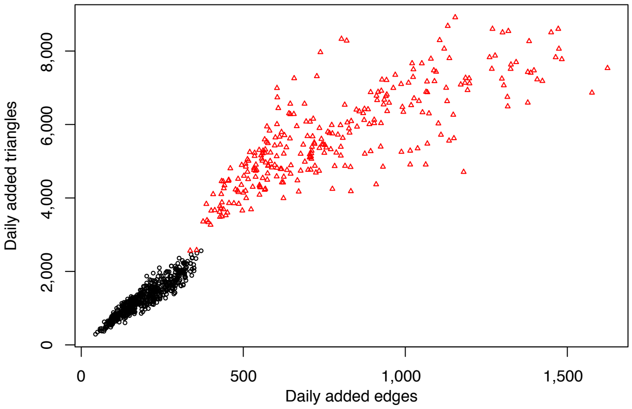


FIGURE 2.8: The daily added edges and triangles have a close relationship in the Facebook data. Black circles are time points before 2008-03-26, and red triangles are time points afterwards.

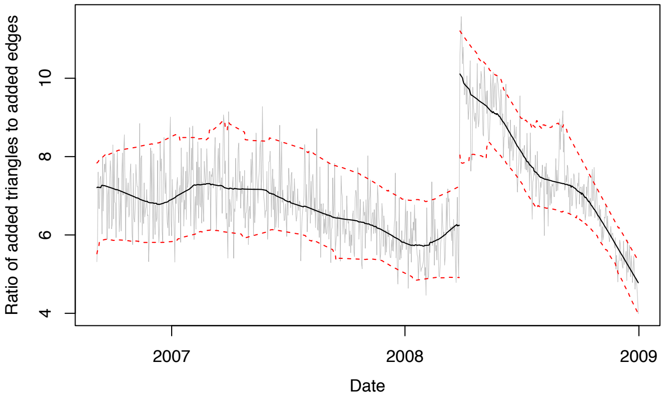


FIGURE 2.9: A local linear fit for the median daily ratio of added triangles to added edges in Facebook New Orleans. The dashed lines give a fitted 90% tolerance interval, from local linear quantile fits to 5% and 95%.

2.5 Conclusion

For much of data analysis, discontinuities (such as from abrupt platform changes in social media) are seen as incidental, or annoyances to be corrected (Roggero, 2012). Indeed, they appear in the literature as curiosities or asides. However, given the theoretical concerns about the nature of social media data, they can give valuable insights. My finding about the change in average Netflix ratings echoes work in survey research about response item wordings in a different setting and with different sort of data, quantifying how much we might expect a platform change to shift a baseline, and the sizes of 3% matches the size of causal estimates of a different process, that of recommendation systems, described in Sharma et al. (2015). For the Facebook New Orleans data, the finding is even more dramatic and widely applicable: we now have a sense that the introduction of a triadic closure-based recommender system can nearly double the rate of link creation. Furthermore, it changes the nature of the created links (focusing on closing triads), which has repercussions for the graph structure, seen for example in the changes in density. This is far above previous estimates of the effect of recommendation systems, which could either mean there are additional confounders, or that applying recommendation systems on a *network* rather than on an individual is something qualitatively different.

This also provides an empirical extension of a concern raised by Schoenebeck (2013) about how variation in technology adoption creates online social networks that differ systematically from the underlying social network: from my results, we see it is not just the process of joining social networking sites that creates observed network properties, but also the ways in which platforms design influences users. Multiple works have considered whether network metrics of large online social networks differ from those of previously studied social networks (Corten, 2012; Quercia et al., 2012; Ugander et al., 2011; Mislove, Marcon, et al., 2007); we can continue to theorize how differences result from platform effects, usage patterns, and demographic representation, rather than from online platforms being a superior way to measure social networks.

There are concerns about what social media ties even represent (Lewis et al., 2008), with some authors pointing to interactions over ties (Viswanath et al., 2009; Romero et al., 2011; Wilson et al., 2012; Jones et al., 2013) as more meaningful than the existence of ties. But my results show that the problem is not just one of ties not being a rich enough measure, but that they a non-naturalistic measure of social relationships, and furthermore, their existence determines visibility and access and thereby what activity happens. As people accept suggested links and begin interacting, the underlying phenomenon (the relationships and the network effects) changes, whether for good (Burke and Kraut, 2014) or ill (Kwan and Skoric, 2013). On Netflix, if changes affect different movies differently, it has consequences for modeling user behavior preferences. Beyond research concerns, there are economic benefits for the creators of movies that benefit from platform changes. Lotan (2015) observed this potentially happening in Apple's App Store, where what appeared to be an (unannounced, undocumented) engineering change in the search results ranking led to changes in app sales.

Regression discontinuity design has a rich literature, and there will likely be many future cases where we can apply RD design or interrupted time series in social media data. In geotags collected from the US in 2014, there was a sudden decrease (fig. 2.10) on September 18th, the same day Twitter released significant updates to profiles on Twitter from iPhone.[4] The recent increase in character limit on Twitter from 140 characters to 280 was, after being trialled with a small number of users, [5] rolled out en masse on November

[4] "A new profile experience on Twitter for iPhone", September 18, 2014, https://blog.twitter.com/2014/a-new-profile-experience-on-twitter-for-iphone, accessed 1/2016.

[5] "Giving you more characters to express yourself", 26 September 2016, https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html, accessed 8/2018.

7, 2017.[6] Indeed, one of the justifications for the change was that the artificiality of the 140 character limit was clear in how 9% of tweets came up against this limit, and the distribution of tweet character length being bimodal; changing the character limit to 280, the announcement on Twitter blog noted, reduced the number of tweets hitting the limit to 1%, and in the distribution of tweet length the bimodality disappeared. But more importantly, people who came up against the 140 character limit were likely adopting conventions of abbreviations, extensively editing, or splitting thoughts into multiple tweets (as also noted in the announcement), and looking past just the distribution of characters per tweet to how certain tweeting conventions become less common would give an idea of the causal impact of platform constraints. It would also be interesting to look at whether only 1% of tweets hitting a 280 character limit is a persistent effect; as people get used to a longer limit, will more and more tweets start coming up against the new 280 character limit? There is also an effect of language; the announcement noted that there would be no change in the 140 character limit for Chinese, Japanese, or Korean, as far fewer tweets in these languages were hitting the 140 character mark. In an example that ties in with Chapter (1), Tasse et al. (2017) note a sharp dropoff in the number of geotagged tweets in May 2015, which they attributed to a change in the user interface that made place-tagging, rather than geotagged, the default. In their survey results, they found geotag tweet users who were unaware about the level of precision of geotags, as their intention was only to provide a general location and not a coordinate. The change in the default behavior brought geotagged tweets more in line with users' understandings of what the platform feature was actually doing. These show how there has begun to be a public body of knowledge about the ways in which platform design are responsible for observed behavior, a body of knowledge that outside researchers can continue to build on. Extensions to regression discontinuity are also relevant, for example in how Porter and Yu (2015) develop specification tests into tests for unknown discontinuities.

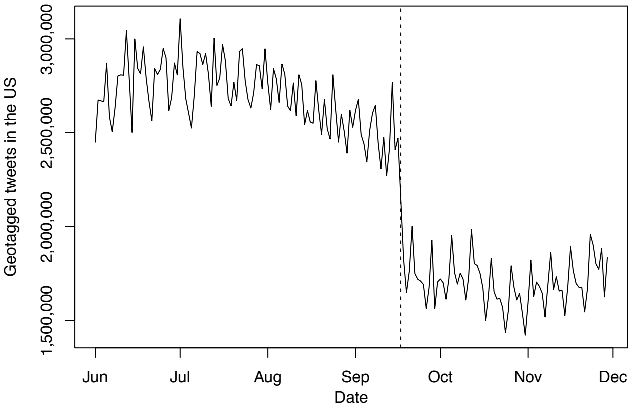


FIGURE 2.10: Another potential discontinuity, seen amidst cyclic behavior in the volume of geotagged tweets collected in the US in 2014.

Social media data have been compared to the microscope in potentially heralding a revolution in social science akin to that following the microscope in biology (Golder and Macy, 2012). This metaphor may have a deeper lesson in a way that its advocates did not expect: history of science has shown (Szekely, 2011)

[6] "Tweeting made easier", 7 November 2017, https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html, accessed 8/2018.

that it was not a simple process to connect the new instrument, with its multiple shortcomings, to the natural objects it was supposedly being used to study. It took centuries of researchers living with the microscope, improving the instrument but also understanding how to use it (e.g., recognizing the need for staining, or the importance of proper lighting), that microscopes became a firm part of rigorous, cumulative scientific research. I would hope that social media data will not take as long, but at the same time, it is as necessary as ever to question the relationship between the novel instrument and the object of study.

Chapter 3

Sensors and social network data: Measurement, models, and meanings[1]

Summary. Here, I review and critique the work that has been done using sensors to gather social network data. I focus on several large sensor data collection projects, and examine how they describe sensor data, and what sort of models they use on such data. I argue that casual treatments of constructs has led to a conflation of "interaction" and "proximity", and that distinguishing these helps identify the kind of research questions for which sensors are most useful, as well as identify needed demonstrations for establishing certain kinds of validity. I also argue that models that do not incorporate network processes are ultimately unsatisfying, and that models developed within social network analysis are fruitful for use with sensor data.

3.1 Introduction

Sensors, which include accelerometers, gyroscopes, barometers, Radio-Frequency Identification (RFID) chips, radio antennae using the Bluetooth and Wifi communication standards, and Global Positioning System (GPS) antennae, are present throughout consumer electronics. Mobile phones in particular are equipped with a bevy of sensors, but increasingly, 'Internet of Things' (IoT) functionality involves putting inert RFID chips in every conceivable object (from coffeemakers to clothing to pencils to plants). These chips can then be read by RFID readers and, in conjunction with external databases, can be used to track objects over time (Hildner, 2006). Like data from online activities before them, sensor data are an emerging type of 'big data'—and will perhaps in the future even represent the dominant form of data both in volume and in value.

Since 2002 (Choudhury and Pentland, 2002), sensors have been used to study social networks. However, sensors have not yet been systematically considered within social network analysis; for example, major recent introductions to the field (Borgatti, Everett, et al., 2013; Hennig et al., 2013; McCulloh et al., 2013; Scott, 2012; Kadushin, 2012; Marin and Wellman, 2011; Prell, 2011) have not mentioned sensors alongside other social network data collection methods, or only briefly mention some future possibilities (Robins,

[1] This is work done in collaboration with Jürgen Pfeffer, Afsaneh Doryab, Michael Merrill, and Anind Dey.

2015). Only three papers in *Social Networks* have used sensor-based data collection (Stehlé, Charbonnier, et al., 2013; Starnini, Baronchelli, et al., 2016; Hertzberg et al., 2017).

Certainly, one reason for this is that thus far, the use of sensors for research has involved a high technical overhead, making it inaccessible to everybody outside of specialized engineering researchers. A second possible reason is that there have not been clearly articulated theoretical frames for the use of sensors for social network data collection, which risks making sensor data a novelty rather than a substantive research tool. As the technical overhead drops, having a clear understanding of the technical possibilities and limitations of sensor data, along with having clear theoretical frames, will help inform possibilities for future investigation.

In this chapter, I review the nature of sensor technology and data, and the ways in which it has been used so far to collect social network data. I also argue that there has been insufficient theorizing about sensor data: that is, there is a need to identify specific constructs of interest, systematically consider the validity of sensors for measuring these constructs, and design studies around possible causal relationships between such constructs and various covariates and/or outcomes of interest. I argue that the common usage of many different kinds of sensors to capture *face-to-face interaction* is not precise enough; the measurements are all *proxies* for the construct of interaction rather than ways of measuring it directly, different sensors are more or less effective proxies, and there are other constructs of interest for which sensors can be used. I also discuss models that are relevant for processing sensor data, including machine learning approaches to data reduction in place of using simple aggregate counts.

To demonstrate my arguments of theoretical possibilities and for modeling approaches from this chapter, in Chapter (5), I present the results of an original study comparing sensor data to self-reported friendship (sociometric choice) data. In that chapter, I also discuss the implications of technical decisions that go into how to collect data and how to process collected data.

3.2 Background

3.2.1 Sensors

The word ‘sensor’ has been used broadly, so it is worth defining the term and specifying what I mean by it. I take a *sensor* to be any device that takes measurements of a physical quantity. This encompasses many familiar devices associated with measurement or detection (i.e., a binary measurement), for example thermometers, odometers, smoke detecters, geiger counters, barometers, sound level meters, gyroscopes, metal detectors, and motion sensors. In some cases, measurement can be through a direct process, such as thermometers using the expansion/contraction of alcohol under different levels of heat; in other cases, such as with ionization smoke detectors, detection is through a specialized methods (using internal ionization chambers to distinguish smoke from other air). Global Positioning System (GPS) receivers are sensors that measure geolocation, but they do so indirectly, requiring a global network of satellites whose signals they can triangulate.

There are also *biometric sensors* that measure physical quantities related to bodies, which are usually examples of more indirect measurement. For example, breathalyzers use alcohol in breath as a proxy for blood alcohol content, and heart rates monitors typically use either skin electrical activity or shine a light on skin and then measure absorption of red light and infrared light (this process can also measure blood oxygen saturation).

Sensors can also detect something seemingly trivial: the presence of another sensor of the same type. This is done using radio waves using a specific protocol, such as Bluetooth, Radio-Frequency Identification (RFID), or even the WiFi standard. There are 'sensor nodes' that do nothing but transmit their unique ID and record the unique IDs transmitted from other devices. But physical *networks* of such nodes can collectively give information such as about relative positions of moving objects—or of moving people.

'Sensor' can also refers to systems built on top of (potentially specialized versions of) generic sensors, such as specialized cameras for fingerprint or retinal scanning, or using reflective markers with cameras and 3D software to do and motion capture or to capture gait. I do briefly discuss some of these. However, I exclude metaphorical uses of 'sensors' and 'sensing,' such as around activity on social media platforms (Goodchild, 2007; Sakaki et al., 2010; Christakis and Fowler, 2010).

3.2.2 Relational sensor data

It is possible to use sensor-based measurements as node-level covariates; Kitts and Quintane (forthcoming) give a review:

> "...researchers may monitor galvanic skin response, pupil dilation, or heart rate (Palaghias et al., 2016; Salah et al., 2011), employ brain imaging (O'Donnell and Falk, 2015), monitor hormones in saliva or urine samples such as oxytocin (Doom et al., 2017; Grebe et al., 2017); or cortisol and testosterone (Ketay et al., 2017; Kornienko et al., 2014; Mehta et al., 2017). They may also automatically monitor sentiment-related nonverbal behavior, such as eye gaze or body posture (Mast et al., 2015), response latency (Iyengar and Westwood, 2014), analyze speech features in audio recordings (Rachuri et al., 2010), use accelerometers to detect laughter (Hung et al., 2013), use chest bands to monitor breathing patterns during conversations (Rahman et al., 2011), or even reflect radio frequency signals off of the body to detect emotional states through physiological responses (Zhao, Adib, et al., 2016)."

For example, such readings could be compared to sociometric choice data, as Parkinson et al. (2018) did for brain activity. But my interest here is in how sensors may themselves gather *relational* data, that is, data reflecting some sort of dyadic relation.

Smartphones have become a dominant tool for sensor-based studies, because of how they are equipped with over a dozen sensors and are used widely. I review these sensors, and consider which are relevant for gathering relational data.

- Accelerometers. These measure acceleration, which on its own is not useful, but can help infer the trajectories in-between detected locations such as through models like state space models, and algorithms on such models like the 'Kalman filter' (Hendeby et al., 2014).

- Audio. Technically, microphones are a type of sensor; work has looked at how to use audio recordings, such as from cell phones, to detect conversations between individuals (which involves both detecting the start and end of when a single person is speaking, and identifying specific individuals by the waveforms of their voices) and thereby create ties between them (Wyatt et al., 2011; Basu, 2002). If the content of communication is not considered (as is sometimes done as a privacy-preserving measure), this effectively becomes communications metadata which have been previously theorized (Butts, 2008; Monge and Contractor, 2003). But *in-person* conversation is potentially theoretically distinct from conversation in other media like radio. Conversely, rather than recording full waveforms, only the audio volume level can be collected; in this case, the data is not relational, but it can provides some information about an individual's context (whether it is noisy around them or not).

- Barometer. An increasing number of smartphones come with barometers; this can be useful for detecting height above sea level, but this is not wholly reliable.

- Battery. Internal logs record the battery level either whenever it changes (such as by 1%) or at regular sampling intervals, as well as the status of charging (phone is not charging, phone is charging, or phone is fully charged). This is not a sensor of the environment, but monitoring this can be a useful and important way to check compliance in a sensor study (e.g., if there is no sensor data recorded right after the battery was recorded as being at 1%, then it is likely any subsequent missing data is because the phone was off).

- Bluetooth. Technically, Bluetooth is a *standard* for encoding and decoding radio signals, not a sensor in itself (and Bluetooth capabilities may go through an existing antenna rather than having a dedicated Bluetooth antenna), and Bluetooth is used for transmitting specific data (e.g., Bluetooth headphones or speakers wirelessly transmitting music from a device to an audio output), but we consider the capability of devices to detect the presence of other Bluetooth-enabled devices to be a sensor. This is an important case; detecting a known other Bluetooth device tells us that two devices are proximate, and indeed several significant mobile phone studies use Bluetooth as the method of detecting proximity between individuals.

  Bluetooth detections also record the Received Signal Strength Indicator (RSSI), a measure of how strong the signal of the other Bluetooth device is. In theory, this can be used as a proxy for the specific distance of two devices; however, environmental interference can weaken signals in arbitrary ways, making RSSI a highly unreliable proxy for distance (Hossain and Soh, 2007). Bluetooth is thus best used only for detecting proximity.

- Communications, which are the logs and/or metadata of calls and SMS (text messages), are often included as a type of sensor. Communication data are inherently relational, and communications logs from multiple devices can be used to build networks when the identity of both senders and receivers are known;[2] however, unlike audio data, the dynamics of and processes present in telecommunications-mediated communications are thoroughly theorized, for example as communications networks (Monge and Contractor, 2003) or as relational events (Butts, 2008).

[2] A common privacy-preserving measure is to *hash* phone numbers of recorded alters, which is running phone numbers through a function that is one-to-one but extremely hard to invert from knowing only inputs and outputs; hashing allows tracking repeat communications of ego with their alters, but prevents identifying common alters for different egos.

- Gravity. This is detected through accelerometers, but is also not useful for relational data.
- Gyroscopes. These track changes in a phone's orientation, and are not useful for getting relational data.
- GPS. Mobile devices will have a dedicated antenna for communicating with GPS satellites in order to calculate the phone's current geolocation in latitude and longitude. The geolocations of a dyad can, in turn, be used to calculate pairwise distance, which is relational. The GPS present in mobile phones is accurate to within about 10m, but has a high power consumption and quickly drains batteries, which can give co-location but not interaction. Recent technical advances claim the use of minimal power to achieve accuracy to within 1 cm (Chen, Zhao, et al., 2015); this technology is not yet present in commercial devices, but might make GPS data particularly useful. Still, buildings block satellite signals, such that GPS has limited usefulness indoors regardless of accuracy. Using cell phone towers and WiFi (see below) can be a more energy-efficient way of getting geolocation.
- Light. Phones record the level of ambient light to regulate screen brightness, but this is not useful for gathering relational data.
- Magnetometer. This serves as a compass, tracking magnetic fields to help get the absolute orientation of the phone. While knowing whether two people are facing each other could supplement relational data of proximity or co-location, this is not practically useful since unless a person is using a phone for navigation (i.e., is pointing the phone in the direction they are facing), the phone can be oriented any which way relative to the person carrying it.
- Processor. Tracking the level of processor usage can be a proxy for when and how much a phone is being used, but this does not give relational data.
- Proximity. A phone's actual 'proximity sensor' detects the proximity of the phone's screen to immediate objects, like a user's face (this is how phones know to shut off their screens when held up to talk, so that a user's face doesn't accidentally press things on the screen). This too does not give relational data.
- Rotation. This is done through gyroscopes, and is also not useful for getting relational data.
- Screen. This is the on/off status of the screen. This does not give relational data but is surprisingly informative, as for many people, their phones are the last thing they look at in the night and the first thing they look at in the morning; this means screens being on or off can be a proxy for when people sleep (Min et al., 2014).
- Temperature. A thermometer is not present in all phones, and is of limited use anyway as because phones are commonly kept in (warm) pockets and also generate their own heat. For knowing the weather (e.g., for considering possible effects of sunlight versus lack sunlight), it is better to use detected location and consult an external source of weather data, such as historical NOAA data.
- Telephony. This is the record of the cellular network antennae to which a phone connects. The geolocation of these antennae are known (whether they are placed atop cell phone towers or, for many of the antennae found in cities, simply on the sides of buildings); in theory, triangulation between multiple antennae could be used to get location, but in practice, only the most available tower communicates

with a given phone. This is usually the nearest tower, but if the nearest tower is busy, a signal will bounce to a more available nearby tower (the telecommunications engineering of this makes the process happen efficiently and automatically, but not in a way that is recorded). Thus, with locations of cell phone towers treated as centroids, a phone may be located to within a Voronoi cell of the antennae to which it is connected. Antennae are installed far more densely the greater the population density and urbanization, and so locations determined thusly is more fine-grained in cities (to within a few city blocks) than in rural areas where a single antenna may cover many miles. This type of data can give co-location within a Voronoi cell.

- WiFi. Like Bluetooth, WiFi is a standard for radio signals rather than a sensor, and may not have a dedicated antenna. Referring to WiFi as a sensor means the detection of WiFi hotspots by devices with WiFi enabled (or, conversely, the detection of devices by the hotspots). Each WiFi hotspot has a unique ID, a MAC address. Manually recording the physical location of these hotspots can help determine the approximate location of a mobile device that is in range of that WiFi hotspot.[3] Being in range of multiple Wifi hotspots can help triangulate a more precise location, although this is more difficult and requires calibration for a specific set of WiFi devices. Mutual detection of a WiFi device can give relational data; or, WiFi can be used to first get absolute location of two devices, from which can be calculated pairwise distance.

Telephony antennae or WiFi hotspots, with phones that connect to them, form two-mode networks which can be projected to one-mode co-location networks between phones. Continuous-valued location data, such as geolocation (in latitude and longitude) can be used to calculate a pairwise distance matrix between sensors, which can be thresholded to make a *co-location network*. Of course, both these measures of proximity and pairwise distance are over time; discrete temporal measurements at regular intervals naturally produce longitudinal data, but if taken at irregular intervals, measurements need to be aggregated in time bins or otherwise smoothed, such as with a sliding window from which regularly spaced samples can be taken.

The majority of large-scale sensor projects looking at social networks use the sensors in mobile phones, such as the MIT Human Dynamics groups' 'Reality Mining' study (Eagle and Pentland, 2006; Chronis et al., 2009), 'SocialfMRI' or 'Friends and Family' study (Aharony et al., 2011), and Social Evolution study (Madan, Cebrian, et al., 2010; Dong et al., 2011), as well as the Lausanne data collection campaign (Kiukkonen et al., 2010), the Copenhagen Networks Study (Stopczynski, Sekara, et al., 2014; Sekara and Lehmann, 2014), and other less extensive and/or primarily non-relational projects (Kostakos and O'Neill, 2008; Do and Gatica-Perez, 2011; Li et al., 2012; Yan et al., 2013; Ghose et al., 2013; Wang, Chen, et al., 2014; Jayarajah et al., 2015). The main mobile phone sensor in these cases is Bluetooth for detecting proximity, although location data (from GPS, cell towers, or a combination of sensors) or WiFi (Montjoye et al., 2014) for detecting co-location are also sometimes used.

However, there is also a significant amount of work using self-contained sensor 'nodes' (also called 'badges' or 'beacons' or 'tags'), which are worn hanging from lanyards around people's necks and detect the presence of other proximate sensor nodes by various radio signal protocols. In fact, the very first sensor studies were done with sensor nodes. Specifically, the first sensor work was with the 'sociometer' (later, 'sociometric

[3] In a practice known as 'wardriving', companies have vehicles drive around and record detected WiFi devices and record the relative geolocation (the Google street view vans do this, for example) which are then sold to various consumer service providers. Such services are how turning on WiFi helps mobile devices calibrate their geolocation.

badge') of Choudhury and Pentland (2002), from the Human Dynamics Group at the MIT Media Lab, although this (along with the UbER-Badge, also out of the Human Dynamics group; Laibowitz et al., 2006; Paradiso et al., 2010) is different from subsequently used devices. Sociometric badges are a custom-made devices, unlike the generic sensors used in subsequent work; the badges combined a number of sensors, most notably a microphone for detecting conversations along with sensors for proximity detection, unlike the single-sensor nodes of subsequently used devices; and it did proximity detection through infrared, rather than through radio signals like subsequent devices.

There have been sensor nodes using a number of different standards (Hsieh et al., 2010; Friggeri et al., 2011; Angelopoulos et al., 2011; Förster et al., 2012); and there are Bluetooth beacons as well that use the energy-efficient Bluetooth Low Energy (BTLE) standard (Ahmetovic et al., 2016). Such sensor nodes allow precise control of infrastructure, but involve high participant burden and require recharging, such that they cannot be used for studies longer than a few days. For this reason, sensor nodes have been replaced by mobile phones in all cases but one: the RFID badges used by the SocioPatterns group (Barrat, Cattuto, Colizza, Pinton, et al., 2008; Cattuto, van den Broeck, et al., 2010; Stehlé, Voirin, Barrat, Cattuto, Isella, et al., 2011; Van den Broeck, Cattuto, et al., 2010; Szomszor et al., 2011; Isella, Romano, et al., 2011; Van den Broeck, Quaggiotto, et al., 2012; Fournet and Barrat, 2014; Mastrandrea, Fournet, et al., 2015; Voirin et al., 2015; Pachucki et al., 2015; Génois, Vestergaard, et al., 2015; Kiti et al., 2016). They found that the RFID signals, when configured to a low enough power level, are blocked by the water content of human bodies (Cattuto, van den Broeck, et al., 2010). This means that proximity is only detected in the direction a wearer's torso is facing, and is the basis for the claim that such RFID sensors measure *face-to-face interaction* rather than just proximity like Bluetooth. Measuring a construct of interaction rather one of proximity provides a major theoretical advantage, as the theorized causal effect of proximity on relationships is through making interactions more likely (Festinger et al., 1950), although I examine this closely below.

The sociometric badge's use of infrared light achieves the same end of limiting detections to the direction the wearer's torso is facing, although it does so by how infrared light is blocked by physical objects just like visible light. Also, testing showed individual detections to be fairly unreliable (Choudhury, 2004), and infrared has not been adopted in other sensor nodes still in use.

As one note on the use of mobile phones versus sensor nodes, I found in my own use of a mobile phone-based platform for data collection that the ability of programmers to write software that can access data from the sensors in mobile phones is increasingly being restricted. If and as this continues (and indeed, it is likely a good thing for protecting consumer data, even if bad for research), sensor nodes over which researchers can have complete control may again become a competitive choice for Bluetooth-based detection.

Lastly, audio recordings, as mentioned above, along with video recordings, are also 'sensor data.' Here, what is novel is not the sensor (as we have long had audio and video recordings), but models and algorithms that can potentially extract conversations (for audio) or interactions (for video) without the need for hand-coding, potentially making such measurements scalable in a way that human coding of recordings is not. Furthermore, the ubiquity of cameras and of microphones, such as in cell phones, raises the possibility that relational data collection can be done unobtrusively (as well as raises the danger that such collection can be done without participants' awareness or informed consent), although so far this has not been the case; successful cases of extracting relational data from video have relied on lab settings or other highly idealized backdrops where individuals are clearly visible, in high quality, against a plain background and with no potentially confounding other activities happening in-frame. Successful cases of audio processing have involved participants individually equipped with conspicuously placed boxes that include microphones along with other sensors (Basu, 2002; Wyatt et al., 2011), they have not been from audio recordings from cell phones in pockets or from ambient audio recordings. If participants were to be equipped only with microphones, such as clip-on microphones or as in ‘wearing a wire’, it would be less conspicuous but would still require participants to put it on. The burden involved with this makes longer-term data collection (over weeks, months, and longer) impractical.

3.2.3 Existing work

Overview

Out of the hundreds or perhaps thousands of papers that use sensors to study human behavior, there are upwards of at least a hundred that use sensors for social network data. Compared with work in social network analysis, these sensor works have generally not theorized the nature of the ties, and are published either in engineering and computer science venues, particularly around the subfields of *pervasive computing* and *ubiquitous computing*, or in general scientific venues like *PNAS*, *Nature*, and physics spinoff journals (*EPJ Data Science*, *Physical Review E*). Works in engineering and computer science venues focus on the sensor technology or on developing services (consumer or public) with that sensor technology. The works published in general venues are often of the ‘social physics’ variety (Hidalgo, 2016), involving univariate modeling of time series and distributions, and seeking to identify generative processes visible through a single variable or in a bivariate relationship. This is in contrast to social statistics approaches, which seek to identify variables that contribute to some outcome and model the relative ways in which they contribute to the outcome through multivariate statistical inference (Scott and Carrington, 2011; Borgatti, Mehra, et al., 2009).

In general, most of the studies are remarkable in terms of the technical effort involved in creating devices and/or setting up the data collection infrastructure (including also gathering survey and demographic data to compare to sensor data), data cleaning and processing. They have good study procedures with boundary specifications, participant recruitment and retention, and make innovative explorations of what sensor-based data may be compared to. My contribution is to try and sharpen the social science interpretation of what this work shows, and to make “a clear theoretical link between the questions we ask and the means of data collection” (Hogan et al., 2007). With a clear link, I also identify a set of unexplored possibilities of study with sensor data.

I also seek to identify the kinds of models for sensor data that are appropriate for certain goals. Taking the categories of statistical modeling described by Shmueli (2010) (echoing those of Breiman, 2001), models can either seek to be *explanatory*, which involves modeling the data-generating process to gain insight about a phenomenon represented in data (the type of modeling that has been almost the entirety of social statistics), or they can seek to be *predictive*, which seek to find the best-fitting model (for example, to use in engineering applications, and the almost exclusive focus in machine learning). For models that seek to be predictive, it is not necessary to have theoretical clarity about constructs, nor to have statistical models that respect the nature of the data; all that matters is establishing external validity.[4] However, rigorously demonstrating external validity in this setting is nontrivial, and requires validation via methods like held-out data, with splits in data carefully chosen. Furthermore, for binary response variables (as the presence/absence of network ties are), neither accuracy (the number of correct predictions normalized by the number of cases) nor variance explained (via pseudo-R^2 metrics) are meaningful ways to judge the predictive performance of models in highly 'imbalanced' cases (again, as the presence/absence of network ties usually are); for a network with a density of .05, a trivial classifier that always predicts non-edges would have 95% accuracy, but would be useless for any practical use (called the "accuracy paradox"). Precision (the positive predictive value, number of true positives over total predicted positives) and recall (the true positive rate, number of true positives over total number of positives), among other metrics, are more meaningful.

For works that seek to explain something about the underlying system, statistical models should seek to model the networked nature of data. Conversely, for models that make claims about goodness-of-fit or predictive performance, proper validation techniques should be used. I further discuss and demonstrate such techniques with my own study.

There are also ethical concerns about the use of sensor data; studies so far (including my own), aside from being careful to obtain informed consent, have been with relatively privileged populations for whom the risks that come with being surveilled are relatively low. But as the use of sensor data expands, there are distinct dangers, which I will discuss in the 5.7.

Projects

Above, I introduced some of the noteworthy projects studying social networks with sensor data: several from the MIT Media Lab's Human Dynamics group (Reality Mining, SocialfMRI/Friends and Family, Social Evolution), the Copenhagen Networks Study, and the work of the SocioPatterns group. I now describe these further.

The most ambitious early sensor data collection was the Reality Mining study, where researchers gave mobile phones (as this was 2006, smartphones did not yet exist, the phones were Nokia 3650 which have a dialpad and a small LCD screen) to 100 participants, 75 of which were MIT Media Lab students or faculty, and the remainder were incoming MIT Sloan students. Tracking software developed at the University of Helsinki collected, over one academic year, "call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle)," which was combined with "web-based surveys regarding their social activities and the people they interact with throughout the day," along with a

[4] It is not intuitive that models do not predict well can offer meaningful explanations, nor that models can predict well without being faithful to the data-generating process; the recent development, especially within nonparametric statistics and machine learning, of "black box" models do exactly the latter. As for whether prediction alone is a useful task, or in what circumstances it is useful, there is a large debate, especially when it comes to social science and public policy, where (unlike for mechanical or physical systems), what is ultimately desired is usually *interventions*, which require knowledge of causality or at least true (non-spurious) associations (or, in statistical and econometrics terms, unbiased parameter estimation in correctly specified models rather than models that minimize over the bias-variance tradeoff). See Athey (2017), Cohen and Ruths (2013), Gayo-Avello (2011), Gayo-Avello (2012), Hofman et al. (2017), Hindman (2015), Lin (2015), Kleinberg et al. (2015), Mullainathan and Spiess (2017), and Wallach (2018).

survey taken two months into the study asking participants “who they spent time with, both in the workplace and out of the workplace, and who they would consider to be within their circle of friends.”

This resulted in a number of papers. One was the spectral clustering described above. Another, Eagle, Pentland, and Lazer (2009), was the first to compare mobile phone data proximity self-reported friendship ties. Eagle, Pentland, and Lazer (2009) first calculated a ‘probability of proximity’ score over the range of a week as an average frequency of proximity over nine months of data. They then gave plots showing systematically different patterns for each of reciprocated self-reported ties $(A_{ij}, A_{ji}) = (1, 1)$, non-reciprocated ties $(A_{ij}, A_{ji}) \in \{(0, 1), (1, 0)\}$, and no ties $(A_{ij}, A_{ji}) = (0, 0)$. For making a predictive model, they reported using MRQAP (although with few details), and reported that they could use the probability of proximity to detect ties with 95% accuracy.

The study was critiqued by adams (2010) [sic] on the grounds first that there are “close strangers and distant friends”, such that the task of using proximity to predict friendship lacked theoretical clarity. Second, adams noted that while the model fit well when looking at individual ties, the network-level properties of the networks formed from each set of ties were systematically different (demonstrated using hypothesis testing with draws from a Bernoulli random graph to generate distributions). Indeed, one of the disadvantages of MRQAP (and any nonparametric permutation test generally) is that they can control for network dependencies, but cannot substantively model them, although I would also note that Bernoulli random graphs are also generally a poor null model.

In response, Eagle, Clauset, et al. (2010) argue that some causal network processes might happen unconsciously, and sensor measurements might be able to detect these. Although this was not a motivation for the initial paper, it is certainly true. I will return to this debate, as the issues it raises are major themes in my argument for theoretically appropriate uses of sensors data.

There are also a number of problems in this study from a machine learning perspective, which I will also discuss below.

The Reality Mining data were also later used to look at obesity and exercise in the presence of contact between people (Madan, Moturu, et al., 2010a), but with basic hypothesis testing that did not control for network processes. Similarly, Madan, Farrahi, et al. (2011) used the Reality Mining data to model adoption of political opinions based on the opinions of others who were proximate, although in a simple linear model with political opinion as a dependent variable, and network properties as independent variables (not considering the possibility of co-evolution). Staiano et al. (2012) used ego networks in call logs, Bluetooth-based proximity networks, and surveys to predict Big-5 personality traits; they used triadic and transitivity measures of ego networks, although modeling ego networks is not statistically ideal because of how such networks are not independent.

Lastly, Dong et al. (2011) modeled the co-evolution of location behavior and social relationships from the Reality Mining data, employing a Markov jump process model that they remarked had similarities to temporal Exponential Random Graph Models (tERGMs). However, this similarity was mainly the in exponential-family form, their sufficient statistics did not include network-level measures. And the results of their model were unclear: they did not use usual network goodness-of-fit metrics, instead visually presenting the predicted and actual adjacency matrices and qualitatively arguing for their similarity. The only performance metric reported was that the binomial model explained 22% of overall variance, of which 6% was due to sensor data. This presumably from a pseudo-R^2 metric, but the specific metric is not given.

The Reality Mining study was followed up a few years later by the ‘Friends and Family’ or ‘SocialfMRI’ study and dataset (Aharony et al., 2011), also from the MIT Media Lab. In contrast to Reality Mining, there was a better-defined boundary specification, 130 adults in a “young-adult living community”, and in the intervening years the smartphone market had emerged, such that the study was able to provide Android smartphones to the study participants. For one year, tracking software collected data of “location, accelerometry, Bluetooth-based device proximity, communication activities, installed applications, currently running applications, multimedia and file system information”, which were combined with financial information (receipts and credit card statements), Facebook activity, “daily polling of mood, stress, sleep, productivity, and socialization, as well as other health and wellness related information, standard psychological scales like personality tests, and many other types of manually entered data by the participants.”

These data have been used to look at the connection between interaction and financial status (Pan et al., 2011) and interaction and sleep and mood (Moturu et al., 2011). Again, the models employed were of basic hypothesis testing.

A related project, although with fewer associated publications, was called the ‘Social Evolution’ project. Out of a dorm of 70 undergraduate students, the study enrolled 80%, reporting that the remaining 20% were “spatially isolated” (Madan, Moturu, et al., 2010b). Students had self-selected into the dorm, were predominantly male (54%), and were about equally split among freshmen, sophomores, juniors, and seniors. Collected data was a combination of monthly self-reported surveys, and records from cell phones of Bluetooth, sampled every 6 minutes, along with communication logs. These data were used to model influence, specifically. Madan, Farrahi, et al. (2011) create a composite measure of ‘dynamic homophily,’ calculating exposure (via proximity and communication) to people with the same or different political orientations. Using this exposure as a covariate, they build a simple linear model of opinion change, finding the effect significant. Similarly, Madan, Moturu, et al. (2010b) find associations between exposure to peers who gained weight and weight gain, again in a regression model. While the research questions were meaningful, the models employed used summaries of network effects rather than directly modeling networks and controlling for various effects. They acknowledge limitations of a small sample size, but surprisingly, neither Madan, Moturu, et al. (2010b) nor Madan, Farrahi, et al. (2011) consider that causation might go from political opinions or weight gain respectively to exposure, as later explicitly articulated in Shalizi and Thomas (2011). This is especially surprising considering that Madan, Farrahi, et al. (2011) has an explicit discussion of the causal effects of homophily; but in the remainder of the paper, homophily is treated as an index whose value is caused by interaction, rather than a causal process in itself. The most intriguing modeling proposal from the Social Evolution data is that of Dong et al. (2011), who set out to model the co-evolution of location behavior and communications data using a Markov jump process that models changes in location based on interactions. However, their results consist of summarizing the ‘proportion of variance’ in self-reported friendship ties explained by shared dorm, shared year, change in friendships, and sensor data. It is unclear what the proportion of variance explained would be, since logistic regression is associated with an analysis of deviance, and pseudo-R^2 metrics rather than variance explained. Usual metrics of precision and recall were not provided, although the paper did note that the base rate was 95% (i.e., 95% of self-reported ties were non-friendships). They note a link between their Markov jump process and temporal exponential random graph models, but it was unclear if and how their model incorporated network topology (including dyad-dependent effects).

The Copenhagen Networks Study picks up where some of the MIT work left off. This gathered Bluetooth proximity data and WiFi logs from 134 students over 119 days in 2012 and 2013 (Stopczynski, Sekara, et al., 2014), in addition to 95 questions from psychology survey instruments given to each participant in 2012 and 310 questions in 2013. Facebook networks were collected for those who opted in. The 2013 deployment also included an anthropological field study, with an anthropologist embedded with 60 participating students to get both qualitative data and feedback about participation. The initial publication (Stopczynski, Sekara, et al., 2014) was focused on describing the infrastructure and giving descriptive results. Stopczynski, Sapiezynski, et al. (2015) considered the time scale at which the sensor data should be modeled, Sekara and Lehmann (2014) considered Bluetooth RSSI as a proxy for distance, and Mones et al. (2017) and Mollgaard et al. (2016) compared properties of communication networks and proximity networks. However, there have not yet been multivariate models of network formation, or models looking at the co-evolution of networks and some node covariates of interest.

The SocioPatterns (Cattuto, van den Broeck, et al., 2010) group has a number of studies, and depart from other major works in using RFID badges instead of mobile phones. These deployments have varied from 2 days to 6 weeks, and most frequently are set at academic conferences, schools, or hospitals Barrat and Cattuto (2013). Modeling here focuses largely on characterizing distributions and time series (Barrat, Cattuto, Colizza, Gesualdo, et al., 2013; Barrat and Cattuto, 2013; Isella, Stehlé, et al., 2011; Kiti et al., 2016; Panisson et al., 2013; Starnini, Baronchelli, et al., 2016), and of modeling possible transmission of infections on collected networks (Barrat, Cattuto, Colizza, Isella, et al., 2012; Barrat, Cattuto, Tozzi, et al., 2014; Vanhems et al., 2013; Ciavarella et al., 2016; Fournet and Barrat, 2016; Fournet and Barrat, 2017; Gemmetto et al., 2014; Génois, Vestergaard, et al., 2015; Isella, Romano, et al., 2011; Mastrandrea, Soto-Aladro, et al., 2015; Starnini, Machens, et al., 2013; Stehlé, Voirin, Barrat, Cattuto, Colizza, et al., 2011; Voirin et al., 2015). The first sensors work published in *Social Networks*, is from this group, Stehlé, Charbonnier, et al. (2013), who look at the connection between gender homophily and 'spatial behavior', measured by the RFID badges, by way of a 'shuffle test' to generate a null distribution. There has also been more methodological work from the SocioPatterns group as time has gone on, explicitly comparing different forms of data collection (Génois and Barrat, 2018; Mastrandrea, Fournet, et al., 2015), which I discuss further below. Works from SocioPatterns also include the only two uses of Stochastic Actor-Oriented Models (discussed further below, under 'Models') with sensor data: Pachucki et al. (2015) deployed the RFID badges in sixth-grade classroom, and compare sensor data to both demographics and mental health outcomes in a SAOM. Eberle et al. (2017) apply a SAOM to a multidisciplinary scientific conference to look at the relationship between that data, discipline, and career level. While such modeling is the most interesting way to use sensor data from a social science perspective, the ways in which these two studies summarized sensor data in order to put into the model are unsatisfactory. I discuss this further below, including taking up an alternative approach (which is also the topic of Chapter 5).

3.2.4 Interaction or proximity?

The quantity that sensors supposedly measure are referred to many different ways, even for the same sensors and even for different papers from the same group. For example, from the Bluetooth in the Copenhagen networks study,

- “Proximity data” (Sekara and Lehmann, 2014)
- “Face-to-face interactions” (Stopczynski, Sekara, et al., 2014)
- “Close proximity interactions” (Stopczynski, Sapiezynski, et al., 2015)
- “Face-to-face contacts” (Mollgaard et al., 2016)
- “Physical contacts” (Mones et al., 2017)

I also note that earlier versions of Sekara and Lehmann (2014) and Sekara, Stopczynski, et al. (2016), posted on arXiv, frame Bluetooth as measuring “face-to-face interactions,” but the published paper only mentions this construct once, and even then in reference to SocioPatterns.

Across SocioPatterns work with RFID badges, the terms used are

- “Person-to-person interaction” (Cattuto, van den Broeck, et al., 2010)
- “Face-to-face contacts” (Barrat, Cattuto, Colizza, Isella, et al., 2012)
- “Close-range interactions” (Cattuto, Quaggiotto, et al., 2013)
- “Face-to-face interactions” (Barrat, Cattuto, Colizza, Gesualdo, et al., 2013)
- “Face-to-face proximity” (Barrat, Cattuto, Tozzi, et al., 2014)

This inconsistency suggests a lack of theoretical clarity. I address this in five clarifying points.

First, ‘interaction’ is ambiguous. If we had human coders and were establishing a codebook, we would need to decide whether or not a passing “hello” is an interaction, or whether an interaction requires some minimum length of time. Or, if interactions were coded by length of time, should it still be binary? Should an intense, one-on-one personal conversation for two hours be coded the same way as superficial interaction as part of a a boring two-hour group meeting, or should there be coding for different types of interaction? In a discussion of coding for *conversation*, Wyatt et al. (2011) have a similar discussion of the ambiguity:

> “For example, imagine two officemates *a* and *b* who work mostly in silence for two hours while occasionally talking. *a* makes a comment, *b* responds, and a short exchange ensues before they fall back into silence. When does the conversation start and when does it end? If *a* makes a comment later but *b* does not explicitly respond, is that a conversation? If a third person *c* enters the room and speaks to *b* but only *a* responds, who is in conversation with whom?”

Alternatively, we could try to define an objective notion of interaction via criterion-related validity (Babbie, 2010), i.e., based on a measure of interaction consistently predicting a certain outcome. But this requires a subjective choice of an outcome of interest, and different outcomes of interest may lead to interaction being characterized in different ways.

Second, depending on the phenomenon we are interested in studying, ‘interaction’ may not be precise enough or even the right construct. For example, if we are studying disease transmission, we might care about interactions through physical contact like a handshake, no matter how brief and whether or not there was conversation involved. And for certain types of infections, we would not be interested in interaction at all but whether a person walked through an area where an infected other individual sneezed or coughed up to 45 minutes ago (Johnson, Knibbs, et al., 2016). For looking at persuasion, interaction through *in-person conversation* would be measurement of interest. For studying common environmental exposure or latent homophily expressed geographically, proximity (at ranges further than face-to-face) should be the target of measurement. Studying memorability in something like a marketing context may require only line-of-sight visibility, or in the context of establishing future scientific collaborations, memorability would be better tied again to in-person conversation. For studying loneliness and social isolation, it would be a topic of investigation whether close-range proximity to others is sufficient to affect loneliness, or whether conversation is necessarily (and whether the existence of conversation is sufficient, or if the content of the conversation matters).

In fact, the SocioPatterns did not originally set out to study social structure; they originally conceived their RFID badges as for measuring interactions in the context of disease transmission. But in the early studies, they realized that the badges were effective at capturing *social* interactions (Mathieu Génois, personal communication, July 1, 2018) which led to studies using the badges for this purpose.

Third, the construct of interest is not face-to-face interaction, but *in-person* interaction. There are no causal processes that specifically require interaction to have two people’s torsos facing one another. When using face-to-face interaction as a proxy for in-person interaction, false negatives could result from people walking side-by-side as they talk, or interacting while sitting in a car where everybody is facing the same direction.

Fourth, the measurement of RFID sensors are not of face-to-face interaction, but of face-to-face *directional proximity*. In using directional proximity as a proxy for interaction, false positives could result from people siting across from each other on a subway car without interaction, or from squeezing into a cramped elevator.

We can imagine that, in such edge cases, human coders would correct label face-to-face non-interactions as non-interactions, and non-face-to-face interactions as interactions. Human observation is capable of more directly observing the multimodal forms of interaction, from conversation to physical touch to body language and microexpressions, that could make up a meaningful overarching definition of interaction. For this reason, there is a need to compare RFID to human coding, work which is only now being done.

In one forthcoming work, Génois et al. find that the confidence interval of RFID detections are contained completely within the confidence interval of inter-annotater agreement. This suggests that human observers may suffer from inconsistencies or biases that are greater in magnitude that those of face-to-face directional proximity for measuring interaction. But in an independent study also comparing human coding of video recordings to RFID, Elmer et al. (Under review) find that RFID badges do not detect all interactions, having a sensitivity (i.e. recall, or true positive rate) of only 50%. However, merging detections within 75s of each other, they find, raises the sensitivity to 65%. This suggests that quality of human coders may be an issue, or again potentially the specificity with which how human coders define an interaction in a given setting.

Further work can focus more on the edge cases where human coders and RFID disagree, as understanding edge cases is essential for proper use of a measurement apparatus. For example, in settings like academic conferences, most in-person interactions will involve two people's torsos facing one another and vice versa, such that there should be few false positives or false negatives due to the discrepancy between the construct and the measurement. However, on road trips in a car or morning commutes on subways, we would avoid using directional proximity as a proxy for measuring in-person interaction. Alternatively, human observation is limited to line-of-sight (whether for in-person observation or in coding video recordings), and may involve difficulties when trying to distinguish pairwise interactions in the midst of large groups.

One clear place where the validity of RFID has been established is in comparison to contact diaries. This is investigated by Elmer et al. (Under review), who find high correspondence between RFID and self-report, as well as two SocioPatterns works. The SocioPattern works, Mastrandrea, Fournet, et al. (2015) and Smieszek et al. (2016), have specific findings: compared RFID data to contact diaries, both works found that short detections were not recorded in contact diaries, and contact durations recorded in diaries were longer than sensor-detected durations. This is again potentially an issue with the construct of interest; for processes like memorability or persuasion, contact diaries are potentially a better instrument. But for looking at disease transmission, or for seeing if forms of influence can happen without people being consciously aware, RFID would be the superior instrument. In another critical consideration, Mastrandrea, Fournet, et al. (2015) also note that contact diaries had much higher nonparticipation and dropout. The lower participant burden associated with RFID badges, once they are equipped, may prove decisive in favor of RFID even if the process of interest is memorability or persuasion.

Alternatively, rather than comparison, Pachucki et al. (2015) use self-reported friendships to *calibrate* sensor data. They determined that the correlation between friendship ties and sensor-based measurements rose when using a minimum threshold of 60s for RFID sensor data (the sensors sampled every 20s), and in their supplementary information they elaborate that they ultimately "decided to use the 80-second threshold for RFID network enumeration because across all three observation periods, it appeared to have the greatest overlap with student self-report on the basis of a number of metrics (% overlap of dyads in each network, density, mean degree)." This is likely an effective heuristic for dealing with sensor data, but it is theoretically unclear, since friendship surveys and sensors are measuring two fundamentally different constructs (proximity, or directional proximity, and sociometric choice). Indeed, it is worth modeling the relationship between these two constructs, which is the focus of my analysis below.

Fifth, Bluetooth sensors capture proximity only. On the level of face validity, Choudhury (2004) and Olguín et al. (2009) argue for how infrared captures interactions over just proximity, as do Cattuto, van den Broeck, et al. (2010) for RFID, but there is no such argument for Bluetooth. In terms of construct validity, now that work has established that RFID sensors are a valid proxy for measuring interactions, we can rely on the results of Génois and Barrat's (2018) comparison of Bluetooth and RFID data. Bluetooth, which captures (non-directionally constrainted) proximity, gives far more pairwise detections than does RFID. Génois and Barrat (2018) propose several downsampling methods that can extract data corresponding to RFID measurements, focusing on reproducing the statistical properties of the RFID networks (i.e., they treat it as a network goodness-of-fit problem rather than a classification problem of predicting the value of RFID dyads with Bluetooth data). Some properties they can recover moderately well, but across different tested settings, the success of various sampling strategies vary widely, suggesting that Bluetooth is not effective for measuring in-person interaction.

Sixth and finally, in-person interaction is not the only construct of interest. In the Supplementary Information of Sekara, Stopczynski, et al. (2016), the authors write:"Bluetooth scans do not constitute a perfect proxy for face-to-face interactions. In fact multiple scenarios exist where people in close proximity do not interact and vice versa, nevertheless Bluetooth can successfully be applied in order to sense social networks." This implies that Bluetooth detections are valuable only insofar as they capture interaction, but this is not the case. As argued above, for studying geographically expressed latent homophily, and common environmental exposure, the construct of interest is proximity. Bluetooth, and indeed other sensors that measure co-location (WiFi, GPS, cell towers) are ways to measure proximity. Note also that in Festinger et al.'s (1950) theory of propinquity, the way in which proximity is causal is in giving opportunities for interaction. This suggests possibilities for using Bluetooth or other proximity-measuring sensors in conjunction with RFID sensors or human observation to build a more complete picture not only of interactions, but of the background opportunity structure created by proximity.

I have focused primarily on RFID and Bluetooth, but there is also another promising option that has not received much recent attention: the use of audio data. If interaction is the construct of interest, on the level of face validity, audio would avoid the false positives and negatives found in RFID measures of directional proximity, although audio may have false negatives associated with nonverbal interactions. And, for persuasion or influence, causal processes associated with interaction happens through conversation, making conversation itself the construct of interest. As Wyatt et al. (2011) argues, conversations as recorded in what they call "situated speech data" (i.e., in-person conversations occurring naturalistically) are a more effective way of describing social systems than is co-location.

Detecting conversations from audio data involves each participant wearing a microphone, and then recording both the wearer's speech as well as that of others with whom the wearer converses. This redundancy helps in noisy settings. The dyadic entity of a 'conversation' is extracted through modeling 'turn-taking' between speakers, and in the case of both Basu (2002) and Wyatt et al. (2011), require some initial human labels to calibrate models. Recognizing the ambiguity of what should count as a conversation, (Wyatt et al., 2011) set out criteria that coincide perfectly with the labels of human coders.

Surprisingly, there has been little follow-up to this work, taking up the use of audio to measure social systems. There undoubtedly needs to be more exploration of the use of audio, although there are limitations. Perhaps the greatest limitation is that recording people's actual conversations is extremely intrusive. Basu (2002) proposed an ingenious solution: scrambling audio in a way that disguises the content of speech, while retaining aspects of the audio signal that can be used to uniquely identify the speakers. This is a feature that remains in sociometric badges. Similar privacy-preserving measures were taken up by Wyatt et al. (2011). However, to be truly privacy-preserving, such processing needs to happen at the point of contact, which creates an additional computational burden (personal communication, James A. Kitts, July 1, 2018). Audio data is also extremely large; indeed, the instrument used in Wyatt et al. (2011) was a bulky shoulder-worn badge connected to a large PDA in a bag. However, this was from multiple sensors being bundled together (Ibid.). If only audio is of interest, an external microphone worn by participants could take the place of the bulky shoulder badge, and a modern smartphone would be a far smaller device than the PDA that could do the privacy-preserving processing, although audio storage may still be an issue depending on the quality of stored audio data. Overall, the setup would be as lightweight for participants as are RFID badges. While audio from smartphone microphones has been used alongside other sensor data (location, accelerometry, temperature) for determining whether two phones share an overall context (de Freitas and Dey, 2015), to my knowledge, there has not been work that has continued this idea of using audio for detecting when two specific individuals are conversing via smartphones and external microphones. This remains a promising avenue for future research.

Note that if study participants are willing to let their conversations be transcribed, then the dyadic interactions can be characterized in far more detail. Work by Ranganath et al. (2009) and Ranganath et al. (2013) (Jurafsky et al., 2009) looked at 1000 recordings of 4-minute speed dates, where the participants rated each speed date in terms of their own and their partner's behavior along 'conversational style' dimensions of friendly, flirtatious, awkward, and assertive. The recordings were transcribed, from which lexical, prosodic, and other features were extracted from transcripts to build a classifier for the conversational styles. But people are less likely to give permission for recording entire conversations outside of such formulaic interactions, i.e., it is likely less feasible to collect the content of situated speech data.

3.3 Theory

Looking at what sensors actually measure sets up one of my main goals: theorizing the nature of sensor data, and articulating the kind of investigations that are appropriate with such data.

3.3.1 Relational phenomena

As with any discussion of different methods of social network data collection, I start with Bernard, Killworth and Sailer's landmark series of papers (Killworth and Bernard, 1976; Bernard and Killworth, 1977; Killworth and Bernard, 1979; Bernard, Killworth, and Sailer, 1979; Bernard, Killworth, and Sailer, 1982) on the problem of "informant accuracy." Relevant here is that they used various 'objective' forms of data as the baseline or "gold standard" (Marsden, 2011) against which to compare survey data: communication (logs of HAM radio communication) or interaction (from human observers making observations at 15 minute intervals). Similarly, sensors give a type of 'objective' data: but just as in Freeman et al.'s (1987) reexamination finding long-term consistency in self-report, and more importantly as Krackhardt's (1987) argument about survey instruments capturing psychological perceptions that may be more important than objective data in explaining how people act, objective data may not be what we want for certain applications.

Such theoretical links have only begun to appear in sensors literature, which so far has not been familiar with the notion of a 'social network' as a category of representation, rather than narrowly just as ties related to socializing. Barrat, Cattuto, Colizza, Gesualdo, et al. (2013) argued that what sensors measure are not 'bona fide social networks,' suggesting calling them *behavioral social networks* instead. But as Borgatti, Mehra, et al. (2009) point out, social network analysis is concerned with dyadic phenomena more generally, and it is not only social relations (such as kinships, friendships, hierarchies, affect, and knowledge of) that form networks, but also similarities (such as location, membership, and attributes), interactions (such as advice, help, talking, sex) and flows (such as information, beliefs, and resources). Within this, similarity in proximity, or interactions of conversation, are social networks as much as are networks of friendship—or at least do not need to be theorized as a new form of dyadic relationship.

Another theorizing of social networks that helps illuminate the appropriate place of sensor measurements is Kitts and Quintane (forthcoming), who identify four *conceptualizations* of networks: networks as role relations (socially constructed categories) like friendship, marriage, kinship, co-membership, coauthorship, etc.; as behavioral interactions like face to face conversation, sex, money lending, citations, phone calls, texts, online messaging, emails, etc.; as interpersonal sentiments (social evaluations) like liking, disliking, loving, hating, trusting, distrusting, respecting, etc.; and as opportunity structures like access to exchanges, access to information, or access to support (even if such access is not actually used).

Propinquity

The first key to how to theorize sensor data lies in the insight, from the work on informant accuracy and subsequent debates, that objective data does not access psychological states. Sensors are thus appropriate for studying physical phenomena, but for studying processes where causality happens through perceptions and psychological states, they can only be a proxy.

The second key is related to my argument above, that what sensors actually measure is *proximity*. It may be possible to constrain the range and direction of detection (as with the RFID sensors used by the SocioPatterns group), and/or to process the data (e.g., with thresholds on signal strength or detection length), to use proximity as a proxy for interaction.

The third is to look to Borgatti, Mehra, et al.'s (2009) argument about the rich possibilities that come with studying the connection between different types of dyadic relations, along with Kitts and Quintane's (forthcoming) categories of networks as conceptualizations of behavioral interaction, role relations, interpersonal sentiments, or opportunity structures.

Putting these together, I propose that the most theoretically appropriate and promising use of sensors is for *studying the relationship of interaction and proximity to each other and to other processes of interest*.

I take *propinquity* as a central example of a social theory that can guide questions we ask with sensors. Propinquity was theorized in the foundational work of Festinger et al. (1950), also known as the 'Westgate study'. Leon Festinger and colleagues used new, relatively isolated dorms built at MIT to house the influx of graduate students returning to school on the GI bill as an opportunity for a natural experiment: given random assignment to units in the Westgate complex, how did friendships form? The close relationship between living close by, or even just passing one anothers' houses during daily commutes, led to the theory of *propinquity*: proximity gives opportunities for interaction, and repeated interaction can lead to friendship. In the conceptualization of Kitts and Quintane (forthcoming), propinquity is the process of proximity being an opportunity structure for the behavior of interaction, which leads to the construction of role relations of friendship.

The 'Newcomb-Nordlie fraternity' study (Nordlie, 1958; Newcomb, 1961), also known as the 'Michigan Group Study Project', provides some details about the link between interaction and role relations: here, proximity was effectively held constant, to show the relationships between other factors of demographics, attitude, and network processes. In 1954 and 1955, the study recruited two waves of 17 male American undergraduates entering the University of Michigan, chosen to not know each other and to have geographic and religious variance (although explicitly not racial or other types of variance). These men were given free ‘fraternity-style’ housing, and the researchers studied how their personal characteristics, political positions, and interests affected their eventual friendship formation. Some of the findings were that ‘interpersonal attraction’ became more stable as time went on, and that the cohorts could be divided into subgroups with high inter-group attraction and low intra-group attraction (i.e., ‘community structure’ in the graph).

Both of these are examples of basic research, and indeed, propinquity as a contributor to friendship formation is of inherent interest in social psychology and sociology (Fehr, 1996). But there are important questions in application settings, for example for policy and public health. One aspect of the debate over the Framingham Heart Study (Wasserman, 2013) was precisely whether the causal mechanism was social influence as the original study claimed or, as Cohen-Cole and Fletcher (2008) proposed and argued for as a more plausible alternative explanation, environmental factors were the cause. When trying to design interventions around health, it is important to know whether environment or social influence is causal (or whether the cause is simply latent homophily; Shalizi and Thomas, 2011). Otherwise, network-based interventions are misguided and risk being ineffective and wasteful (Aral et al., 2009). Fine-grained measurements proximity would help identify possible common exposure, and along with knowledge of some form of social ties along which influence could happen, such measurements could disentangle causality for outcomes under high correlation of friendship and proximity.

Surprisingly, the only sensor studies that have connected to any of this work are those of Oloritun et al. (2012) and Oloritun et al. (2013), who cite Festinger et al. (1950). However, these two papers do not distinguish between proximity and interaction, and both use Bluetooth, aggregated with simple summation and thresholding,

3.3.2 Models

Using large amounts of data for specific research questions requires explanatory modeling, and explanatory models should respect the theorized data-generating processes.[5] Neither the Westgate study nor the Newcomb-Nordlie fraternity study had the available statistical machinery or computational capacity to perform statistical inference on a multivariate model that incorporated network covariates and that looked at the co-evolution of networks and opinions/behaviors (allowing the network to be both a dependent and an independent variable). Thus, while these works established major theoretical directions, their specific numerical findings do not control for the multiple factors (both network and non-network) that we observe in network evolution, and thus are not directly usable.

Fortunately, Stochastic Actor-Oriented Models (SAOMs; Steglich et al., 2010) provides exactly such machinery for statistical inference, and is an appropriate modeling end goal (although I suggest there are some necessary intermediate modeling steps, which I demonstrate). Stochastic Actor-Oriented Modeling was first realized in the SIENA (Simulation Investigation for Empirical Network Analysis) framework, which is a set

[5] Here I consider only statistical modeling, as it can directly take in data and give interpretable parameters or predictions, as compared to simulation modeling or other types of conceptual mathematical models that we only generate outputs from and qualitatively match to observed data.

of functional forms (relating to networks observed at discrete intervals), network terms representing sociological processes, and estimation procedures, implemented in the `RSiena` software package (Ripley et al., 2017). Duijn et al. (2003) used SIENA on data gathered from a retreat for all incoming sociology majors at the University of Groningen; these students did not know each other prior to the retreat, providing an opportunity to study friendship formation. They were able to quantify how friendships developed due to a mixture of four main effects: physical proximity, visible similarity, invisible similarity, and network opportunity.

For much survey data or other data longitudinally collected at discrete intervals, the SIENA model is appropriate. Alternatively, if our goal is to simply model the internal dynamics of an evolving network with the use of known sociological processes (e.g., transitivity, popularity, etc.), or if the process of interest is some other densely measured process (potentially something else measured in sensor data, like activity recognized from accelerometer data, or other biometric readings), then we may use a Relational Event Model (REM; Butts, 2008), which combines parameterizations of event modeling (e.g., using an exponential distribution to parameterize the waiting time between events) and network models (using the Exponential-family Random Graph Model framework) in order to model timestamped sequences of relations. This is also the recommendation made by Kitts and Quintane (forthcoming). Additionally, a recent analog of REMs that is more explicitly an Actor-Oriented Model, called Dynamic Network Actor Models (DyNAM; Stadtfeld and Block, 2017; Stadtfeld, Hollway, et al., 2017) may be used. Currently, there are explorations of DyNAM developed specifically for use with sensor data (Timon Elmer, personal communication, July 4, 2018).

Currently, I have only identified two studies that use a SAOM with sensor data, both from the SocioPatterns group and both using SIENA: Pachucki et al. (2015) and Eberle et al. (2017). Eberle et al. (2017) used the duration of RFID detections as an edge covariate, and collaborations as the network. Self-reported previous collaborations formed the first wave, and self-reported "potential for future collaboration" formed a second wave. Pachucki et al.'s (2015) operationalization of RFID sensor data is intriguing: as discussed above, they used self-reported friendship data to *calibrate* the sensor data, finding a high correspondence between friendship networks and the sensor network thresholded at detections of 80s. Friendships are directed ties, and RFID detections are undirected; but in their supplementary information they discuss how one of the SIENA model types for undirected networks (Snijders and Pickup, 2016) was a perfect fit for modeling interactions. Specifically, they used the "unilateral initiative and reciprocal confirmation" model type in RSiena, where the dynamics are that "one actor takes the initiative and proposes a new tie or dissolves an existing tie; if the actor proposes a new tie, the other has to confirm, otherwise the tie is not created; for dissolution, confirmation is not required" (Ripley et al., 2017). Pachucki et al. (2015) used the analog of interacting through sitting together at a lunch table: permission must be given for a person to join others at a lunch table, but a person can get up and leave without permission.

However, there is a mismatch in timescale; there would be many 'interaction ties' with the dynamics described above over the course of a day, whereas they used ties representing periods of several days. Specifically, "Social ties from Days 1-3 of the first week were then aggregated to constitute Period 1; days 4-6 comprised Period 2; and days 7-9 comprise Period 3". Whatever is represented by this aggregation is potentially no longer appropriate to model as a "unilateral initiative and reciprocal confirmation" dynamic.

This points to a larger challenge: sensor data is incredibly dense in time, whereas many processes of interest (especially those collected through survey instruments), for which SIENA is used, are measured at far more sparse intervals. In Pachucki et al. (2015), the sparse survey data were node covariates and the sensor data were the network ties, whereas in other situations (including my own study), the sensor data give node and/or edge covariates and ties come from survey data, but the problem of a mismatch in timescale is the same. Not every outcome of interest can be measured at a dense enough resolution that we can use an REM or DyNAM.

In such cases, the question becomes how to summarize the sensor data into a form commensurate with the other variable(s) of interest. Most studies that face this problem take some form of simple aggregation, such as counting the number of detected interactions, or the total length of interaction, and then thresholding to form a binary indicator for a network tie. Pachucki et al.'s (2015) aggregation, described in their supplementary information, is "A given day's 80+ second ties are then aggregated with the next two days' 80+ second ties to obtain a cumulative 3-day weighted network, which represents a given Period's binary matrix" (they do not specify what thresholding on the weighted network they used to binarize edges). Eberle et al. (2017) "considered only pairs of individuals with a total measured interaction time [aggergation] of at least 100 seconds [threshold] during the total event".

But why should an aggregation and then a binary thresholding be the correct way to summarize the distribution of sensor measurements over time? Assuming that we measure interaction, perhaps it is the number of *repeated* interactions over some period of time that matters. Perhaps the times at which interactions occur, e.g., earlier in the day versus at lunchtime versus at the end of the day, are significant. Here, the richness and density of sensor data overwhelms theoretical guidance: we have some ideas, such as Eagle, Pentland, and Lazer's (2009) observing that time spent together during the daytime is systematically different for friends and non-friends, and reciprocated friendships from non-reciprocated friendships, a competing suggestion from Oloritun et al. (2013) that the duration of proximity on *weekend nights* is most important, and we have the suggestion from Latané et al. (1995) that inverse-squared distance may transform distance into something with a linear relationship with survey data. But if we consider different weight thresholds for time or distance, summaries over specific times of day, distributions of *lengths* of detections and of gaps between detections, and different functional forms (e.g., logarithmic or other variance-stabilizing transformations for heavy-tailed distributions), we arrive at hundreds or even thousands or possible measures that might be the best summary of sensor data for some outcome of interest.

Machine learning

Methods to address this task can be found within machine learning. Machine learning frequently deals with data sets where some signal to be used as a predictor has measurements that are far more dense than the response. The process of finding meaningful summaries is called *feature extraction* ('features' are the machine learning term for covariates or predictors). There are few cases where feature extraction can be done systematically; the typical approach is to heuristically generate a huge set of features, and use feature selection (i.e., variable selection) methods on the generated set (Christ et al., 2016).

The feature selection methods used by machine learning draw on multiple sources, both those with statistical justification like the lasso (Tibshirani, 1996; Tibshirani, 2011) or stability selection (Meinshausen and Bühlmann, 2010), as well as methods from computer science literature on data mining and pattern recognition that may not have such statistical exploration of their theoretical properties, but that are demonstrated to work on specific data sets and in specific use cases.

In order to manage issues like the multiple comparisons problem, machine learning relies on *cross-validation* rather than on tests of significance for determining success. In theory, if a model can do well on predicting the outputs (e.g., the class labels, or the response variable value) of unseen, out-of-sample data, then this directly establishes external validity, bypassing any concerns about any other type of validity like internal or construct. As it turns out, it is possible to get external validity without interval validity, or in other words, models that predict well on the basis of spurious correlations (or of biased estimates, or of other associations that are not causal, or of modeling assumptions that do not match known or theorized aspects about the data-generating process). Cross-validation (Arlot and Celisse, 2010) consists of splitting the data into partitions, and 'holding out' one partition in model fitting. Once the model is fit, the values of the explanatory variables from the observations in the test set are used compute fitted (predicted) values, and these fitted values are compared to the (known, true) value of the response. Done correctly, cross-validation simulates out-of-sample data, providing an estimate of how well the model generalizes. Prediction-only models are appropriate for detection tasks, although for interventions that change the system that produced the original data, knowledge of causality or at least of unbiased estimates become important.

Cross-validation is not perfect. First, if test data is re-used to test multiple models, it can create distributions over models (Dwork et al., 2015). Second, while the tendency of dependencies between observations to produce biased estimates and too-small standard errors are irrelevant since predictive models do not compute standard errors nor do they seek unbiased estimates, dependencies matter in another way. Dependencies between training and test sets (including network dependencies) may effectively share information across the training/test split, which will inflate predictive performance (Hammerla and Plötz, 2015; Bergmeir and Benítez, 2012; Chen and Lei, 2018; Dabbs and Junker, 2016). Thus, dependencies must be considered when partitioning data for cross-validation.

My interest is ultimately in explanatory models, not in finding predictive features. Feature selection can be used as an exploratory tool for substantive application (Lin, 2015; Dhar, 2013), since variables that are highly correlated with a response are more likely to have some substantive relationship, albeit with caution and many caveats (Yang and Yang, 2016; Mullainathan and Spiess, 2017): for example, a 'true' variable being selected out in favor of a correlated variable of which it is the cause. For feature selection methods with statistical justification, the justification is usually around predictive performance rather than which variables are selected. For example, given a data generating process, a feature selection method used on one data set observed from that process will come up with one set of variables, whereas the same method applied to a second data set observed from the same process might come up with a completely different set of variables, but the predictive performance of the two sets is, on average, the same (Mullainathan and Spiess, 2017). Nevertheless, in the presence of a huge number of possible ways to summarize sensor data, the exploratory process of feature extraction and selection is a more principled approach than simple aggregation.

3.4 Conclusion

Having endeavored to clarify what I see are confusions around the use of sensor data, and having identified a set of meaningful theoretical questions, appropriate constructs, and valid modeling approaches, I will proceed to carry out a sensor study in Chapter (5).

Part II

Responses

Chapter 4

Social media for applied research and interventions: Hookah sentiments on Twitter[1]

Summary. Instead of using Twitter for public health monitoring, we should be using it for public health campaigns. This uses Twitter the way it is meant to be used; a platform for engaging with others who are on Twitter, rather than as a tool for taking measurements of characteristics of larger populations. In cases where audiences receptive to certain messages are on Twitter, it becomes more appropriate to engage there than through print or broadcast media. Taking up the case of sentiments around hookah smoking on Twitter, I demonstrate a rigorous approach to model construction and evaluation, respecting both the purposes of the model and validation that simulates use cases.

4.1 Public health and social media platforms

Recently, there has been an explosion of interest in using digital trace data, and specifically Twitter, for public health research (Paul and Dredze, 2011; Sadilek, Kautz, and Silenzio, 2012; Sadilek, Kautz, DiPrete, et al., 2016). However, in light specifically of the failure of Google Flu Trends (Lazer, Kennedy, et al., 2014) and of the demographic biases I highlight in Chapter 1, relying on Twitter for public health information may provide a very biased, and not reliable or generalizable, view of public health trends. Furthermore, norms like commiseration or exaggeration present among various sub-communities might make data difficult to take at face value.

However, do these limitations mean that public health researchers eschew Twitter? I suggest a different approach: in the past, public health messaging and outreach has taken the form of public service announcements, aired during commercial times on broadcast television. This, rather than epidemiological research,

[1] This is a version of work done with Kar-Hai Chu, Jason Colditz, Tabitha Yates, and Brian Primack. The primary authors of the original work are Kar-Hai Chu and Jason Colditz. I have redescribed the results, and adapting the framing for this thesis but the main results, including the tables and figures, are due to Kar-Hai and Jason.

is the correct pre-Internet analog of Twitter. Just as public service announcements were broadcast on certain shows with the goal of reaching certain key target audiences rather than trying to be representative, Twitter can be used to reach those who use it. This approach uses Twitter for what it is built for, being a means of *engagement*, rather than trying to adapt Twitter to be a research or monitoring platform. Twitter is potentially superior to static broadcast and print public service announcements, as unlike those media, it can help researchers monitor ongoing trends, dynamically adjust, and interact around trends, adopting strategies of social media marketing and the style of engaging in many-to-many communications in place of previous top-down models (Nitins and Burgess, 2014). Social media-based interventions are in their infancy (Maher et al., 2014), but public health researchers are beginning to see online spaces as sites for possible intervention for behaviors like smoking (van den Heerik et al., 2017). Such uses avoid the severe (and demonstrated) threats to external validity posed by the non-representative and biased nature of Twitter data involved in using it as a tool for monitoring prevalences of diseases and behaviors.

4.2 Social media monitoring and interventions

The specific substantive problem addressed is that of waterpipe tobacco smoking (WTS), commonly known as hookah or shisha. As Rose et al. (2017) points out, menthol cigarettes are disproportionately used by demographics similar to those who use Twitter, and similarly, while cigarette smoking has been decreasing, WTS use has grown (CDC, 2016; Singh et al., 2016; Jamal et al., 2017) particularly among college students—a demographic that is also likely to use Twitter, making Twitter an appropriate platform for planning out public health messaging. Public health messaging through print and television has played a key role in raising awareness of health risks of cigarettes and decreasing consumption (Farrelly et al., 2002; Siegel and Biener, 2010), and it is appropriate for WTS because there is a lack of awareness to address: in WTS, water cools the tobacco combustion, which makes users think it is not harmful (Chan and Murin, 2011), but it has similar toxins as does cigarette smoke (Primack et al., 2016) and is associated with cancer, dependency, and other harms of cigarettes (Haddad et al., 2016; Kadhum et al., 2015; Sidani et al., 2015).

There have been several examples of looking at tobacco-related topics on Twitter. Rose et al. (2017) studied sentiment towards menthol cigarettes, Chu et al. (2015) looked at the diffusion of e-cigarette advertising, and most directly, Krauss et al. (2015) found content normalizing hookah use. Other work (Depue et al., 2015; Yoo et al., 2016) has found that social media exposure to tobacco products is associated with attitudes towards smoking and future smoking behavior. The unrepresentative nature of Twitter data means that we cannot generalize these findings (sentiment towards menthol cigarettes, diffusion of e-cigarette advertising, normalization of Twitter) to the public at large, but so long as Twitter remains a locus of inherent interest, characterizing what happens there is useful for taking action.

Health campaigns are already using social media platforms as media on which to engage the public, like smoking awareness campaigns (CDC, 2018; Chung, 2016) and the “ShishAware” campaign (Jawad et al., 2015) using Facebook, YouTube, and Twitter. However, campaigns that mimic the top-down one-to-many and day-to-day communication (i.e., not timely or about specific events) model of print and television are failing to take advantage of the nature and capacities of social media platforms, and specifically the possibility of issue-related communications (that are around a specific event) (Stieglitz and Krüger, 2014) and the general ‘talking back’ many-to-many communication style (Nitins and Burgess, 2014). On Twitter, for example brands that broadcast messages, rather than engaging with users, often find themselves caught unprepared for public relations disasters. Instead, public health can adopt ‘social marketing’ practices developed around media like Twitter (Thackeray et al., 2012).

In the space of WTS, this would mean not just doing a ‘broadcast’ model of putting out public health information, but of identifying specific Twitter users with less established opinions (Brennan et al., 2017). People who are already staunch hookah smokers are unlikely to be swayed by information, and people who are opposed to smoking do not require messaging.

But we do not necessarily know how people talk about WTS on Twitter in order to be able to identify users relevant to target, nor what sentiments they may have about WTS. So, in preparation for a future intervention, in this study I show how to lay the groundwork of a monitoring system for Twitter sentiments around WTS. I use a supervised learning approach, hand-tagging sentiments in a training set of tweets and using those to scale up inferences about sentiment to all tweets collected as relevant to WTS. Panger (2016) argues that this is an approach superior to using automatic (unsupervised) tagging systems; and indeed, the analysis shows that WTS has context-specific language use (e.g., the “dash” emoji, U+1F4A8, is used to represent blowing smoke) that general-purpose systems will not catch. Conversely, a supervised learning approach is far more practical than a standard social science approach of either being only able to deal with a sample, since we are dealing with a *detection* problem that seeks to identify all relevant instances from a population, rather than an estimation problem that seeks to characterize population-level parameters (for which sampling would be appropriate). And, human coding would not apply to labeling incoming tweets in realtime, which may be necessary for acting on the basis of ongoing WTS-related events.

Public health generally falls into social science traditions of a focus on explanatory modeling (Shmueli, 2010), for which black box prediction models are ill-suited. This is a case, however, where I argue that it is irrelevant how various black box models find correlations between tweet *n*-grams and human labels. We are not interested in causal processes or in estimating the magnitude of certain associations, but only in the ability of a model to reliably replicate human labels. Indeed, the causal processes that cause people with certain sentiments about WTS to use certain words a tweet are psychological and likely not measurable via those tweet words alone, meaning that we could not hope to achieve a correctly specified model with the variables we have available to us anyway.

Given, then, that the outcome of my model are the classifications and not estimated coefficients, the way in which I use the model is also nonstandard for public health. I look specifically at two use cases for what we would do with a trained and validated classifier: first, we can apply the classifier to all collected tweets and look at overall trends in positive and negative sentiment. While the constraints of the Search API again mean that such trends are not representative of Twitter and cannot be used to make solid scientific claims about relative prevalences on Twitter as a whole, for practical applications it would be good enough to go looking for potential causes of positive or negative spikes. For example, there might be a hookah-related event, such as when Prince Harry was photographed smoking hookah in 2014[2], around which tabloid coverage included WHO statistics about the harms of WTS (although possibly more to depict the act as salacious than out of

[2] “Party Prince Harry spotted smoking hookah pipe onboard yacht in Abu Dhabi”, 26 November 2014, *Daily Mail*, http://www.dailymail.co.uk/news/article-2851008/Partying-prince-Harry-spotted-smoking-shisha-pipe-onboard-yacht-Abu-Dhabi.html, accessed 7/2018.

interest in educating the public) and which an active public health campaign could have similarly injected into conversations in social media with appropriate hashtags and replies. Second, we are able to use the classifier to find users who express both positive and negative sentiments around WTS, as they may be good candidates for targeted messaging.

4.3 Data collection and analysis

Data collection was led by Jason B. Colditz, at the University of Pittsburgh School of Public Health. The six terms *hookah, hookahs, hooka, shisha, sheesha*, and *narghile* were used in the Twitter Search API (Gaffney and Puschmann, 2014) from which we collected approximately 560k posts between January 1, 2016 and June 30, 2016. While these six terms will not necessarily be exhaustive in collecting all WTS-related tweets, the limits of the Search API are such that we are are getting a convenience sample anyway of tweets that match a given term (Morstatter, Pfeffer, Liu, and Carley, 2013). For characterizing Twitter, this would not be effective, but it is valid for mimicking an application setting where we would only be able to react to the tweets that are available to us.

“Sentiment analysis” often refers to general-purpose dictionaries; but here, I follow Panger’s (2016) argument for the superiority of supervised learning for classifying sentiment. Contrary to what may be familiar in public health, when our only interest is in finding a model that fits well, there is no need to worry about problems of colinearity, variables not being causal, or even measurement error. As discussed elsewhere, this means we cannot interpret model coefficients for information about underlying relationships, as they will likely be heavily biased in addition to the overall model being misspecified. As is usual with predictive models, we are then able to use models beyond logistic regression or discriminant analysis, including models like Support Vector Machines that do not provide p-values or confidence intervals for doing inference. In particular, Random Forests have been shown to be an excellent general-purpose classifier (Caruana et al., 2008; Fernández-Delgado et al., 2014).

Machine learning, and the supervised learning setting, are becoming more common in public health and tobacco research. Several tobacco-related studies have used machine learning to analyze large datasets to classify topics (Cole-Lewis et al., 2015) and sentiment (Myslín et al., 2013), using the ability of machine learning to scale human labels to large datasets. There is the additional problem of class imbalance; Allem, Chu, et al. (2017) found that hookah-related content collected from social media tends to skew towards positive sentiments.

From the 560k collected tweets were sampled 5000 from the entire six-month period, which were labeled by two trained coders along three independent dimensions: positive or not positive, negative or not negative, commercial or not commercial. These categories emerged after some pre-testing with an earlier collected set of hookah-related tweets, where a large number of tweets were trivially positive because they were advertising hookah-related products or events, which we wanted to be able to separate out from organic discussions of WTS use. While positive and negative sentiment will usually be mutually exclusive, again after pre-testing found examples of tweets expressing ambiguity about WTS that could be coded as both, we chose to treat these as two independent dimensions. Furthermore, in the interest of our specific application, sentiment was defined as positive or negative *towards hookah*, rather than overall positive or negative (as off-the-shelf, dictionary-based or other unsupervised methods would do). Because of the class imbalance in negative vs. non-negative tweets, and our interest in potentially finding people expressing negative sentiments about Twitter to potentially target with reinforcing messages, we were willing to decrease sensitivity for increased specificity for detecting negative sentiment and consequently under-sampled negatively coded tweets in our training set.

In order to make emojis both interpretable and able to be read by text processing tools, they were replaced with appropriate equivalent in words, e.g. "[BLUSH]", although many emojis remained as unicode codes. After some exploration, we decided to code all URLs as the same 1-gram, rather than looking at the target of the URL, as there was too much variability. And, we expected the the presence of a URL to be predictive of commercial content. Commas and other punctuation were also coded as 1-grams.

I used the R packages `text2vec` (Selivanov and Wang, 2018) for extracting n-gram features from tweets, for eliminating sparse terms, and for TF-IDF. After exploring several boosting models, as well as ℓ_1 regularized logistic regression, I found a random forest gave the best performance, and which I focus on here. For random forests, I used the `ranger` package (Wright and Ziegler, 2017).

Language changes over time; on Twitter, the terms with which people talk about things can change in the course of hours. Thus, one drawback of training on tweets labeled in a certain time period is that the resulting classifier may not generalize to future linguistic content. Evaluating how my classifier's performance might drop over time is thus a crucial part of how we can judge its generalizability.

To do this, I temporally split the data in two, trained on the first half, then plotted the daily accuracy, precision, and recall over the second half of the data set. The way in which performance degrades in the test set approximates the quality of classification in tweets that come after the six months' worth of training tweets. For brevity, I focus on the hard task of identifying content with negative sentiments.

For evaluating the overall classifier performance, I held out a fifth of data in a temporal block.

4.4 Results

4.4.1 Inter-rater reliability

For human coders, inter-rater reliability (IRR), is 'substantial' or better. For positive/non-positive, Cohen's is $\kappa = 0.78$, for negative/non-negative it is $\kappa = 0.75$, and for commercial/non-commercial is $\kappa = 0.82$.

4.4.2 Performance over time

Figure (4.1) shows the performance of the classifier over time. Strictly speaking, temporal cross validation should include a 'buffer' block of data between training and test blocks in order to break temporal autocorrelation (Racine, 2000), but we are interested mainly in performance over a longer span of time. We can immediately see the drop from training performance to test performance; but beyond that, we see that both precision and recall overall seem to be decreasing over time. There is still a great deal of variance and so this is not a strong pattern, but it is reason to believe that at least for a length of time equal to the length of time covered by the training data, performance will remain good.

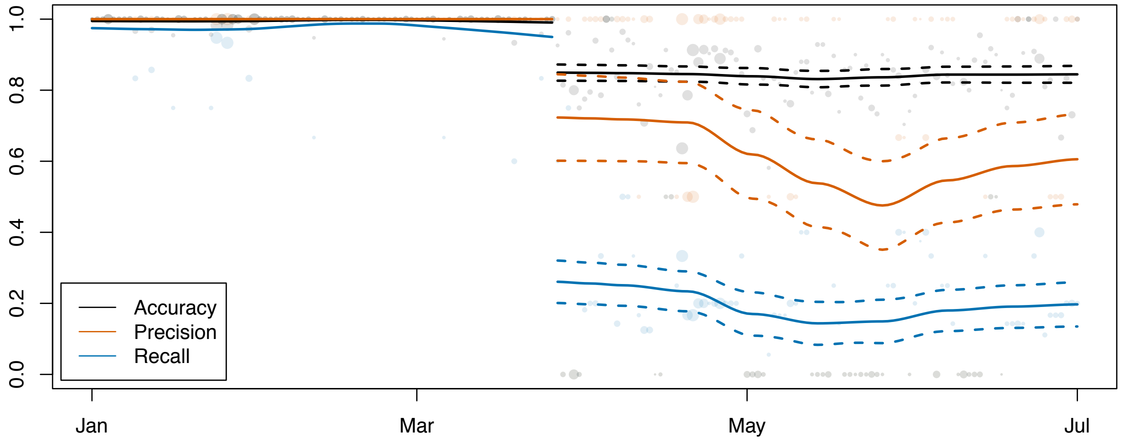


FIGURE 4.1: Performance drop over time in test set (right half). The circles are the daily observations, to give a sense of the overall variance; their size is proportional to the number of tweets observed on a given day, to identify outliers that are down-weighted and have little effect. Fit is a local constant, with 95% confidence intervals.

4.4.3 Classifier test performance

Performance is given in table (4.1). In Chapter (5), I cite arguments supporting the Matthews correlation coefficient as a good overall summary of classifier performance: while the sensitivity is low, in figure (4.2) we can see that if we sacrifice precision, we can increase recall, getting a recall of .8 for decreasing precision to about .4, which may be worthwhile depending on the application.

Accuracy	0.8577
Accuracy, 95% CI	(0.8343, 0.879)
(No Information Rate / Majority class)	*(0.8252)*
Binomial test, Accuracy vs. NIR, *p*-value	*p*=0.003425
Precision (Positive predictive value)	0.8810
Recall/Sensitivity (True positive rate)	0.2151
Specificity (True negative rate)	0.9938
F1 score	0.3458
AUC	0.8558
Matthews correlation coefficient	0.3926

TABLE 4.1: Performance of a classifier for negative sentiment, tested on a held-out temporal block of 1/5 of the data.

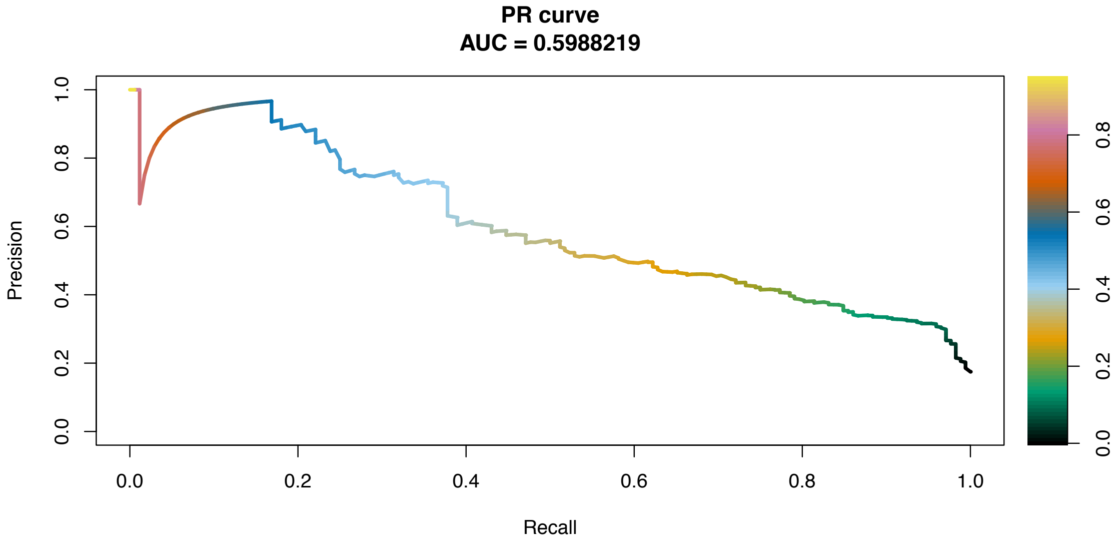


FIGURE 4.2: Precision-recall curve for detecting negative sentiment.

4.4.4 Use case 1: Time series of detected sentiments

The time series output of applying the application of the classifier to the collected tweets is shown in figure (4.3). Note that this is the result of a separate classifier, as well as a scheme that collapsed "commercial" tweets to positive.

4.4.5 Use case 2: Mixed-sentiment users

Using the classification, we are able to identify users with more than one sentiment expressed across multiple tweets. Examples of such tweets across users is shown in table (4.2). Some of them, like user 10, are having adverse immediate experiences with hookah that could make them receptive to messages about long-term harms.

4.5 Limitations

One assumption I am making is that injecting public health messaging into conversations on Twitter would have a positive effect, but it is entirely possible that users would resent this as a paternalistic intrusion and that it would have opposite effect. Indeed, around an e-cigarette counter-campaign, Allem, Escobedo, et al. (2017) found tweets objecting to government regulation, refuting ties between e-cigarette manufacturers and tobacco companies, and touting health benefits of e-cigarettes (although they also note that they did not specifically consider whether tweets were from bot, astroturf, or other engineered accounts, potentially from e-cigarette manufacturers). The next step would be experimenting with ways of making WTS-related

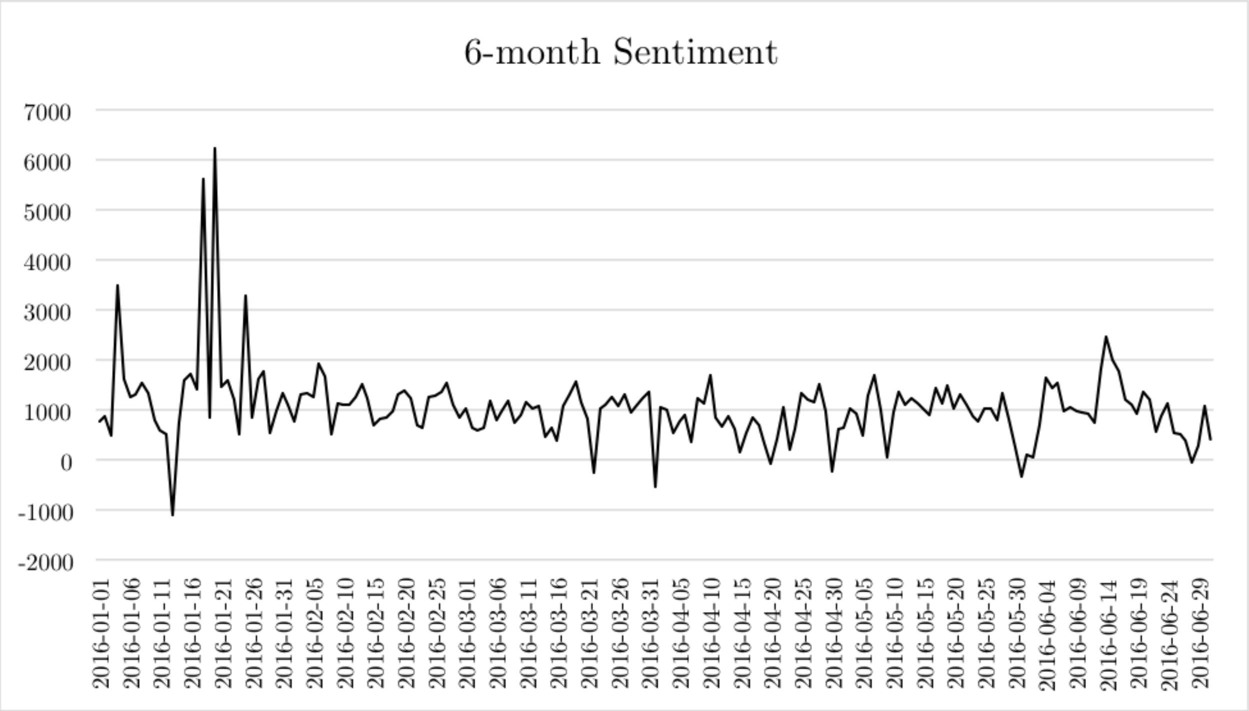


FIGURE 4.3: Time series of sentiments determined from application of the classifier over the 6-month period of collected tweets.

messaging visible but not obnoxious (perhaps by only using hashtags, and not mentions or replies), and experimenting with different rhetorical strategies (e.g., should information be delivered in formal language or humorously, e.g. using memes? Should it be delivered with text, or through infographics?), all of which would potentially have an enormous impact on the success of any eventual strategy.

Second, the stability of the classifier may decrease over time. We assumed a six-month period was a representative block, but this is not necessarily the case. Shifts in overall sentiment about hookah or shifts in the language around WTS would decrease the validity of the classifier without an analyst necessary knowing. For an active system, having occasional audits or coding updates would help maintain the integrity of classification over longer periods of time.

4.6 Conclusion

Public health has recognized that it needs to adapt to new media in communicating with the public (Harris et al., 2014; Grant, 2017). While Twitter is not appropriate for all type of public health issues, with WTS, the demographics that are likely to use, adopt, and have malleable opinions about WTS are the same demographics that are disproportionately represented on Twitter—and who are potentially discussing WTS on Twitter. Adopting techniques from social media marketing, public health officials should not continue top-down

User	Tweet
1	Wednesday about to be lit lmao I need a hookah man
1	I don't want hookah no more dawg lmao
2	Life feels so good when you are smoking hookah.. [BLUSH]
2	SO I TRIED VAPING TODAY [COMMA] ON 108Hz.. HAHAHA fucking hard but such thick clouds [COMMA] vaping is the best ! gotta quit shisha and start vaping now!
3	@[USERNAME] stop smokin hookah then
3	The hookah spot was rockin wit bitches feenin for cancer smh
4	I'm smoking hookah in front of my building right now [URL]
4	My goal is to not DJ any spots with Hookah this summer
5	Almost all my male friends love hookah smh
5	Trying to put plans together for Chandra's birthday and I have to make sure hookah is involved [WEARY]
6	Man y'all be paying 20 dollars at hookah spots to stare at each other [sob]
6	I only smoke shisha once in a while tbh lmao and wth we got jobs and school [URL]
7	She was sent from the heavens... She don't smoke hookah or know about lemonade. #Skinny
7	I gotta find a way to make crab flavored hookah tobacco. #Skinny
8	FAM be proud of me I havent smoked hookah ALL year -@[USERNAME]
8	My ramadan nights bouta consist of me sitting on the porch till 5am skyping and smoking hookah.
9	I wish hookah never existed [URL]
9	There's no hookah so why go [URL]
10	I've done hookah less than 5 times
10	whenever I smoke hookah I wanna throw up

Table 4.2: Sample of users with different-sentiment tweets

communications, but recognize the possibilities and indeed necessity of engaging in targeted messaging that actively engages with the audience and that picks up on timely themes. Doing so requires some guidance, as modeling techniques involved in social media marketing are distinctly different from the modeling approaches familiar to public health (Saeb et al., 2016). Here, using predictive modeling for engagement, with 'black box' models that are misspecified and biased in their individual terms but that overall form effective proxies for human labels, is the appropriate way to use available data. I provide an illustration of what this approach would look like, and some of the uses cases that such modeling enables.

Chapter 5

Sensors and social network data: Detecting friendship with smartphone co-location data[1]

Summary: In order to act on the theoretical concerns and possibilities I raised in Chapter 3, in collaboration with Afsaneh Doryab, Michael Merrill, and Anind Dey, I carried out a three-month study to look at the relationship between friendship, measured by sociometric choice, and proximity, gathered from smartphone data. Towards the goal of making sensor data commensurate with survey data for putting in statistical models, I pursue a feature engineering and selection task, where I try to find summaries of sensor data that can be most meaningfully compared with self-reported friendships. Part of this is carefully employing different cross-validation schema to get more realistic appraisals of out-of-sample performance. In my final results, I find a subset of 19 features (out of about 2,000) that achieve good performance, that can be the starting point for multivariate models.

5.1 Introduction

In Chapter (3), I made the following arguments for theoretical motivation and modeling approaches:

- Mobile phone sensor data can capture the construct of *proximity*.
- The framework of Stochastic Actor-Oriented Models provides a way of modeling complex dependencies in network processes, and is appropriate for use with sensor data collected over time. Thus, study design for sensor studies should be set up like studies using SAOMs, whether at longer-term scales (SIENA) or for modeling event streams (REM, DyNAM).

[1] This is work done in collaboration with Afsaneh Doryab, Michael Merrill, and Anind Dey. A version of this was submitted as the Data Analysis Project for my Secondary Master's in Machine Learning, Machine Learning Department, School of Computer Science, Carnegie Mellon University, on May 15, 2018, with committee members Anind Dey, Afsaneh Doryab, and Nynke M. D. Niezink.

- Proximity data is most interesting for comparing with other forms of data (i.e., rather than characterizing univariate distributions or modeling univariate time series), for example various forms of survey data. However, sensor data is far more dense than survey data, and needs to be reduced to a commensurate scale to fit in SIENA.
- Simple aggregations are not necessarily the best way to extract the relevant information captured in sensor data. However, we do not *a priori* know what the most meaningful extractions will be.
- The machine learning approach of feature extraction can help create potentially relevant dimensions for modeling. Extracting a large number of features and using feature selection, with selection validated by use of data-splitting and cross-validation, can be used as an exploratory tool for making sense of high-dimensional data.
- Dependencies in data complicate the process of cross-validation, and require careful handling.

In order to demonstrate a valid use of sensor data, making use of both theoretical possibilities that have been under-explored, and of connecting modeling approaches that have so far been separate (the SAOM framework with machine learning approaches to data reduction), here I describe an original study carried out with the use of sensor data. In order to avoid data re-use, I focus my analysis on the step of extracting ways of characterizing sensor data, rather than on using a SAOM. However, my setup, from my study design and survey instruments through to my goal of using sensor data to get a single covariate for characterizing proximity, is to determine how to effectively use sensor data in a SAOM.

The output of my study is a classifier that detects friendships from smartphone location data. My work is the first effort to do such detection; previous works either did not validate predictions of friendship from location data (Eagle, Pentland, and Lazer, 2009), looked at ties on location-based online social network services (Cranshaw et al., 2010), or used mobile phone call and SMS logs (Wiese et al., 2014; Wiese et al., 2015). I believe that detecting friendship from mobile phone co-location data is a realistic approach for future mobile applications and interventions that seek to leverage friendship for other tasks.

This paper presents the results of a 3-month study of a cohort of 53 participants, with final analysis performed on 9 weeks of data from 48 participants. I combine mobile phone sensor data collection with established social network survey instruments, and use rich feature extraction from co-location data to see how well such data can be used to detect friendships, *close* friendships, and *changes* in friendship.

My contributions are as follows:

- I present, to my knowledge, the first *pairwise* feature extraction from smartphone location data, and show that a classifier built with the extracted features **performs 30% above random** (Matthews correlation coefficient). This can serve as a baseline for all future work.
- I design a novel evaluation method (using temporal block assignment cross-validation and what I call dyadic assignment cross-validation) to mimic different realistic application settings in order to more rigorously test my classifier's generalizability to these settings, and use it to show that my approach is robust to seeing new pairings of individuals, and to variability in co-location patterns over time.

5.2 Method

5.2.1 Participants and recruitment

We recruited members of an undergraduate fraternity in a research university in the northeastern United States. The fraternity had 60 members at the start of the study, with an additional 21 prospective members going through the 'pledging' process during the study duration, of which 19 completed the process. Of this cohort of 79 men, we recruited 66 participants, of which 53 ultimately participated in sensor data collection, and of which 48 responded to at least one survey wave. Having this sort of well-defined *boundary specification* (Laumann et al., 1983) lets me ask each study participant about their friendships with each member of the fraternity, giving negative examples that are explicit, unlike open-ended solicitation for friendships (such as from 'name generator' instruments) in which individuals are only implicitly not friends by not being mentioned.

The fraternity was relatively loose-knit; about 20 fraternity members live in a fraternity house, with the rest living elsewhere and required to be in the fraternity only one day a week (for a fraternity chapter-wide meeting). Participants were compensated $20 a week for having the passive and automated sensor data collection software, AWARE (Ferreira et al., 2015), installed on their smartphones, with additional $5 incentives for each survey wave they completed.

5.2.2 Study setting

Using SAOMs as multinomial choice models assumes that all possible choices (i.e., for sending ties) are present in data. The best way to do this is to have a well-specified boundary (Laumann et al., 1983). For sociometric choice as the modeling target, having a boundary allows asking participants exhaustively about ties with others within the boundary.

My population is not as clean as the carefully controlled population selected for the Newcomb-Nordlie fraternity study, nor does it have the isolation or feature of having people who were previously strangers as in the Westgate study or the Groningen sociology freshmen study. But this setting did have the advantage that the fraternity listed its members publicly, so I could ask those who participated in the study about their ties with all other fraternity members, not just those members who agreed to participate in the study. I did not have sensor data from non-participants to actually use in the study and so could not use reported ties to these non-participants, but I was able to look at the indegrees of non-participants to get a better sense of the possible consequences of their non-participation, which is better than testing if they are systematically different in their demographics. I further describe the survey instrument below.

5.2.3 Mobile platform

I used the AWARE framework (Ferreira et al., 2015),[2] software developed for both Android and iOS (iPhone) devices to record sensor data. Among my population, ≈90% of participants had iPhones and the remaining ≈10% had Android phones; there were no users of Windows or other mobile operating systems.

Mobile 'development', which is the industry devoted to writing applications for mobile devices (the kind of applications that may be downloaded from the Google Play Stores or Apple App Store, and run both on smartphones and some tablets), takes place within frameworks developed respectively by Google and Apple where developers can use specific snippets of code to access various functions of phones. However, such access is not unlimited; I take as an example detection of Bluetooth devices.

I discovered as part of my study that, unlike previously, both Apple nor Google no longer make available to app developers the 16 hex digit identifier of Bluetooth devices (the Bluetooth MAC addresses) of detected Bluetooth devices. Instead, detected devices are recorded in terms of a 32 hex digit universally unique identifier (UUID), which are assigned by the detecting device uniquely to each detected device and used to recognize those detected devices in the future. This preserves privacy of detected Bluetooth devices by making it impossible for two different devices to know that they have detected a third Bluetooth device in common, but also makes it impossible to determine from the data alone which device was detected.

Early testing showed Bluetooth MAC addresses being recorded by the app, but these turned out to be other devices like Fitbits (detections of which are still recorded in terms of their Bluetooth MAC address). I recorded the MAC addresses of the individual phones in the study, expecting to be able to create Bluetooth-based networks, only to find that only two pairs of phones in the study were able to detect one another's Bluetooth MAC addresses (both of which were Android devices detecting iOS devices). In all other cases, the detection was a UUID that I could not match with mobile devices in the study.

Since a UUID is unique for each device that detects a given device, I was also unable to use mutual detection of a third Bluetooth device as a way to measure proximity. There were cases of devices recorded in terms of MAC address by multiple phones that I tried to use for common detection, but there were only about a half dozen or so such common detections over the entire study period, not enough to meaningfully use.

I was fortunate to have the support of mobile developers (i.e. programmers) who could assist in the maintenance of the software, which proved vital; unfortunately, it may be some time before there is surveillance research software that can be used without dedicated development labor. Not only are there inevitably bugs, but also based on how mobile operating systems (and mobile devices) and the various access permissions and features are almost constantly being upgrading by Apple and Google, sensing platforms must keep apace to retain all functionality.

While I was fortunate to have a cross-operating system platform, such that I did not have to limit study eligibility to either only Android users or only iPhone users, the two development environments have differences that impact the comparability of collected sensors. The AWARE framework needed to be developed from the ground up independently for iOS and for Android, as the way in which equivalent sensor signals are

[2] https://www.awareframework.com

accessed from the phone within both operating systems is different, requiring completely different input and code.

One solution, taken by the Reality Mining and Social fMRI studies is to provide the same device for all participants: given that high-quality mobile devices are expensive, this also serves as an incentive for participation. However, letting participants use their own devices is preferably because it is more lightweight for both researcher and for participants, who do not have to have any interruption in their normal phone usage.

There were also difficulties in writing data from two different versions of AWARE into the same database (and combining those data later on to have a consistent analysis across device types). But overall, while maintaining two separate AWARE implementations requires a large amount of coordination (e.g., adding features to one operating system's version need to be replicated, potentially from scratch, in the other), AWARE has managed to do this fairly effectively, including working out (with some required manual monitoring or tweaking) how to use the same database for the two AWARE versions, and having scripts that resolve differences between iOS data and Android data in preparing the data for analysis.

While Bluetooth was ultimately not usable, it was one of the collected sensors; both Bluetooth and scans of nearby WiFi hotspots were done at 10 minute sampling intervals. I also had continuous monitoring of battery (percentage, and whether the battery was charging or not) and screen status (on/off), barometer (which can be used to calculate height above sea level, for examine to determine whether two individuals are on the same floor of a building—although the barometer measurements were not sufficient for doing this), and complete records of call and message metadata (with hashed values for phone numbers), although none of these are used in the present work.

Location

Location was a key quantity of interest, but GPS is one of the most battery-draining applications; instead, I used functionality built into AWARE.

The Android AWARE client uses the Google Fused Location plugin to collect location data. The PRIORITY_LOW_POWER option prioritizes low power usage, as previous testing with AWARE had showed battery drain was a major cause of participant dropout. This low power option does not actively use GPS, instead using a combination of cell phone towers and detected Wifi hotspot with known geolocations, and is advertised as being accurate to within about 10km.[3] In the iOS client, the accuracy setting corresponding to low power use was to set `desiredAccuracy` option to 1km, with a threshhold for recording new movements of 1000m.[4] In practice, the reported accuracy was usually much better, with a significant portion of readings reporting an accuracy of within 10m.

[3] "LocationRequest", 12 April 2018, *Google APIs for Android*, https://developers.google.com/android/reference/com/google/android/gms/location/LocationRequest, accessed 7/2018, and "AWARE: Google Fused Location", 2018, *AWARE: Open-source context instrumentation framework for everyone*, http://www.awareframework.com/plugin/?package=com.aware.plugin.google.fused_location, accessed 7/2018.

[4] "Location and Maps Programming Guide", 21 March 2016, Apple Developer Documentation Archive, https://developer.apple.com/library/content/documentation/UserExperience/Conceptual/LocationAwarenessPG/CoreLocation/CoreLocation.html, accessed 7/2018.

Importantly, location only updates when a device has moved a ‘significant’ distance away (this threshold is also configurable); this preserves battery, but generates irregular time series. Furthermore, long periods without data could either be because a device has not moved, or because data was not successfully collected, requiring the use of other sensors to determine when data is missing.

Wifi

One candidate for characterizing proximity is when two devices detect or connect to the same Wifi device. Here, Wifi hotspot MAC addresses are unique (unlike the hotspot name/label, which for example with ‘eduroam’ is shared not only across multiple hotspots in the same university, but across multiple cities across the world!), and mutual detection of this picks up when two devices are proximate.

I also conduced *Wifi fingerprinting* in the fraternity house. This involved walking around the fraternity house with a smartphone and collecting all WiFi hotspots detected in each room, along with the received signal strength indication (RSSI) of the respective signals. WiFi fingerprinting can, in theory, be used for indoor localization; however, I found that only about 6.7% of scans for WiFi hotspots throughout the study recorded more than one detected hotspot, and similarly in the frat house, I could tell when a device was connected to one of the frat house’s Wifi hotspots but not which other hotspots were detected in order to determine a specific room.

Thus, I only use as the basis for features whether at least one Wifi hotspot was detected in common at the scan of a specific 10 minute interval from two devices, ignoring the tiny fraction of detections that include multiple devices, and also ignoring RSSI. The frat house has 5 main Wifi devices for about 30 rooms over 3 floors. Based on the size of rooms in the fraternity house and the relative coverage of its Wifi devices, I estimate that at least within the fraternity house, my Wifi localization approach is accurate to within a bit of a smaller radius than its general 32m accuracy, perhaps 20m or so; however, I do not have similar measurements for the rest of campus.

5.2.4 Compliance and retention

As discussed above, devices other than mobile phones have particular issues with compliance; but as I argue that mobile phones have emerged as the *de facto* standard and will be moving forward as well, I focus here on compliance with mobile-phone based sensing studies. Here, having a sensing platform is necessary but not sufficient for running a study. Even if the respondent burden is far lower than in studies that require participants to carry and use additional materials like a journal, the traditional difficulty of ensuring participant retention still applies. In the case of AWARE, people may turn off their phones, go into airplane mode, deactivate Bluetooth, GPS, or Wifi, or not carry their phone with then, all of which will create periods of missing data. Especially in the case of Android phones with their multiple manufacturers, there are problems or variations with individual companies and models (for example, Sony phones have “stamina mode” which interferes with data collection). In the cause of audio, Bluetooth listening/recording devices (e.g., headphones) or other audio recording applications and AWARE’s audio sensing may interfere with one another, causing one or the other to not work.

In some cases (for example, going into airplane mode being an indication of travel via airplane), such gaps may be informative, but in other cases they will cause data to be missing not at random. One advantage to sensing platforms, however, is that collected data is constantly uploaded to a database on a server, where the data stream can be used to monitor compliance. So, for example, if there is one participant without any recent uploads, it might be an indication that they have deactivated the AWARE app and left the study, which can prompt a follow-up to either confirm, or troubleshoot. If one particular sensor is not uploading data (e.g., Bluetooth), the participant may have turned it off on her/his phone, and again depending on the research questions and the importance of complete data, this may be followed up manually or automatically by a reminder to reactivate the sensor.

Note that the default setting for uploading sensing data stored on the phone to a server is to only do so when the phone is charging and has Wifi (to not drain the batteries of participant's devices, and to not risk using up limited data allowances); this may cause large gaps to appear during monitoring , but in principle all the data is retained on the phone until the upload can happen, so no data is lost.

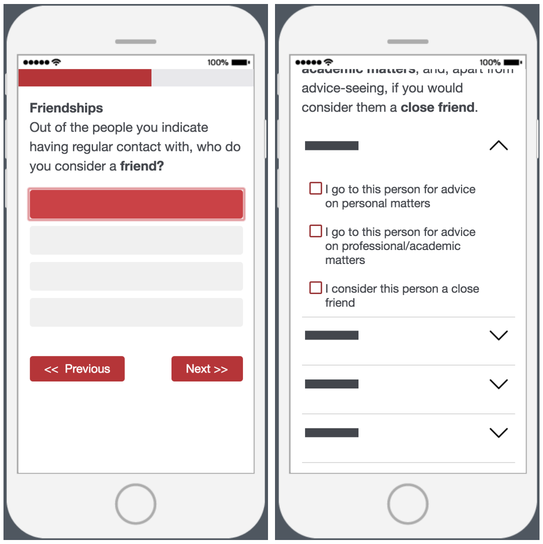
The main threats to compliance and retention were when AWARE interfered with the operation of other applications, when it caused the phone to run more slowly, or if it drains battery (the latter two being problems particularly for older phone models). In addition, if there are long gaps between uploads to the server, if the sampling rate for certain applications is exceptionally high, and if the phone has limited storage space, AWARE may cause the phone to run out of storage; in this case, data may be lost. If this is a concern, the sampling rate of high-volume sensors (e.g., audio) can be reduced, but the study should also only be run with populations that have regular access to Wifi (or that have lots of data to spare). Working on optimal data compression are an important part of development, as better compression means less space is taken up on the phone (during a previous AWARE study, an update improving the data transmission helped considerably reduce the data retention burden on participants' phones). In a worst-case scenario, a mobile device may be collecting more data than it can upload, e.g. if in 1 hour it collects data that takes more than an hour to upload, which is a definite possibility for many types of data and/or with high sampling rates.

A certain amount of increased battery usage is inevitable, but an enormous amount of development effort has gone into minimizing this drain as well as to eliminating or minimizing interference with other applications or the normal functioning of the phone. However, with older phones in particular, there may be no way to limit battery usage or interference with normal functionality (or, the development effort required to do so is prohibitive). In practice, I did not find the issue of *surveillance*, i.e. concerns around privacy, to be a major consideration in participant retention; I discuss this further in the 5.7.

5.2.5 Survey instrument

I based survey questions on the instruments used in SAOM studies, and particularly that of Duijn et al. (2003). While at first I explored writing a custom plug-in for AWARE that would give popup social network survey questions, this proved unwieldy, and I ultimately delivered the questionnaire outside of AWARE. Specifically, I used Qualtrics, and using the "carry forward choices" functionality, were able to ask about the population of 79 in a lower-burden way: I first asked participants to nominate people "you have had regular interactions with." For only the subset of alters thus nominated, I asked if they go to that person for advice on personal matters, if they go to that person for advice on professional/academic matters, and if they consider that person a close friend. This structure assumes that only those with whom participants frequently interact can be considered friends, and only those who they consider friends can be sources of personal or professional/academic advice, neither of which are necessarily the case (I do believe it a reasonable assumption that only friends can be considered close friends); but I found this to be a low-burden way of collecting multiple types of network ties within the available survey software functionality.

While the survey was not integrated into the mobile software, Qualtrics does provide a mobile-friendly interface for answering questions. Screenshots of how my survey appeared on phones are given in figure (5.1).



Friendships
Out of the people you indicate having regular contact with, who do you consider a **friend?**

<< Previous Next >>

academic matters, and, apart from advice-seeing, if you would consider them a **close friend.**

[redacted]

□ I go to this person for advice on personal matters

□ I go to this person for advice on professional/academic matters

□ I consider this person a close friend

[redacted]

[redacted]

[redacted]

FIGURE 5.1: Two views of how the survey appeared on mobile devices.

The extent to which such self-report accurately captures the underlying psychological construct of friendship is a much larger topic; here, I defer to previous work in social science discussing the validity of such measures (Fehr, 1996; Krackhardt, 1987; Duijn et al., 2003). Suffice to say, self-report is not ideal even if our goal is a construct that is a psychological, not physical, entity; but there have been both theoretical arguments and demonstrations of criterion-related validity, i.e., of self-report predicting outcomes that we should expect the construct of friendship to predict (Freeman et al., 1987), that establish the overall validity of self-report.

5.2.6 Research goals

The main goal of my research was to understand the relationship between proximity and friendship. I operationalize this as asking, what characterizations of proximity are most related to friendship? More specifically,

1. How well can we detect friendships from co-location? In other words, if all we know about two people is their location patterns, how accurately can we say if they are friends?
2. Do the detected friendships capture a greater proportion of self-reported *close* friendships?
3. How accurately can we detect whether a friendship is likely to change? Will co-location patterns over time provide information about the creation or dissolution of friendships?

In preparation for a SAOM, which models the co-evolution of behavior (in this case, proximity behavior) and sociometric choice, I am interested in how well we do if we treat changes in friendship purely as a detection problem as a baseline.

5.3 Data collection

The surveys were collected three times over 9 weeks: shortly after the beginning of the study, then four weeks after, and lastly at the end of the study five weeks later (I made the second period longer, as one of these five weeks was spring break, when many study participants were away from campus). The completeness of the survey data is shown in fig. (5.3a). The response rate dropped in each survey round; compared to survey 1, survey 2 had a response rate of 59%, and survey 3 had a response rate of 51%. In total, there were 48 participants providing network data, 34 of which responded to 2 surveys giving me longitudinal network data (the minimum requirement for looking at evolution or co-evolution), including 20 participants that responded to all 3 surveys. In total, out of $\binom{48}{2} = 1128$ potential pairs, I was able to train and/or test on 830 pairs.

Apart for non-response, there was also non-participation: those members of the fraternity who did not participate in the study. As mentioned above, the public listing of fraternity members allowed me to ask participants about friendship ties with non-participants as well; these responses were not usable for modeling using sensor data, since I did not have mobile phone sensor data from non-participants, but we can get a sense in *network* terms of the relative importance of missing individuals. It would be ideal if non-participation were correlated with being peripheral to the social system. Unfortunately, this is not the case (fig. 5.2a). One network, the interaction network from survey wave 1 (breakdown by individual survey waves not shown in figure) was particularly troublesome: in it, one non-participant received 18 nominations (i.e., 18 study respondents reported frequently interacting with them), another study non-participant received 20 nominations, and one survey non-respondent received 17 nominations. In subsequent survey rounds of the interaction networks, these participants received respectively 15 and 13 nominations, 6 and 10 nominations, and 9 and 11 nominations, compared to, for survey respondents, mean interaction network nominations of 8.46 in survey 2 and 7.46 in survey 3. Looking only at the friendship network, the focus of my modeling, does not improve things much; again, there are study non-participants and survey non-respondents with high degrees in various waves of the friendship questions. This remains a limitation in my work; future data collection can only try to achieve more complete data.

As another issue, *spring break* may be extremely informative, for example if two people are proximate to each other but far from everybody else it may be that they are more likely to be friends. However, spring break is systematically different from every other week, such that if we train on spring break, we have no

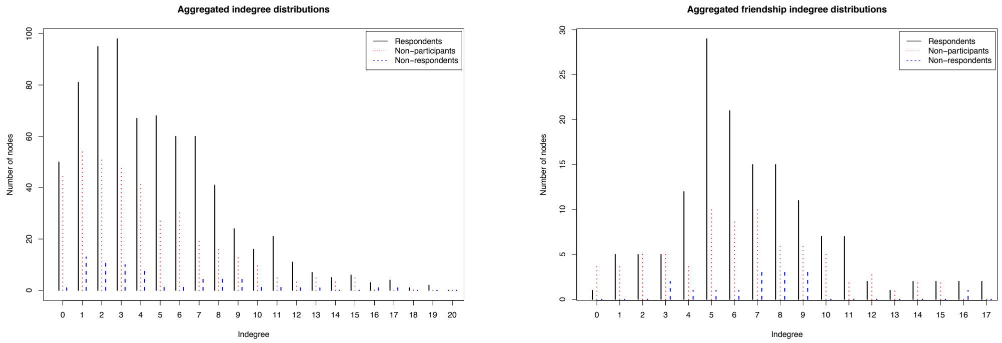


FIGURE 5.2: Indegree distributions (non-normalized), aggregated across all three survey waves and all network question types (left), and across the three survey waves for friendship questions only (right). If study non-participants and survey non-respondents were peripheral to the social system, they would have far greater mass at low indegrees. Instead, they have similar indegrees to those of those study participants who responded to at least one survey.

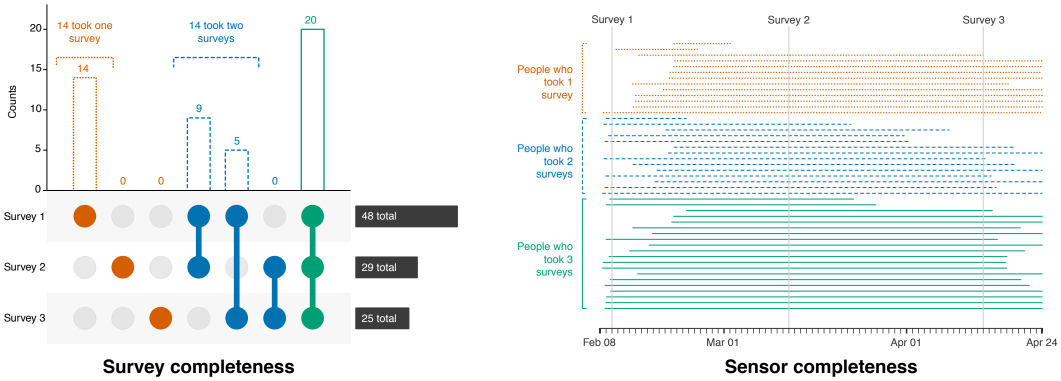
meaningful test set. Thus for the detection task, I removed spring break from the data set, which turns the unequal number of weeks between waves (4 weeks between survey waves 1 and 2 versus 5 weeks between survey waves 2 and 3) into an equal number of 4 weeks each, as spring break fell between survey waves 2 and 3, such that removing it leaves 4 weeks in each period.

Sensor data were collected throughout, and their completeness is shown in fig. (5.3b). Some logistical problems prevented all participants from starting smartphone data collection on the first day, and some participants discontinued the use of the app because of technical issues (battery life, sporadic interference with certain external Bluetooth devices, etc.).

5.3.1 Missing data handling

The main sensor data I used is location. Bluetooth, Wifi, and barometer were all set to sample at 10 minute intervals. Both Bluetooth and Wifi can be turned off, and several devices did not have a barometer; thus, I aggregated these and all available sensors for determining when AWARE was not active (again, could happen if the device were off, if AWARE were manually stopped, or if there was an error in data recording or transmission).

While figure (5.3b) shows completeness of sensor data collection over time periods, it doesn't show when periods of data might have been missing between when participants joined and when the study ended (or when they dropped out). I find that only small *number of readings* have gaps greater than the sampling rate of 10 minutes. However, when I account for the *proportion of time* accounted for by gaps between readings, we find that only about 60% of the total participant-hours are present in the data (fig. 5.4). We can see that interpolation (specifically, last-observation-carried-forward interpolation) would do little good; interpolating one 10-minute-period only gives about a half a percentage more coverage of the total time; interpolating up to an hour only gives an additional 1 percentage point; and so on. Only interpolating periods up to 8



Survey completeness **Sensor completeness**

FIGURE 5.3: (Left) Looking at longitudinal completeness, 14 people completed survey 1 only, and none completed surveys 2 or 3 only. 9 people completed surveys 1 and 2, 5 people completed surveys 1 and 3, none completed 2 and 3 only, and 20 people completed all three waves. This is shown in the vertical bars at the top. This comes out to 48 respondents for survey 1, 29 respondents for survey 2, and 25 respondents for survey 3, shown in the solid horizontal bars on the right side. (Right) I show the time periods in which sensor data was collected for people who answered one survey (dotted lines), two surveys (dashed lines), or all three surveys (solid lines). The times of the three surveys are marked with vertical lines.

hours (480 minutes) gives a meaningful additional 20 percentage points of coverage. Given the inevitable assumptions involved in interpolation, and how it anything short of extensive interpolation would have a negligible impact on overall coverage, I chose to not use any interpolation, and retained missing values in my data when training and testing.

5.4 Method

As is common in supervised learning settings, the ‘ground truth’ data, the survey responses, are far more sparse than the sensor data. I have one survey measurement every 4 weeks, but with sensor sampling once every 10 minutes (i.e., six times an hour), we potentially have $6 \times 24 \times 7 \times 4 \approx 4000$ data points per survey response. We need to summarize these data points in a way that extracts some essential information in order to compare with the response variable of self-reported friendships. I explain the details of my feature extraction, and the data processing required before extracting features.

5.4.1 Data processing

I consider three sensors for giving proximity or a measure of geographic similarity: location from GPS, for which I relied on the ‘Google fused location’ plug-in for a processed signal, to detect geographic similarity; Bluetooth, which unfortunately turned out to be unusable, to detect proximity through mutual detection; and Wifi, for two devices mutually detecting (or connecting to) the same Wifi hotspot.

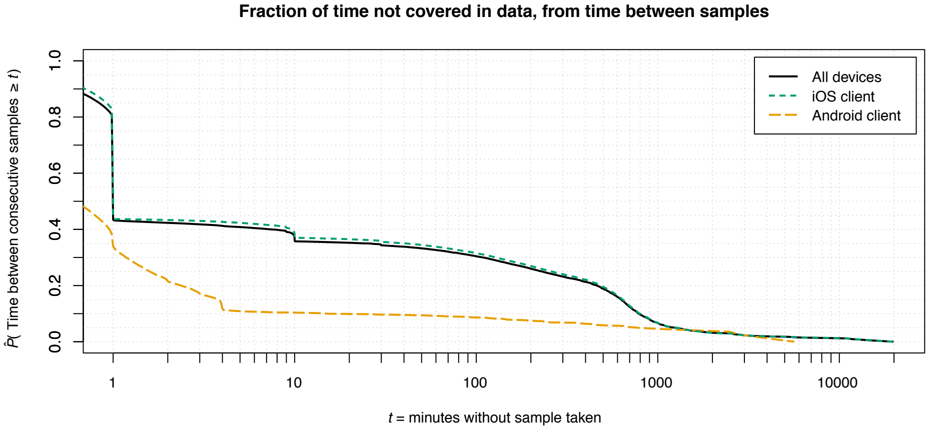


FIGURE 5.4: A plot of the survival function (the empirical complementary cumulative distribution) for time not covered by sensor data. 40% of the study time across all participants is covered by sensor readings less frequent than 10 minutes, meaning that 60% of the data has no gaps in coverage.

The location signals, as given, are irregularly spaced point observations of individual device locations. I treat these points as representing intervals; that is, for record $(i, t_1, longitude_{t_1}, latitude_{t_1})$, I assume that device i stayed at the point $(longitude_{t_1}, latitude_{t_1})$ until t_2. As noted above, a lack of a record between two times can either be because the device did not move, or because there was a failure in data collection for some reason. The first step I take, then, is in manually insert missing data values into these irregular time series; for the regularly sampled sensors (Bluetooth, WiFi, and barometer), if there were no readings for over 10 minutes, then I took that time t'_1 and inserted a record $(i, t'_1 + 10, \text{NA}, \text{NA})$ into the location data, and treated as missing data the interval from that time until the next location observation.

Next, for every pair of participants, I merged their irregular time series, looking like what is shown in table (5.1). This representation is simplified; from the raw data, I also carried forward provided accuracy measurements (given in meters, such that smaller numbers were better). After this, I used last-observation-carried-forward (including whether that last observation was missing or not) to fill in the empty cells for each time series, to make them commensurate.

Next, I used the measurements of longitude and latitude[5] to calculate the *haversine distance* between them, a measure of distance for two points on the surface of a sphere (this may not have been strictly necessary as, on a sphere as large as the earth, the distance between very close points can be approximated by Euclidean distances), producing a irregular time series of pairwise distances. I carried forward missing values into this time series. These time series of pairwise distances is what I ultimate performed feature extraction on, a process conceptually visualized in figure (5.5).

While I explored retaining the time series of pairwise distances as irregular for performing feature extraction, this gave little benefit; I regularized the time series, aggregating into one-minute intervals, because at one

[5] We usually refer to geolocation as ‘latitude and longitude’, or ‘lat-lon[g]’, but confusingly, many mathematical operations assume that longitude is given first.

$time_i$	$time_j$	$longitude_i$	$latitude_j$	$longitude_i$	$latitude_j$
17		0.487752478318	0.440915407684		
	23			0.476752391604	0.455347458639
	25			0.476277618459	0.455413653127
	29			NA	NA
33		0.488107678368	0.440784360037		
	49			0.476758742665	0.455328492764
59		0.487849721147	0.440819631710		

Table 5.1: This shows an intermediary step of merging two irregular time series towards getting measurements of pairwise distances.

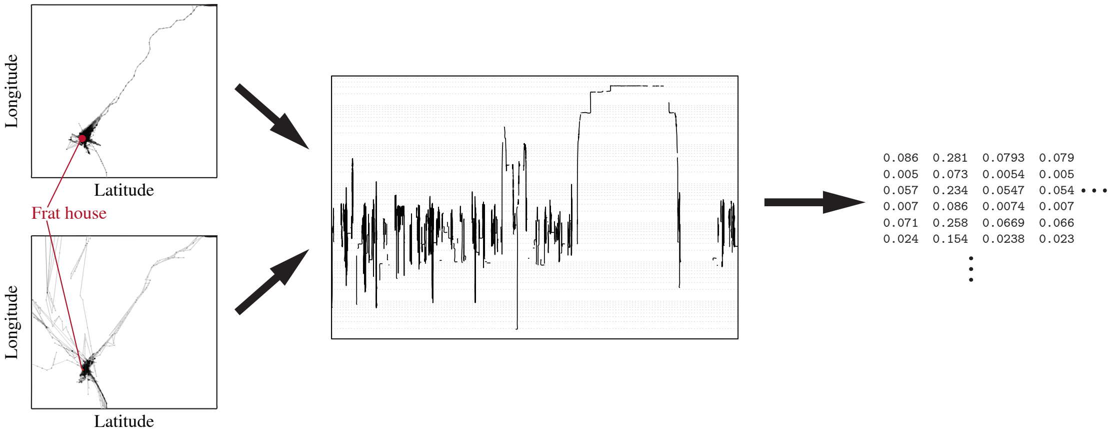


Figure 5.5: A visual representation of what my data processing looked like: I merged data about location over time (left) into pairwise irregular time series (center), on which I performed feature extraction to get various numerical summaries (right).

minute intervals (unlike even five minute intervals) there were few observations that needed to be averaged. In case there was more than one observation within a one-period bin for one of the two people, I chose the more accurate observation for the one person (usually, multiple observations within such a small time period was because of low-accuracy observations). If both individuals had multiple observations within a one-minute period, I chose the row of data with the lower average accuracy for both individuals. If there were multiple equally-accurate measurements, only then did I take the average of the pairwise distances.

Note that I regularized the pairwise distance time series, rather than the original time series of latitude and longitude, because the former approach would have meant two regularizations that potentially would have added more bias.

The pairwise distance time series were the main target of feature extraction. However, as part of creating these time series, I also included columns for whether the location of both individuals fell within a geobox around the fraternity house, and if the location of both were within a geobox around the university area (using the same handling for multiple observations within a one-minute period, carrying through missing values, etc.). Each of these columns gives a binary time series.

The WiFi time series were simpler to process; here, instead of haversine distance, I used an indicator function that was 1 if the two smartphones detected a WiFi hotspot in common (calculated as if the intersection of two lists was nonempty), and 0 otherwise (there were few instances of multiple WiFi devices being detected in common, so I determined it was not worthwhile to make the time series be of the count of hotspots detected in common). To speed up the calculations, I pre-reduced the WiFi devices listed as part of each individuals' time series to only those WiFi devices detected in common (i.e., forming a two-mode network aggregated over the entire time period, and removing WiFi devices of degree 1). Because sampling was at regular intervals, the processing of regularizing the time series was also cleaner (there were no conflicting rows that needed to be aggregated with some rule; sometimes multiple WiFi devices detected within 10 minute periods in multiple rows of data, but I just aggregated them into the same list). While my attempt at WiFi fingerprinting proved unhelpful, I did generate an additional binary time series of whether both people detected in common a WiFi hotspot visible from within the fraternity house.

Distance thresholding

One particularly common task in simplifying sensor data is to set some kind of *threshold*: for sensor nodes, this might be of the signal strength necessary to count a detection as a co-location event, or it might be a minimum span of time, or a choice of width of a sliding time window. In the sensors literature, these decisions are usually not explained in any detail; it is likely that these are hand-tuned to a point that generates 'reasonable' results (i.e., not too dense, and not too noisy), although I note that such heuristic reduction and thresholding is not uncommon in social network analysis, for example when choosing a threshold weight below which to not visualize weighted edges of a completely connected network. An exception to the usual heuristic approach is that of Clauset and Eagle (2012), who seek out a 'natural time scale' for discretizing the dynamic sensors measurements of the Reality Mining data, estimating the rate at 4.08 hours.

In my case, I had continuous-valued pairwise distances, so I wanted to explore whether some threshold or thresholds for *distance* may be particularly informative for detecting friendships. We could hypothesize that a small enough threshold would pick up times of interactions and thus become a reliable proxy for friendship processes; but given the relatively coarse geographical accuracy of the location measurements, a low threshold might prove too noisy. Alternatively, maybe a threshold corresponding to living on-campus versus living off-campus would fall within the scale of accurate detection, and would pick up another type of latent similarity that correlates with friendship processes. Given the scale of the data, I take a data mining approach to explore, rather than generating and testing specific hypotheses about what might be meaningful thresholds. I empirically chose thresholds in order to generate candidates whose predictive power I can then explore through feature selection.

First, I plot an empirical complementary cumulative distribution function (i.e., a survival function) in log-x scale (fig. 5.6a) to better see the overall distribution of pairwise distances over all pairs. There is an 'elbow' around 2000m, which is about the size of the university and surrounding area. Within 2000m, I do one-dimensional clustering (Wang and Song, 2011) of the distances, weighted by the amount of time spent at those distances to find thresholds. I show these empirically fitted clusters over a kernel density estimate of the top 2000m of the distribution, this time plotted in linear scale (fig. 5.6b). Discretizing at each of these thresholds gives another binary time series, again measured at a frequency of 1 minute.

Note that I am re-using data; I am using the same data to do the distance clustering as I am to do the actual modeling. This is not best practice as any sort of data re-use can lead to overfitting in subtle and unexpected ways; I justify it primarily by appealing to the overall exploratory nature of the process, and by noting that there isn't a clear way to split out a portion of the data to use only for this clustering task that would both give a good enough representation of the distribution of pairwise distances and that would not take away too much data for use in the actual model.

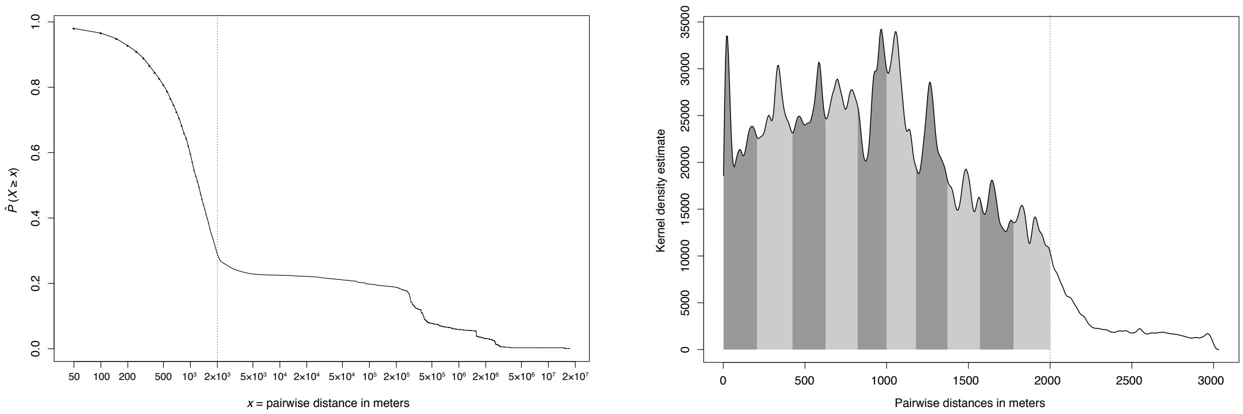


FIGURE 5.6: A survival function (left), plotted in log-x scale, shows pairwise distances over time. Based on the 'elbow' around 2000m (approximately the size of the university and surrounding area), marked with a vertical dotted line, I only found clusters for pairwise distances below 2000m. Below 2000m, I clustered distances (again weighted by the time spent at that distance). The fitted clusters are shown on top of a kernel density estimate (right) that gives a detail of the head of the distribution. The cluster breaks are at 207m, 422m, 626m, 822m, 1001m, 1178m, 1373m, 1570m, 1776m, and then my cutoff of 2000m. These are also listed in table (5.2).

After data processing and the thresholding, I have the following:

- Continuous-valued time series of *pairwise distances*
- Binary time series of whether both members of a given pair were within a geobox around the university campus
- Binary time series of whether both members of a given pair were within a geobox around the fraternity house
- 10 binary time series of whether the two members of a given pair were within a certain threshold of each other, for each of 10 thresholds
- Binary time series of whether both members of a given pair detected at least one Wifi hotspot in common

- Binary time series of whether both members of a given pair detected a Wifi hotspot visible from the fraternity house in common

Next, it is necessary to summarize four week spans of relational geolocation data (4000 observations for each distance-based time series, or 800 observations for the WiFi-based time series), into a set of features that I can feed into a statistical model.

5.4.2 Feature extraction

My feature extraction is summarized in table (5.2).

The first set of features are calculated from the continuous-valued time series of pairwise distances. I both calculate various usual summary statistics on the distributions, but also on logarithmic transformations of these distributions, as I found many were right-skewed (heavy-tailed) but became reasonably symmetric under a logarithmic transformation.

The second set of features are calculated on each of the 10 binary time series I get from thresholding, and are based off of summary statistics for Bernoulli variables (e.g., the maximum likelihood estimator of the variance of draws $X_1, ..., X_n$ of a Bernoulli random variable is $\overline{X}(1 - \overline{X})$, rather than $n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ for normal distributions, although if the random variable is overdispersed and not Bernoulli, the usual estimator for standard deviation may capture different information than does the estimator of Bernoulli variance).

Lastly, for each binary time series, we can consider the length of sequences of consecutive 1s (*spans* of co-location at the given threshold) and of consecutive 0s (*gaps* between co-location at the given threshold). These are integer-valued but I treat them as continuous, and calculate summary statistics accordingly. I found these distributions to also frequently be right-skewed, so I calculate usual summary statistics on a logarithmic transformation.

Each of these feature types are crossed with time periods: weekdays only and weekends only, nights only (12am - 6am), mornings only (6am - 12pm), afternoons only (12pm - 6pm), and evenings only (6pm - 12am). The choice of 6 hour (quarter-day) spans was heuristic, based on my sense of hours that characterize different activities in campus life (sleeping, classes, homework and extracurriculars, etc.), with the location of break points at 6am, noon, 6pm, and midnight. I did explore other intervals, like 12 hours (half-day) and 8 hours (one-third-day), finding that they did not give better results, but this exploration was not systematic. Trying different time intervals could be an avenue for future testing, or even for using the finding of Clauset and Eagle (2012) that 4 hours is a natural time scale.

In total, there are $9 + 12 \times (5 + 20) = 309$ features, each taken over seven settings, for $309 \times 7 = 2163$ candidate location features. WiFi features were $2 \times (5 + 20) = 50$, and $50 \times 7 = 350$ for an additional 350 features, for a total of 2,513 features. I extracted these over two 4-week periods, corresponding to the 4 weeks between surveys 1 and 2, and the 5 weeks between surveys 2 and 3 with the week of spring break subtracted out.

There were two sources of missing values in the calculated features: either artifacts relating to no observations fulfilling a certain criteria (e.g., no co-locations within 626m on mornings), or else actual missing data

Distribution to summarize		**Statistic**		**Timeframe**
Pairwise distances (continuous-valued)	⊗	Mean Median Standard deviation Mean of logarithm Median of logarithm Standard deviation of logarithm Mean of inverse-squared distance Median of inverse-squared distance Standard deviation of inverse-squared distance	⊗	4-week period Weekdays only within 4-week period Weekends only within 4-week period Mornings [6am - 12pm) only within 4-week period Afternoons [12pm - 6pm) only within 4-week period Evenings [6pm - 12am) only within 4-week period Nights [12am - 6am) only within 4-week period
Pair is within 207m (binary) Pair is within 422m (binary) Pair is within 626m (binary) Pair is within 822m (binary) Pair is within 1001m (binary) Pair is within 1178m (binary) Pair is within 1373m (binary) Pair is within 1570m (binary) Pair is within 1776m (binary) Pair is within 2000m (binary) Pair is within geobox around campus (binary) Pair is within geobox around fraternity house (binary) Wifi device detected in common (binary) Wifi device in fraternity house detected in common (binary)	⊗	Count of times within threshold Mean of count Standard deviation of count Standard deviation squared of count Count(1 - Count) *Spans (distribution of lengths of consecutive 1s)* Mean of lengths Median of lengths Standard deviation of lengths Minimum length Maximum length Mean of logarithm of lengths Median of logarithm of lengths Standard deviation of logarithm of lengths Minimum of logarithm of lengths Maximum of logarithm of lengths *Gaps (distribution of lengths of consecutive 0s)* Mean of lengths Median of lengths Standard deviation of lengths Minimum length Maximum length Mean of logarithm of lengths Median of logarithm of lengths Standard deviation of logarithm of lengths Minimum of logarithm of lengths Maximum of logarithm of lengths	⊗	

TABLE 5.2: Extracted features. "⊗" indicates taking all pairwise combinations. The thresholds are irregularly spaced because they are empirically derived from 1-dimensional clustering; see fig. (5.6b) for these clusters.

(one or both mobile devices were not providing a certain sensor's data during a given period, e.g., mornings of a given period). For the former (artifacts), I replaced missing values with appropriate substitutes, such as 0s or the maximum possible value. For logarithmic features, some of which could be less than 1 (but always greater than 0), I replaced $-\infty$ with zeros. For inverse-squared features, I replaced ∞ with a value, 200, slightly larger than the largest observed inverse-squared value. For the missing values resulting from an actual lack of data, I kept the missing values in the cells of the feature matrix. This necessitated using classifiers that can handle missing values among the features, like the R random forest implementation `rpart` (Therneau and Atkinson, 2018) which has procedures for handling NAs when constructing decision trees, and other packages built on top of `rpart`.

Alternative approaches

In addition to the main approach described here, I also looked at interpolation alternatives. Time series interpolation, performed prior to feature extraction, did not improve performance. Interpolation on the extracted features also did not improve performance. Trying thresholds of equal width (specifically, trying thresholds from 50m to 500m in 50m increments, or 200m increments from 600m to 2000m as looking at fig. 5.8 below would suggest) also did not improve performance.

5.4.3 Modeling targets

In the models I set up, I use sensor data from weeks 1-4 to predict the self-reported friendships in survey wave 2 (given to study participants at the end of week 4), and sensor data from weeks 5-9 (excluding the week of spring break), to predict the self-reported friendships in survey wave 3 (given to participants at the end of week 9). This is a standard binary classification task. In this task, I do not make use of survey wave 1.

Next, I can treat changes in friendship (whether tie creation or dissolution) as another binary classification task, with targets

- $P(A_{ij}^{(t)} = A_{ij}^{(t+1)} \mid X^{[t,t+1]})$: No change in friendship (either no friendship, or maintained friendship)
- $P(A_{ij}^{(t)} \neq A_{ij}^{(t+1)} \mid X^{[t,t+1]})$: Change in friendship (either tie creation or tie dissolution)

While I ideally would be able to separately model tie creation and dissolution, as they are distinct processes Snijders, Bunt, et al., 2010, in my data only a small proportion of ties changed in either direction such that modeling became difficult. We will see below that my results for this task were poor, although treating it as a multiclass problem over the direction of change only led to worse performance.

Again, I collected data specifically to use in SAOMs; the level of change here makes the machine learning extremely difficult for changes being the specific modeling target, but are fine for a SAOM. Specifically, the RSIENA manual (Ripley et al., 2017) suggests that for SIENA estimation, “Jaccard values of .3 and higher are good; values lower than .2 indicate that there might be difficulties in estimation; values lower than .1 are quite low indeed”, and changes are all well above .75 (fig. 5.7).

5.4.4 Cross validation schema

As described above, machine learning approaches can suffer from overfitting and the multiple comparisons problem. To address this, machine learning uses held-out data as a way to simulate out-of-sample data; if a machine learning classifier were to perform well on out-of-sample data, it would directly establish external validity and therefore the validity of using the model for prediction, even if it lacks construct validity, interval validity, or other types of validity of concern for explanation and intervention. Alternatively, no matter how well a model fits to the data on which its parameters are fit,

But also as described above, dependencies within data can cause information from held-out data to leak back into the data used for modeling fitting. To address this, I try out multiple cross-validation schema: that is, different ways of splitting data to try and respect some of the dependency structures. The rules and use cases of these schema are detailed below.

Cross validation with unrestricted assignment

This is independently assigning each observed A_{ij} to a fold. It corresponds to a use case where a model is trained on a population ($n - k$ pairs) and then applied back to k pairs from same population (potentially seeing the same people multiple times, or the same dyad in multiple directions).

Cross validation with dyadic assignment

This groups all values associated with a pair of individuals (a dyad), that is, $(A_{ij}^{(1)}, A_{ji}^{(1)}, A_{ij}^{(2)}, A_{ji}^{(2)}, A_{ij}^{(3)}, A_{ji}^{(3)})$, and assign the entire 6-tuple to a single fold. Some values in the tuple will be missing, causing folds to be of different sizes; But since assignment to fold is not dependent of the number of missing values, sizes will be the same in expectation.

Such assignment controls for reciprocity and temporal autocorrelation. For reciprocity, if $A_{ij} = A_{ji}$, then the label-feature pair (A_{ij}, X_{ij}) and (A_{ji}, X_{ji}) are identical and should not be split between training and test. Similarly for temporal autocorrelation, if two people's friendship and co-location patterns do not change over time, then $(A_{ij}^{(t)}, X_{ij}^{[t-1,t)})$ and $(A_{ij}^{(t+1)}, X_{ij}^{[t,t+1)})$ would also be very similar and should not be split between training and test.

Cross validation with dyadic assignment corresponds to a use case where we have not previously seen the labeled co-location patterns of a given dyad, whether previously in time or in one direction, to have included it as a training instance.

Cross validation with temporal block assignment

This splits data by whether a class label is from survey 2 or survey 3 (for detecting friendship and strength of friendship) or is the change from survey 1 to 2 or the change from survey 2 to 3 (for detecting change in friendship). In other words, for detecting friendship and strength, I train on $(A^{(2)}, X^{[1,2)})$ and test on $(A^{(3)}, X^{[2,3)})$, and for detecting change, I train on $(A^{(1)}, A^{(2)}, X^{[1,2)})$ and test on $(A^{(2)}, A^{(3)}, X^{[2,3)})$.

As a note, here I can only split into 2 folds as I only have two observation spans between different surveys.

Cross validation with temporal block Bergmeir and Benítez, 2012; Racine, 2000 assignment accounts for temporal variation in co-location. If there is a great deal of variability in co-location patterns, then my classifier would have little generalizability over time. In this case, if I train with instances with features from both $X^{[t-1,t)}$ and $X^{[t,t+1)}$, it would even out the temporal variation and obscure the lack of generalizability. But if I train only on instances associated with features $X^{[t-1,t)}$ and then test only on instances associated with features $X^{[t,t+1)}$, it simulates how well out classifier will do in predicting friendships from future patterns of co-location data.

5.4.5 Evaluation metric

To summarize classifier performance, I rely on the Matthews correlation coefficient (MCC). This is the same as Pearson's ϕ, or mean square contingency coefficient, an analog for a pair of binary variables of Pearson's product-moment correlation coefficient, but was rediscovered by Matthews Matthews, 1975 for use as a classification metric. For the count of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the MCC is

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(FP + TP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

The MCC has several desirable properties. First, like the F1 score and area under the ROC curve (AUC), it summarizes the performance on both classes in a single number. Unlike AUC and F1, however, it has an interpretable range: 0 for random predictions, -1 for perfect misclassification, and 1 for perfect classification. Most helpfully, it is a good summary of performance in cases of class imbalance Boughorbel et al., 2017, which have here (about a 25:75 split). I include other metrics, but rely on the MCC as the single-number summary of how far I am above a random baseline of MCC = 0. Note that if I predict the majority class for all instances, the MCC is also zero.

5.4.6 Feature Selection

As mentioned previously, feature selection can be a useful diagnostic and exploratory analysis, although interpretations must always remain speculative and cautious. To produce a selected set of features, I use Correlation-based Feature Selection (CFS) Hall, 1999, which selects features that are both correlated with the class label, and uncorrelated with one another. Unlike other feature selection methods, such as methods that use the variable importance scores from random forests, or feature selection from the lasso or stability selection (Meinshausen and Bühlmann, 2010), CFS to my knowledge has never been analyzed from a statistical perspective, for example to examine its consistency. It is preferable to use models whose statistical properties are understood; however, implementations of the lasso do not have ways of handling missing data which I felt it was substantively important to preserve; but as discussed above, I did also try interpolation on the extracted features in order to be able to use usual variable selection methods, but the selected features were those that had been most heavily interpolated, which was not a confidence-inspiring result. In contrast to other methods I tried, CFS gave results that were both excellent in terms of performance, and that cut down the features to a subset small enough to be interpretable.

Implementations of CFS in R also require interpolation; the implementation in the Java-based machine learning library WEKA (Hall et al., 2009; Frank et al., 2016), however, handles missing data. I imported the training instances of the feature matrices from temporal block assignment into WEKA, had CFS run to get the names of selected features, and re-imported these into R to continue the modeling. I chose to use the training set built temporal block assignment as it proved to be the most conservative CV schema (see below).

For using CFS, I took the half of data with features extracted from the first four weeks, and further divided it into 10 folds. After running CFS on each fold, I look at the features that were selected in multiple folds, taking inspiration from stability selection. I choose those features that appeared in CFS runs on at least 9 of the 10 folds.

5.5 Results

5.5.1 Networks comparison

Looking at the changes in the networks from survey to survey, we see that close friendships and both types of advice-seeking relationships are much less variable over time than are friendships or self-reports of frequent interaction.

5.5.2 Robustness of apparent patterns

Eagle, Pentland, and Lazer (2009) present a visually striking graph that shows systematic differences in the ‘probability of proximity’ between mutual friendships, non-mutual friendships, and non-friends, when that probability of proximity is aggregated into a one month interval from 9 months of data. I replicate this plot in figure (5.8), although what I plot is the median distance, rather than the probability of proximity. The image is equally striking, appearing to give a strong pattern.

However, when trying to use this apparent pattern as a basis for classification, and specifically for training on data in one temporal block and testing on data in another (chronologically later) temporal block, proved to give poor results. The aggregation over a long period, and the averaging process that disguises the amount of variation of those averages, is an ineffective basis for results that would be useful in an application setting, for example actively trying to predict unknown friendship status/perceptions based on previously collected co-location data. The pattern is visually striking but is not good enough for classification.

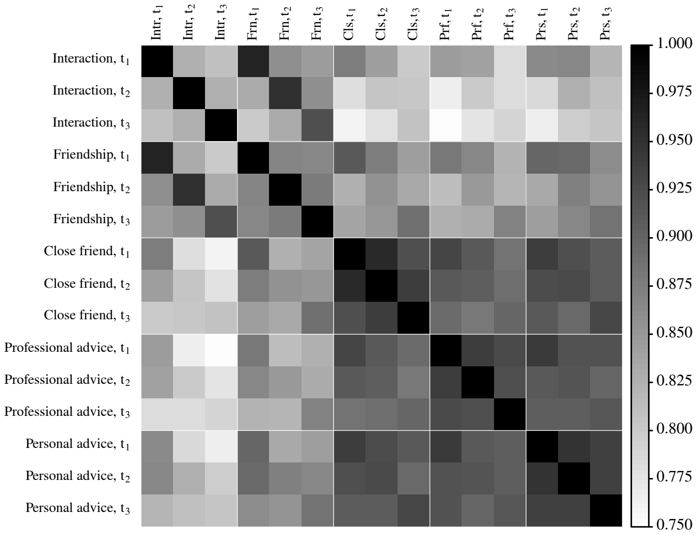


FIGURE 5.7: The similarity between networks (5 types of ties, each collected 3 times), measured via the Jaccard index. The self-reported frequent interaction and friendship networks are more similar than the other networks, and both also exhibit more variation across the three waves.

5.5.3 Friendship detection

Results for the three cross-validation schemes are given in table (5.3). In each case, the no information rate corresponds to the proportion of the majority class, 0, and would be the accuracy I would get if I always predicted no tie.

The unrestricted assignment gives better results than either of the other two CV schema, showing that labeling a previously unseen dyad is indeed a more specific and difficult task than what is evaluated by unrestricted assignment, and that there is a significant amount of variation in co-location patterns over time—and that while my classifier performance does drop, it still generalizes across patterns in time.

I use a one-sided binomial test of the accuracy against the No Information Rate (NIR), equal to the frequency of the majority class, and find that both unrestricted and dyadic CV are significant at the usual $p < 0.05$ level. Under temporal block CV, the classifier is only significantly better than the NIR at the $p < 0.1$ level.

In my classifications, the MCC ranges from .30 in CV with unrestricted assignment, to .26 in CV with dyadic assignment, and .21 in CV with temporal block assignment. This indicates that the classifier performance is between 30% and 21% better than baseline (for which MCC=0).

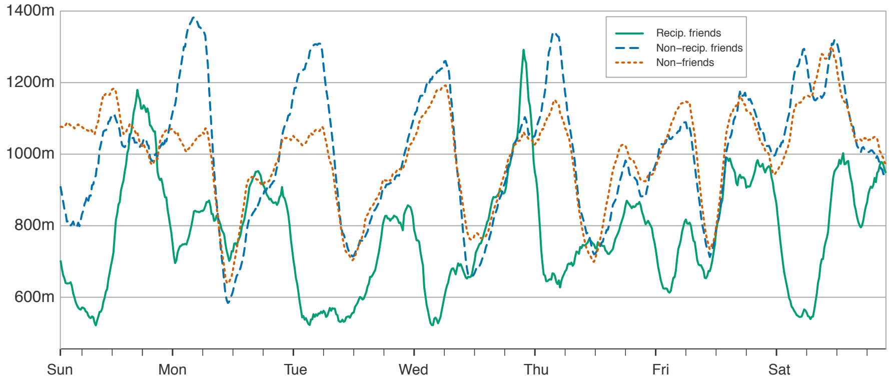


Figure 5.8: The median weekly pairwise distances between reciprocated (mutual) friendships, $A_{ij} = A_{ji} = 1$, non-reciprocated friendships, $A_{ij} = 1 \neq A_{ji} = 0$ or $A_{ij} = 0 \neq A_{ji} = 1$, and non-friendships, $A_{ij} = A_{ji} = 0$, for times when pairs are within the area of the university, and aggregated over the entire period of data (i.e., no training/test split). This is analogous to the approach of Eagle, Pentland, and Lazer (2009), and this figure reproduces their figure 2 (except with median distance, rather than mean frequency of proximity). While it appears there is a strong pattern, it is a result of an aggregation that obscures the variance between weeks and among various pairs, such that this seeming pattern proved ineffective as a basis of classification in testing.

Cross validation	Unrestricted	Dyadic	Temporal block
Accuracy	0.8006	0.7920	0.7913
Accuracy, 95% CI	(0.7882, 0.8125)	(0.7794, 0.8042)	(0.7726, 0.8091)
(No Information Rate / Majority class)	*(0.7740)*	*(0.7740)*	*(0.7785)*
Binomial test, Accuracy vs. NIR, *p*-value	*p*=1.5e-05	*p*=0.0025	*p*=0.0901
Precision (Positive predictive value)	0.6918	0.6508	0.6812
Recall/Sensitivity (True positive rate)	0.2122	0.1723	0.1088
Specificity (True negative rate)	0.9724	0.9730	0.9855
F1 score	0.3248	0.2724	0.2964
AUC	0.7148	0.7039	0.1876
Matthews correlation coefficient	0.3039	0.2562	0.2120

TABLE 5.3: Friendship detection, test performance across the three CV schema. The no information rate corresponds to a baseline accuracy given by predicting no ties; in the case of networks, this is 1 minus the density of the network.

5.5.4 Detecting close friendships

I repeat the assessment of the above models, conditioning on the presence of a friendship, and making my detection target whether or not a friendship is reported to be *close*. In this case, the network of close friendships has a network density of .41, making the no information rate .59.

Cross validation	Unrestricted	Dyadic	Temporal block
Accuracy	0.6817	0.6670	0.5741
Accuracy, 95% CI	(0.6511, 0.7112)	(0.6361, 0.6969)	(0.5259, 0.6212)
(No Information Rate / Majority class)	*(0.5861)*	*(0.5861)*	*(0.5185)*
Binomial test, Accuracy vs. NIR, *p*-value	*p*=7.6e-10	*p*=1.8e-07	*p*=0.0117
Precision (Positive predictive value)	0.6904	0.6711	0.7069
Recall/Sensitivity (True positive rate)	0.4188	0.3832	0.1971
Specificity (True negative rate)	0.8674	0.8674	0.9241
F1 score	0.5213	0.4879	0.3083
AUC	0.6997	0.6695	0.5889
Matthews correlation coefficient	0.3250	0.2906	0.1777

TABLE 5.4: *Close* friendship detection, conditioned on the presence of a friendship, test performance across the three CV schema.

We see a similar pattern of performance, with temporal block CV being the most conservative (18% better than baseline), and unrestricted CV being more optimistic (32% better than baseline).

5.5.5 Detecting changes in friendship

Detecting loss in friendships could be particularly important for social interventions, such as preventing the onset of isolation. However, the rarity of changes in friendship (only 13% of ties change, either being created or dissolving) complicates modeling.

My approach in meaningfully detect changes in friendship proved to be challenging. AdaBoost failed to predict any positive test cases for any CV schema; a random forest performed better with a Matthews correlation coefficient of .07 for the unrestricted CV and .03 for the dyadic-based CV (see table (5.5). The classifier output does not pass a statistical test for being significantly better than the No Information Rate. One of the reasons for the poor performance may be the type of features used in the classification. I used the same aggregated features used for friendship detection to detect change. However, change in friendship may be reflected in the feature values and thus a feature set that contains change values may better capture change in friendship.

Cross validation	**Unrestricted**	**Dyadic**
Accuracy	0.6842	0.8645
Accuracy, 95% CI	(0.6692, 0.6989)	(0.8532, 0.8752)
(No Information Rate / Majority class)	*(0.8710)*	*(0.8710)*
Binomial test, Accuracy vs. NIR, *p*-value	*p*=1	*p*=0.8902
Precision (Positive predictive value)	0.1676	0.2093
Recall/Sensitivity (True positive rate)	0.3651	0.0183
Specificity (True negative rate)	0.7315	0.9898
F1 score	0.2297	0.0336
AUC	0.5483	0.5167
Matthews correlation coefficient	0.0720	0.0256

TABLE 5.5: *Change* detection, random forest test performance. AdaBoost made only negative test classifications, but random forests (performance shown here) did make some positive classifications under unrestricted and dyad-based CV, although under temporal block CV again there were no positive classifications.

5.5.6 Feature Selection

While I applied CFS to select features from the training set in all tasks, the features selected were not always consistent across folds, and across cross validation schema. So, I focus on the features selected in the case of the most conservative cross validation schema, and the extent to which feature selection improved model performance here.

Applying CFS to only the training data from temporal block assignment and splitting it into 10 folds, I find 19 features that are selected in 9 or 10 of the folds. Using only these features leads to improved test performance from temporal block assignment, shown in table (5.7), which also includes the test performance with this set of features under each cross validation scheme.

While the test MCC of CV with unrestricted assignment goes down, with this fraction of only 19 features the test MCC of CV with dyadic assignment rises slightly, and the test MCC of CV with temporal block assignment does far better, going from an MCC of .21 to .27. These 19 features, then, seem to be picking up a significant portion of the pattern in co-location data, and a pattern that is more robust to changes over time.

While again, it is dangerous to substantively interpret the selected features as causal or even as necessarily stable (Mullainathan and Spiess, 2017; Yang and Yang, 2016), it is a useful exploratory step to see the features that are effective for the detection task. The features are listed in table (5.6) ,with the pairwise correlations given in figure (5.9). While there are groups of highly linearly correlated features, many of the features are not correlated, giving an independent signal.

There are some patterns that emerge in this well-performing subset of features. Threshold 2 (422m) shows up frequently, as do measures related to variance (standard deviation measures), nighttime, and the distribution of inverse squared distances. This generates several hypotheses: first, that Latané et al. (1995) finding that

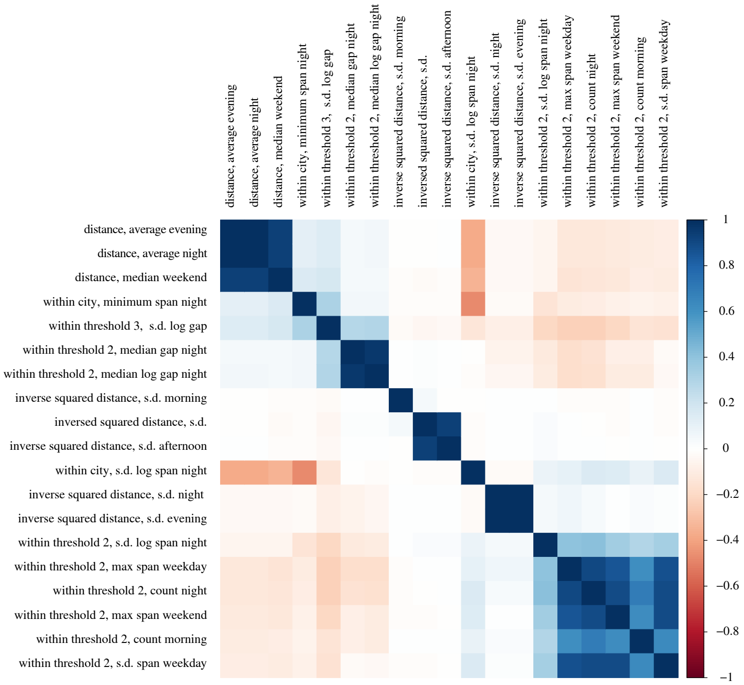


FIGURE 5.9: Correlations between the features selected via CFS on the training set of a temporal block cross-validation scheme. The ordering is from the angular order of eigenvectors.

inverse-squared distance fits well to reports of memorable social interactions may be effective for friendship detection as well. Second, the threshold at 422m seems particularly relevant versus others: this specific value might not be what is important, but perhaps this captures some relevant radius around the frat house. Otherwise, features associated with where people are co-located at night appear most frequently, which is in contrast to the finding by Eagle, Pentland, and Lazer (2009) that the daytime probability proximity is what was discriminative for friendships.

5.6 Discussion

For interpreting my results as the extent to which co-locations can detect friendships, we can treat the MCC as a correlation coefficient, and say that I have demonstrated that I can find a 0.2 to 0.3 correlation between a signal extracted from sensor data and self-reported friendships. For this application setting, this is probably a more useful description than the accuracy to which co-locations can detect friendships, or from seeking to identify the percentage of variance explained by my model (e.g., with pseudo-R^2 metrics). It would be imprecise to say that I find that 20-30% of friendships can be detected using co-location data, but intuitively, in a more general sense we can think of this finding saying that 20-30% of the overall phenomenon of friendship (although measured dyadically only, not with whole-network metrics like in goodness-of-fit testing, and measured via self-report) can be captured by co-location alone. This provides a strong argument for

Feature	Distribution	Summary statistic	Timeframe
1.	Distance	Mean	Evening
2.	Distance	Mean	Night
3.	Distance	Median	Weekend
4.	Within city	Minimum span	Night
5.	Within threshold 3	Log gap	All
6.	Within threshold 2	Median gap	Night
7.	Within threshold 2	Median log gap	Night
8.	Inverse squared distance	S.D.	Morning
9.	Inverse squared distance	S.D.	All
10.	Inverse squared distance	S.D.	Afternoon
11.	Within city	S.D. log span	Night
12.	Inverse squared distance	Standard deviation	Night
13.	Inverse squared distance	Standard deviation	Evening
14.	Within threshold 2	S.D. log span	Night
15.	Within threshold 2	Max span	Night
16.	Within threshold 2	Count	Night
17.	Within threshold 2	Max span	Weekend
18.	Within threshold 2	Count	Morning
19.	Within threshold 2	S.D. span	Weekday

TABLE 5.6: The 19 features selected via CFS on the training set from temporal block assignment: what they measure, how they summarize it, and the timeframe in which they summarize it. Ordering is from angular order of eigenvectors on the correlation matrix (fig. 5.9).

studying propinquity alongside other processes, as a contributor to—but not the sole cause of—friendship formation and maintenance, and demonstrates the ways in which sensor-based measurement can help study propinquity. It seems that Latané et al.'s (1995) inverse-squared transformation of distance is useful, and furthermore that explanatory models might explore focusing specifically on proximity at nighttimes as more important than proximity at other times.

I can also say that using co-locations, it is possible for companies to make predictions about friendship ties, although such predictions would be extremely noisy and only capture about 10-20% of friendships ('recall' is number of true positives versus number of positive cases), and conversely, it will be wrong about 30% of predicted friendships. This is not perfect prediction, but is likely a sufficient basis for commercial experimentation, although even this is assuming first that there was some training set for a given social system on which to train a model, and second that the boundary of a social system is either clear or can be identified. In practice, iterating on an initial model by using implicit feedback from users (e.g., seeing how they react to recommendations or proposals made on the basis of predicted friendship ties) to refine the model, might lead to better performance over time.

5.7 Conclusion

In Chapter (3), I argued that Stochastic Actor-Oriented Models have the most value for modeling sensor data, as they are able to capture multivariate associations, can express network processes, and can model

CV assignment method	**Unrestricted**	**Dyadic**	**Temporal block**
Accuracy	0.7975	0.793	0.7923
Accuracy, 95% CI	(0.785, 0.8095)	(0.7804, 0.8051)	(0.7736, 0.8101)
(No Information Rate / Majority class)	*(0.774)*	*(0.774)*	*(0.7785)*
Binomial test, Accuracy vs. NIR, *p*-value	*p*=0.0001	*p*=0.0016	*p*=0.0734
Precision (Positive predictive value)	0.6602	0.6370	0.5799
Recall/Sensitivity (True positive rate)	0.2143	0.1954	0.2269
Specificity (True negative rate)	0.9678	0.9675	0.9532
F1 score	0.3236	0.2990	0.3261
AUC	0.6837	0.6804	0.6767
Matthews correlation coefficient	0.2921	0.2682	0.2658

Table 5.7: Friendship detection with CFS feature selection on the temporal block assignment training data.

co-evolution the of networks and behavior. However, a major challenge is in the scale of sensor data: sensor data are far more dense than many outcomes of interest, so in order to use sensor data in a SAOM or other network model, we need to summarize the sensor data. In contrast to some of the other approaches, such as how Pachucki et al. (2015) conflate constructs captured in friendship self-report and in RFID data by their use of friendship data to calibrate sensor data, I work explicitly with theory about the relationship between friendship and proximity, and find ways of characterizing sensor data in ways that best match friendship data. This both has intrinsic value, for application in friendship detection tasks, and in forming the first step of being able to use sensor data with network models.

Conclusion

Thesis statement, revisited

Returning to my thesis statement:

> Social media and sensor data do not give unbiased, generalizable findings about human behavior: inferences about constructs are complicated by selection bias, medium-specific norms and culture, and algorithmic user manipulation, and raw measurements are of physical quantities rather than of causal underlying social constructs. But by studying these forms of bias and the data-generating processes of such data and understanding their limitations, we can establish proper scopes and study designs within which findings will be accurate, reliable, and fair for use in business decision-making, scientific research, and public policy.

In Part I, I take up the first part of this statement: how inferences are complicated by selection bias, medium-specific norms and culture, and algorithmic user manipulation, and how raw measurements are of physical quantities rather than of causal underlying social constructs.

In Chapter (1), I convincingly showed that geotagged tweets do not represent the US population. While my findings about specific demographic biases are potentially sensitive to model specification, timeliness of data, and choices of filtering mechanisms, there is little doubt that there are statistically significant biases versus a baseline of a random distribution over the population. If we select geotag tweet users to study some construct of interest, we are introducing selection bias into our studies.

The substantive significance is that relying on geotagged tweets will only capture the behaviors of a select, nonrandom few, thereby biasing conclusions we might try to make about either the behaviors of national or international populations, or even of just the Twitter-using population. The dangers are the same as in how nonrandom sampling led *Literary Digest* to, in 1936, predict a Republican presidential victory only to have the Democratic candidate win with 61% of the vote (Gayo-Avello, 2011), or how similarly sloppy sampling (and a haste to publish results) led to the infamously wrong ‘Dewey Defeats Truman’ headline of the *Chicago Daily Tribune* (Ruths and Pfeffer, 2014). Relying on geotagged tweets risks making nongeneralizable conclusions when used for basic scientific research, incorrect predictions when used for planning, and misdirection of resources when used for policy purposes such as in disaster response (as also discussed in Shelton et al., 2014) or urban planning.

While the finding around geotagged tweets does not automatically extend to every source of social media (or every possible use case), I argue that geotagged tweets are intrinsically important because their combination of linguistic, temporal, geographic, and social network data (plus the relative availability of tweets via the Twitter API, versus other social media platforms) has made them enormously attractive as a test bed for any number of topics and themes. Certainly, even without the geographic data of geotags, tweets are highly valued data. Two recent compilations, *Twitter and society* (Weller et al., 2014) and *Twitter: A digital socioscope* (Mejova et al., 2015) take up this theme, and Tufekci (2014) suggests that Twitter is the *Drosophila melanogaster* of big data in the sense that it is a "model organism", but also in the sense that the very factors that make model organisms like *Drosophila melanogaster*, *Escherichia coli*, *Caenorhabditis elegans*, or *Mus musculus* useful for biological research—like rapid life cycles and low variability in development—also include drawbacks, like insensitivity to environmental influence, and sometimes a failure to generalize to other organisms that do not have rapid life cycles or as low variability in development. Beyond the specific case of geotagged tweets, my work serves as an empirical 'existence proof' about demographic biases in rich social media data, supporting the theoretical and smaller-scale empirical argument about why patterns of adoption and adoption mean that our baseline should be that such data are biased and therefore not representative (there have been other such existence proofs, but mine is the first done through nation-wide, multivariate, spatial modeling).

The medium-specific norms and cultures that impact generalizability come through implicitly in this and other chapters; my observational finding about the high rate of geotagged tweets sent from airports dovetails with Tasse et al.'s (2017) finding via surveys that users use geotagged tweets as "postcards, not ticket stubs". Geotagged tweets are selective, deliberate broadcasts about specific locations made for social purposes, and not an automatic and passively made record of all movements. They should not be, for example, taken as analogous to footprints as other forms of digital trace data are argued to be (Golder and Macy, 2014). In Chapter (4), the topic-specific usage of certain emojis (e.g., the "dash" emoji to represent smoke) show a medium-specific norm that, rather than the medium being a social media platform, being a particular *topic* discussed on that and other social media platforms (as we can imagine the same convention being used in, say, text messages).

Next, platform design and engineering act on constructs that may be the interest of scientific study. For example, social network structure, and socializing behavior are examples of classes of constructs that are frequently studied within social network analysis (Borgatti, Mehra, et al., 2009); in Chapter (2), my study of Facebook's "People You May Know" feature via the data gathered by Viswanath et al. (2009) shows how a recommender system had a causal impact on people's networking behavior. This empirically demonstrates theoretical critiques about how platforms are not neutral utilities (indeed, as they frequently aspire to and sometimes claim to be), such as from van Dijck (2013), Gehl (2014), Tufekci (2014), Ruths and Pfeffer (2014), and Healy (2015), among others. The data produced on platforms conflate governance and management processes (or, put more cynically, manipulation) with the intentions and expressions of users, in how user behaviors are constrained and shaped by deliberate platform features.

In Chapter (3) I had an extended discussion of what sorts of relational sensor data there are by which we might study social relationships, and what constructs such sensor data capture. Looking into the language used by various sensor studies, as well as the descriptions of the actual sensor technology, I argued that sensors have been used to measure interaction, when in fact they actually measure proximity (potentially directionally constrained). There are times where proximity is a construct of interest, but it has seldom been recognized as such. The focus has largely been on the less precise (but more theoretically rich) construct of interaction. Measurements of proximity, even if directionally constrained, do not automatically capture a construct of interaction; there is a good argument that directionally constrained proximity is a good *proxy* for measuring interaction, but a proxy is always imperfect. It is easy to think of physical scenarios where such constrained proximity would produce false negatives and false positives when used to measure interaction (I gave the examples of people interacting while sitting side-by-side and facing forward in a car for a false negative, and people sitting across from one another on a subway but not interacting for a false positive). Such thought experiments, where we can hypothesize scenarios where we would expect misclassification, show that human judgement about what counts as an interaction is still the measurement standard to which we are aspiring, and it is a standard against which sensors are not necessarily more accurate. Of course, variability among human observers (and disagreement about how we should demarcate 'interaction') may ultimately make sensors a more *reliable* form of measurement, but we must establish this via study and understand the bounds of this new proxy (identify the situations that lead to false negatives and false positives), rather than taking it for granted.

Once we start thinking in terms of constructs, it also becomes more clear that sensors are not a substitute for getting at constructs that are psychological entities. In cases where such psychological entities are the constructs we care about, sensor data are not superior, but in fact inferior to self-report.

The next part of my thesis statement concerns our ability to understand biases, and follows from considering Part I as a whole. The fact that I and others have been able to study these types of bias in social media data, and that I have laid out a precise understanding of the nature of relational sensor data, means that we can understand limitations of data and correctly design studies and condition our conclusions. For example, given that I can be aware of the presence of recommender systems that seek to change network structure, I can condition claims about networking behavior made on the basis of data from social media platforms on such manipulation. Given knowledge of demographic limitations of geotagged tweets and motivations for their use, I can design studies that, for example, look at how people use geotagged tweets as a form of self-expression, or that try to correct for demographic biases (Zagheni and Weber, 2015).

On that latter point, in the introduction I stated my skepticism that the types of fixes designed for survey data, namely sampling frames that can correct for under-reached populations via oversampling, and/or weighting observations by their (known) proportion in the general population to correct for imbalances in a sample, can be replicated for digital trace data. If we shift the scope of research to not use digital trace data for purposes like urban planning, disaster response, or basic research into mobility, then the limitations of digital trace data do not pose (potentially fatal) threats to validity—but are there ways of shifting the scope of attempts to use digital trace data for such purposes? I believe it is possible to do this if we shift from platform-wide studies to locally embedded cases. For example, the Humanitarian Technologies Project has a number of papers discussing the utility of social media data during recovery after Typhoon Haiyan in the Philippines in 2013 (Madianou, 2015a; Madianou, 2015b; Madianou, Longboan, et al., 2015; Ong, 2015). Based on participant observation, interviews with affected populations, and interviews with experts from humanitarian organizations, local civil society groups, government agencies, telecommunications companies, and digital platform developers, the research team members were able to build an understanding of how certain groups managed to use social media to amplify their voices, and how the purposes for which affected populations used various forms of media were not necessarily the same as what was envisioned by aid agencies (Madianou, Ong, et al., 2015):

> "We observe how platforms that were introduced by aid agencies to facilitate information dissemination and feedback were often appropriated for different purposes by affected people. Such is the case of humanitarian radio which used Frontline SMS for feedback but was largely used for song requests and dedications to friends and family members. We recognize this as an important social function of humanitarian radio and interactive media. Such practices represent a need to affirm relationships in the post-disaster context and a way for people to regain control over their social lives after the disruption of disaster. The uses of media for sociality and recreation are vital for our participants' well-being. While not fulfilling the expectations of 'humanitarian technology', such uses also express a more modest politics of reconnecting to the fabric of public life. Yet, we remain aware that the ordinary uses of new as well as old media, despite their social significance, do not achieve the redistribution of resources which is vital in the aftermath of disasters."

They ultimately recommend shifting "from feedback fetish to cultures of listening", noting that "Digital technologies make it easier to collect and catalogue feedback but can only work alongside processes of needs consultation and agencies' immersion on the ground. Cultures of listening cultivate the participation of communities beyond the promotion feedback tools by developing relationships based on respect and trust." They advocate *ethnography* (rather than, say, social media mining or census-type surveys) as the most appropriate approach to understand needs in the aftermath of disasters. In such cases, if on-the-ground workers understand how and why people generate digital trace data, and how social media platforms (or other digital trace data) may amplify certain voices over others, then such data are absolutely usable as part of general recovery efforts. Media and data are meaningful parts of the fabric of people's lives, and in that sense they should be incorporated into any monitoring or communication.

This unfortunately does not necessarily translate into recommendations about how large-scale organizations (like, say, the UN) might want to prioritize research into and development around digital trace data, or what overall standards organizations should set on the use of digital trace data. I anticipate that building up a larger collection of local case studies like this, and then trying to systematize from those, will be a more reliable approach than a top-down, statistically or computationally motivated determination of what sorts of estimators to use, corrections to make, and systems to build for monitoring or feedback. Statistical and computational work should instead be motivated by qualitatively and theoretically built understandings of use cases, usage, socioeconomic structures, and needs—indeed, I have tried to make my work a demonstration of such motivations. The idea of qualitatively studying the context of digital trace data also ties into the proposal of Wang (2013) I quoted in the introduction: to have 'Thick Data', i.e., to pursue are ethnographic understandings of how data are produced as a counterpart to big data attempts to characterize populations via available data. Prioritizing context will be a good safeguard against being led astray by the ways in which techno-cultural structures or socioeconomic structures introduce selection bias, cause changes in behavior, or contain heterogeneity, and can even identify data-generating processes of biases, changes, and heterogeneity that modeling research can then investigate and quantify.

In Part II, I demonstrated the last part of my thesis statement: that we can establish proper scopes and study designs within which findings will be accurate, reliable, and fair for use in business decision-making, scientific research, and public policy. Inspired by Abbott's (2004) discussion of the powerful social science heuristic of "shifting the question", I recommend shifting the *scope* of studies done with digital trace data.

Specifically, I took another part of Abbott's discussion, his division into "case study", "small-*N*", and "big-*N*" levels of analysis, and I suggested shifting the justification of studies undertaken with digital trace data from the justifications of the "big-*N*" level of analysis to the justifications of the "small-*N*" level of analysis. We need not try to characterize entire populations as the size of the data seems to enable, but rather recognize that our studies are of specific populations and that our generalization must be *logical* (as argued in Luker, 2010), i.e., argued via theory and evidence and being qualitative rather than quantitatively and formally following from statistical theory of sampling and knowledge of sampling frames.

The scope I demonstrate in Chapter (4) revolves around shifting from public health monitoring, a task at the big-*N* level of analysis, to public health campaigns and outreach as a form of intervention, which can be effectively pursued at a small-*N* level. Specific populations may be in greater need to messaging around certain topics, and susceptible to certain messaging on those topics. I take up the case of Waterpipe Tobacco Smoking, in collaboration with domain experts from public health research, where the populations at risk of adoption are also populations involved in consuming and generating digital trace data on Twitter. Following a social media marketing model, I discuss the kind of system that would allow public health researchers to effectively make use of Twitter, one of engagement rather than a traditional, top-down, one-to-many marketing approach. I demonstrate the rigorous construction of the foundations of such a system, from annotation being done under the direction of the public health domain experts rather than by modelers, to cross-validation that estimates performance over time.

Lastly, I bring together multiple strands in my study in Chapter (5). I shift the scope of a sensor study from a large population of convenience (as in earlier sensor studies) or short-term assemblies (as in studies done with sensor badges) to a longer-term study of a smaller population within the well-specified boundary of a fraternity cohort, echoing previous social science study designs. I further shift the scope from approaches that seek to use sensors as an objective, superior replacement for subjective self-report to an approach that recognizes how sensors can *complement* self-report, by presenting an opportunity to study the *interaction* between the objective and directly measurable construct of proximity and the subjective and only indirectly measurable construct of friendship. Within this larger research objective, I get into some of the practical details of studying such a relationship, namely the vast differences in scale between sensor data and self-report, and how statistical machinery requires commensurate scales of data for modeling. I suggest that using frameworks of machine learning (feature selection, black box models, cross validation) on sensor data towards what Fisher (1922) identified as the core purpose of statistics—the "reduction of data"—is a principled way of reconciling the scale of digital trace data with that of self-report or other, similarly more sparsely measured outcomes of interest. Some of the exploratory findings I present, about characterizing proximity in terms of an inverse-squared functional transformation, and focusing on evening and weekend proximity, can serve as a baseline for future work and for creating a summary covariate of proximity for use in machinery like Stochastic Actor-Oriented Models that can model the co-evolution of node/edge covariates and network structure/processes.

Just like the Newcomb-Nordlie fraternity study or the Westgate study informed our theoretical understanding of processes of friendship formation without necessarily being conducted with a representative sample of the population, careful study design and proper scopes will similarly let us use digital trace data to advance our understandings of basic social processes, and ultimately to incorporate our understandings in responsible and just business decision-making and public policy.

State of the field

In reflecting on my work, there are a number of implications for the state of the field of digital trace data—which includes "computational social science" (or at least the definition of it articulated by Lazer, Pentland, et al., 2009, rather than the earlier notion of the term around social simulation, as Conte et al., 2012 seek to reclaim), as well as much of the data science and machine learning in the public sphere, such as in public policy, that makes use of such digital trace data.

Ethics

In both the introduction and in Chapter (3), I had deferred a full discussion of ethics, which I now complete. Taking the theme of the last chapter, using sensor data alone to detect friendship, we see that it is possible to improve on a random baseline although performance remains far from perfect. How such friendship predictions might be used is another question altogether. On the positive side, the Social Evolution work suggested using friendships to carry out social interventions around diet and exercise (Madan, Moturu, et al., 2010a) or for preventing disease transmission (Madan, Cebrian, et al., 2010). Or, like in Chin, Xu, Wang, and Wang (2012) and Chin, Xu, Wang, Chang, et al. (2013), applications might try to use co-location to identify possible but not currently existing ties and make suggestions, or suggestions about information dissemination in workplaces (Lawrence et al., 2006). On the negative side, marketing may be able to use knowledge of friendships for psychological manipulation, or for creating dissemination strategies, and such network information could be abused by governments; such uses may have negative consequences for individuals regardless of how good the predictions are.

This returns to the idea of two regimes of danger I raised in the introduction: there is the danger that, if digital trace data is effective in detecting aspects of our lives (like our friendships) that can be used for manipulation and control, and such detection power is concentrated in the hands of people with access to data and the technical skill to make use of it, then it exacerbates inequality. Conversely, if such detections are not high-quality but the lack of quality is not well appreciated, then regardless the data may still enable control and manipulation through feedback loops. Again, it is never possible to prove that inferences are not possible; as I wrote about Twitter data and elections, there is always the possibility that there is *some* feature extraction and *some* sophisticated and complex model that will extract a reliable signal. But based on my investigations, from smartphone location data alone we can only improve friendship detection over a random baseline by 20-30% (and, in a commercial or government surveillance use case where we may not have the controlled setting of cohort membership, this will be probably be even lower), suggesting we are in the second regime. If national security officials become fixated on using flimsy, circumstantial evidence to form network connections (Harris, 2013), and if such circumstantial evidence is used to make decisions about who to kill (Grothoff and Porup, 2016), it would be a grave human rights violation. Critics charge that this is already occurring, with secrecy preventing public knowledge of the extent to which it is true (Robbins, 2016, note that leaks that area rare source of information may be misleading, as they lack the context about how mature certain internal claims are and how sincerely we should take boasts); but theoretical arguments and empirical demonstrations of the the limited ability of digital trace data to in itself recover certain constructs of interest will hopefully help both public and internal arguments against such uses.

Surveillance, power, and control

The work out of MIT has been accompanied by extensive discussions of privacy, ethics, and institutional controls, such as in the "New Deal on Data" (Pentland, 2009; Greenwood et al., 2014; Pentland, 2014) that seeks to set rights of users to own and manage data about themselves, and discuss institutional rather than technological controls. But such controls do not yet exist, nor have the principles been widely adopted; as pointed out by Watts (2014), "we are being manipulated without our knowledge or consent all the time—by advertisers, marketers, [and] politicians" using modeling with digital trace data. Again as in the introduction, I cite Watts' argument that preventing academic research into such inferences and manipulations may only end up ensuring that the public will have no access to knowledge about what private entities are able to know about us from our data.

Mobile phone sensing platforms are, in some sense, currently the ultimate form of surveillance, even beyond social media data; the constant location monitoring alone potentially gives an excellent rough picture of our patterns of life, and mobile phones even have the capability of logging all keystrokes, and recording all audio. Some uses of mobile phone data are straightforward, such as using where phones are at night as a way of finding out where people live. One anonymous tech worker and abuse surviver bemoaned their colleagues in the tech industry "building technologies that make life easier for abusers" (Anonymous Author, 2015), and Freed et al. (2018) conducted interviews with survivors of intimate partner violence, one of whom described Facebook as "a stalker's dream". Certain simple inferences like these fall into the first regime of danger: those few who have access to data and models can exert control in inequitable ways. As successful models become incorporated into publicly (or commercially) available interfaces (e.g., in consumer-facing products and records, we seldom see raw GPS data, and instead only interact with post-processed data; and we can easily imagine APIs in the near future containing prepackaged "home location inference" methods), the fact that raw data is inaccessible does not keep inaccessible the inferences that such data enable.

There is also a question about populations with which sensing studies are run. Currently, sensing studies (including my own) have been run with relatively affluent populations, specifically those in elite higher education institutions and/or people employed in technology companies or in research. On the one hand, this means that less affluent populations are being left out of the research, which violates the justice principle of the Belmont Report. Not recognizing the full range of variation in behaviors and lifestyles may contribute to seeing a potentially narrow set of practices as standard. For example, even having a single, stable "home location" to which a user reliably returns daily is a pattern that may not hold for those experiencing homelessness and/or housing instability. On the other hand, affluent populations are the least likely to be at risk for the negative effects of surveillance, as their behaviors and lifestyles are considered normative (and indeed they may be the source of establishing and implicitly or explicitly enforcing norms of behavior around data sharing). Less affluent and marginalized populations who are disproportionately targeted by injustices such as violence from law enforcement, predatory lending, asset seizure, and a lack of equitable resource distribution have a long history of being surveilled, with that surveillance used to further control and repress them (Harper et al., 2014). Like with ankle bracelets, technologies developed for the purpose of behavioral intervention can be easily changed to officially sanctioned methods of control. As Robert S. Gable, a psychologist who was one of the creators of GPS anklet monitors used to monitor convicted offenders, writes that electronic monitoring is "a form of punishment itself" (Gable, 2017); he also advocates for using smartphones for electronic monitoring because they can be used to "reward rather than just to punish”, although there is the question of whether this would actually happen, or even worse, if (as he again points out has happened for existing technologies) the costs of surveillance are perversely passed on to the offenders themselves. Considering this, extending sensing research to worldwide and to less affluent and vulnerable populations is extremely sensitive, and care must be taken to, say, not use the results of sensing to further criminalize or pathologize youth (Drucker, 2017), to contribute to the dispossession of already-marginalized communities (Buchanan, 2008; Hayes, 2017), or to inadvertently empower repressive regimes by giving access to surveillance technologies that allow such regimes to track and suppress political dissidents and social movements.

Decontextualization

I had mentioned prioritizing context as a safeguard against being led astray by the ways in which technocultural structures or socioeconomic structures introduce selection bias, cause changes in behavior, or contain heterogeneity. The converse, of course, is how we can attribute many threats to validity to decontextualization in the research process. To make this precise, I connect decontextualization to Jasanoff and Kim (2015), who present the idea of *sociotechnical imaginaries*: we built technologies towards what we imagine, and use existing technologies to further imagine, in a process of co-production. What, then, are digital trace data technologies being built towards, and what sorts of imaginations do they inspire? A particularly striking image from Aharony et al. (2011), reproduced in figure (5.10) gives a striking example of an imaginary around sensors. While it is unclear how literally or sincerely the authors intended this image of the all-seeking eye (or how different audiences have taken it), and whether it means to imply that sensor data itself asymptotically approaches the eye, that sensor data is just another step along the way to this eye, or that the eye may appear in reach but will perpetually remain situated outside the axes of reachability, it does visually position a singular and complete ‘truth’ as the goal of research, a universal, neutral, objective, and omnipotent perspective towards which we strive.

As I have argued, for certain outcomes of theoretical interest (including social influence), the causal construct is actually that of psychological perceptions, and neither data with more throughput nor covering a longer duration is itself able to approach what we care about. At best, we can seek better and better correlations in sensor data of causal, subjective states, but we still need measurement of such states to use for calibrating models using sensor data. But so long as there is a market for pursuing and selling the vision of sensors and other digital trace data as contextless, universal, and therefore superior replacements for subjective measurements, there is a need for correctives.

Google Flu Trends might have been a ‘Dewey Defeats Truman’ moment for detection or prediction with digital trace data (Lazer, 2014), but there have not yet been explicit failures around scientific conclusions about social systems based on digital trace data. While the ways in which this might play out in sensors studies is hard to anticipate, the work that I have identified as relevant precedents can provide some possibilities. Cherry (1995), in a reinterpretation of Festinger et al. (1950), questions “whose reality the study reflects”: while the relative isolation of the study population was taken as a benefit by the researchers, Cherry points out that the lessened mobility and extreme isolation had a particular effect on the women, the wives of the men pursuing graduate degrees at MIT, who needed to rely on those nearby especially around childcare. It is possible that the need to keep watch over small children while they were playing in ad hoc spaces may have

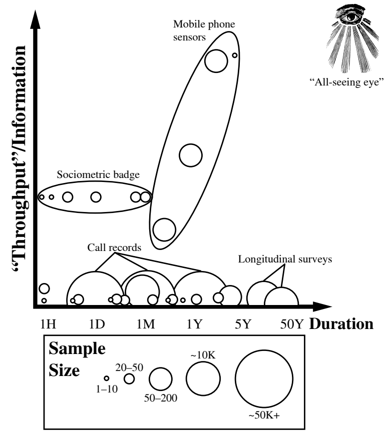


FIGURE 5.10: Reproduction of Aharony et al.'s (2011) schematic representation of their argument about how various types of sensor data compare to previously collected types of data.

driven much of the interaction, and Cherry argues that this would would do much to explain some findings that perplexed the original authors. She goes on to write about the 1950 book,

> "...the text becomes more abstract and mathematized and further removed from the experience of residents. The authors tested out ideas about sub-group formation and cohesiveness. They offered generalized statements such as 'The more cohesive the group, the more effectively it can influence its members' (p. 100). In statements like this and those that run throughout the remainder of the book, women's experiences in forming friendships and their participation in a tenants' organization are further decontextualized. Despite the use of women's experiences and choices as the data base, the language of 'the group' is used to discuss pressures to uniformity and consequences for deviation, and how rumor is communicated through friendship networks. The authors have already accepted that 'women' will represent the couple, and having done so, they introduce increasingly abstract constructs... In the final chapter, 'A theory of group structure', the language reverts to the 'generic' male. Statements are made that are not consistent with my own knowledge of how women, brought together through their common responsibility for small children, actually experience their lives."

The researchers' notions of the relative importance of men to women led them to not consider women as subjects in their own right, such that they failed to recognize how the labor involved in childcare drove many of the aspects of the social system they studied. While we would hope this particular structural bias is not present in digital trace data, such data do not capture psychological states and thus miss critical explanatory aspects of social systems—including structural inequities that have deeper social embeddings than what can be observed in the given social system. For sensors, studying modern student populations as self-contained social systems neglects the labor that supports much of their basic needs. The labor is done by people who few students interact with socially, people who likely would not in included from sensor studies. Resulting findings might be meaningful from the perspective of students, but would reproduce the enforced invisibility of working-class labor in modern social systems.

I hope that this thesis gives reasons to value the contexts in which digital trace data are produced, and provides counterpoints to the possibility of decontextualized, universal knowledge: the importance of context cannot be imagined away, and so pursuing a vision of such universality risks producing error and inequality, regardless of if or when the producers of systems and knowledge become aware of it.

Scientific progress

There is the moral argument about inequality as a reason for caring about threats to validity. There is also the argument in terms of scientific rigor. In Chapter (2), I brought up the metaphor of the microscope that is frequently used as an analogy of how digital trace data is poised to be an instrument that revolutionizes social science (Lazer, Pentland, et al., 2009; King, 2011; Golder and Macy, 2012; Golder and Macy, 2014; Watts, 2014; Mejova et al., 2015) to note that the historical analogy is more revealing than its advocates perhaps appreciate. For example, Golder and Macy (2012) write [emphasis added],

> “Disciplines are revolutionized by the development of novel tools: the telescope for astronomers, *the microscope for biologists*, the particle accelerator for physicists, and brain imaging for cognitive psychologists. *Social media provide a high-powered lens into the details of human behavior and social interaction* that may prove to be equally transformative.”

But there was more than a century between when Robert Hooke first described cells in the 1660s, and the emergency of *cell theory* in the 1830s. Why the ‘gap’? The historical record shows that instruments on their own do not cause new theory; instrumentation changes and is refined as new ideas develop about the phenomenon that the instrument is used to observe. It was not only improvements in the raw power of the microscope in terms of imrpoved lenses that enabled finer observation, but understanding the importance of proper illumination, and perhaps more importantly, the invention of *staining*—a technique of making the phenomenon amenable to observation by the instrument (Szekely, 2011).

We can use the historical lessons to create a competing sociotechnical imaginary to the all-seeing eye: digital trace data on its own (and in its original form) will likely not lead to new, more refined theories of social systems or behavior as is our scientific goal. Only as we live with and use the tool will we learn how to improve it and how to manipulate phenomena of interest to make the tool effective in the ways we desire. For digital trace data, I see understanding the biases—including the frequently commercial contexts of its production—as an analogy of better illumination, and I see designing interventions or using cohort studies as analogous to staining. It is impossible to know, of course, what will lead to new scientific knowledge, but we can take lessons from the past to understand the kinds of challenges we face, and the kind of responses we should expect needing to take to address those challenges.

Concluding thoughts

We are only at the beginning of finding out what we can do with digital trace data, and how we can go astray. To insure valid scientific findings, and to provide a solid base for fair, just and equitable public policy, large-scale trace data need to be considered with far more skepticism. Findings in this thesis validate certain grounds for skepticism, such as around platform design and engineering, or of the constructs captured in sensor data. But I also show ways to go beyond biases, to carve out research directions that do not try to make the data into an ideal measurement but instead respect the processes that generate the data.

Taken together, I believe this thesis has demonstrated the ways in which large-scale behavioral trace data can be biased and misleading, but how rigorous study design and the right set of theoretical considerations can overcome these barriers, successfully harnessing the scientific and policy potential of digital trace data.

Acknowledgements

I see this thesis as one step along a larger journey to understand the nature of modeling, its possibilities and limitations, and how to best use models to approach social phenomena in the age of computers. I have many people to thank along this journey.

First, thanks to my parents, Tariq and Saberah, and my brother Mohsin, for being my grounding throughout life, and the ground from which I have grown. I can scarcely appreciate how everything in my life has followed from the love, support, and foundation they gave me. My sister-in-law Asma has joined this loving foundation, and I can't wait to see what life has in store for my nephew Zia and his brand-new little brother Rami.

Second, I thank my co-advisors and committee. The resources and intellectual, logistic, vocational (programming and design), professional, and even emotional support Jürgen Pfeffer gave me especially early in my PhD career have benefitted me immensely, and I am proud to be a part of and to contribute to a larger research program around biases and 'knowing our data'. When I hit less-than-ideal circumstances in my PhD, Anind Dey was a savior to whom I am eternally grateful, and he has also been a fantastic mentor and supporter. He is, for me, a model of how to be an academic: working on fascinating and important problems, having a social consciousness that extends across his work, professional position, and personal life, and running a lab with compassion and that constantly brims with excitement about the research process. Cosma Shalizi's unapologetically critical work, and his identification and defense of rigor in statistical methodology, has made him one of my academic heroes and I am honored to have him on my committee. I am thankful to have had the chance to interact with him over the past four years; from larger points to offhanded remarks, to the advice and guidance he has given me in response to my (many) questions, to the fantastic reading recommendations he has given me across the breadth of academic disciplines, I have joined the many students for whom Cosma's insights, understandings, and insistence on deep understandings of fundamental statistical concepts have been a source of wisdom and training (although any misunderstandings and misuses of statistics remain my own). Lastly, I thank my external committee member, David Lazer, who has become a model of what computational social science should be in both carving out its potentials and doing excellent critical work that reveals the many challenges. This is also what I strive to do, and I am honored that he has accepted my request to review my work and contribute his expertise, in addition to helping guide me through this brave new field of computational social science.

Next, thanks to Urs Gasser and Sandra Cortesi: their limitless support and the support of the community at the Berkman Klein Center for Internet & Society is the reason I was even able to start down the path of my current career. Also from Berkman, Eszter Hargittai has continued to provide advice about what it is to be rigorous in social science. I am grateful to the many researchers and staff for their mentorship and care at Berkman, including Becca Tabasky, Jon Murley, Karyn Glemaud, Carey Andersen, Colin Maclay, Amar

Ashar, Sebastian Diaz, David Larochelle, and Edward Popko, among others, as well as my colleagues from the Youth and Media team, and especially to Ashley Lee, Meredith Beaton, and Nathaniel Levy, the 2011 and 2012 Youth and Media interns, and John Palfrey. From my undergraduate time at Harvard, thanks to Ann Harrington for first inspiring me to pursue the history of science, and to Mario Biagioli, who showed me where to apply my history of science training outside of the discipline, and who also remains a model to me of what it means for academic work to be an expression of living life. Thanks also to Dick Wolf and Elaheh Kheirandish for being mentors and supporters across my disciplinary journeys, to Parimal Patil for great advice, Dick Gross for looking out for me, and Catherine Shapiro for tireless support. In Rhode Island, thanks to my teacher Mimi Morimura, to Gwenn Stearn and everybody at the RI State Archives, to Janet Freedman, Caitlin Chaffee, and the staff at the RI Coastal Resource Management Council for my first white paper experience, to Hasan Askari and Tim Dahler for some fantastic exposure to and training in design during an internship at RISD Prints by Concept Link, and to my musical partner Bob Fish for much joy.

From my time at the Oxford Internet Institute, thanks to Bernie Hogan and Monica Bulger, who gave me fantastic guidance and advice, both intellectual and professional, and who have continuing doing so in the years since. Bernie's notion of 'digital social research' continues to inform and inspire me, and I highly value his appraisals of intellectual quality. I am grateful to Monica for, among other things, starting me on a path to appreciating the depth and breadth of educational theory and for overall career advice. Thanks to Sandra Gonzalez-Bailon, who supported and encouraged me at OII and has continued to do so since as she becomes a leader in computational social science. Thanks to Grant Blank for solid training in some intuitions of social statistics that have served me well throughout learning econometrics and mathematical statistics, and Ralph Schroeder for challenges that helped me sharpen my own beliefs and arguments, as well as Vicky Nash for the community she contributed to creating, and Eric Meyer for overseeing my thesis.

Thanks to staff of the IT Learning Programme at the University of Oxford Computing Services, and particularly to David Baker, Kathryn Wenczek, and Pamela Stanworth. My internship with them after my master's degree gave me a fantastic opportunity to practice and apply my new skills in network analysis and visualization around a meaningful project, understanding how students were using the ITLP in order to better serve the university community.

A special thanks also to Viktor Mayer-Schoenberger and Kenneth Cukier; working as their research assistant for *Big Data* gave me both an amazing lesson in the worthwhile challenges and compromises involved in public engagement, and provided grounds for me to start to systematically think through implications of big data, machine learning, and causality vs. correlation. Seeing them work and participating in it continues to shape my thinking, even as I gain technical knowledge and an appreciation of the subtleties of modeling.

I also had many fantastic colleagues at OII who I have kept in contact with and who continue to stimulate and encourage me. Cohen R. Simpson, in addition to being a great friend, has been working with ERGMs, SAOMs, and other statistical models as a practitioner long before I got to them, and his tutoring was a great boon in developing my understanding of how these models actually work and how to run them. Devin Gaffney, who I met even earlier at Berkman, has been an excellent colleague; I greatly value our shared interest in interrogating the nature of modeling, and transferring that skepticism into thoughts about what questions to research and how better to do research. Manuel Schäffer has likewise been a fantastic friend and intellectual companion, not only in matters of technology and data but politics and humanism. Amy McLeod and André Avorio are both gentle souls whose experiences in the tech sector I greatly appreciate seeing, and whose respective dreams of helping people's learning and collaboration I hope to see them reach. Thanks to Kath Tolentino for many discussions, and for her perspective now from the arts. I continue to have fantastic conversations with Scott Hale, and look forward to continuing them across conferences and meetings. Thanks also to Joshua Melville, who is both a brilliant programmer and a brilliant and true social scientist. He is responsible for having challenged much of my thinking when I inadvertently slid towards positivism and other such unsavory worldviews, and for pointing me towards sources that have forced me to think about the *phenomenology of ties* in network representations as a fundamental concern that precedes and supersedes statistical modeling and mathematical manipulation.

I am extremely grateful for the Sunbelt community, which for me is a model of unpretentiously and naturally interdisciplinary work. Many thanks to Tom A. B. Snijders for his mentorship to the community and many contributions to the field; while my interactions with him over the years have been brief, every time I see him seem to get a nugget of wisdom that vastly furthers my understanding of statistical modeling. Thanks also to Brooke Foucault Welles, who I first met when she gave me advice on graduate school, and who I have continued to look to for having an appreciation of the ethical issues involved with online research and with whom to talk through the challenges therein. Noshir Contractor has been a mentor to me even without any official relationship, and I thank him for his generosity and for being a living embodying of the very best of networking. I appreciate the help I have gotten from other colleagues I have met at Sunbelt, including jimi adams, Timon Elmer, Mathieu Génois, and James A. Litts, the latter three of whom helped provide me with some advance notice of results around sensors that helped me make my Chapter (3) more accurate and timely. I am also grateful for the communities of ICWSM and IC2S2, which overlap with both each other and with Sunbelt, as places where I see statistical modeling and social theory genuinely coming together in meaningful and fascinating ways. Desi Hristova has been a fixture of my ICWSM experiences, and I look forward to her future work. Derek Ruths has been endlessly patient and kind, and inspires me as an example of a trained computer scientists who has arrived at a nuanced, enlightened view of the nature of data and modeling, using that view to do amazing work. A special thanks to Abby Jacobs for bringing to my attention several important articles, and in general for being a like-minded collaborator in thinking critically about social media data and a supporter of my efforts. I also appreciate her work, particularly as her paper around the "Facebook 100" is a perfect example of how to make meaningful statements about and with social media data. And, while a meeting at Cornell rather than at a conference, I thank Phoebe Sengers for helping me discover 'critical technical practice'.

Thanks to the ARCS Foundation, whose award made my graduate career much more comfortable, and particularly to Carol and Paul Stockman for both their named award and for wonderful company over the three years of the award.

In the Institute for Software Research, many amazing administrators and staff have made my experience at CMU smooth despite some exogenous shocks. Thanks to Connie Herold for putting up with so much, from me and others, in running two PhD programs, being an advocate for students, and managing every logistic nightmare that we bring; Sharon Blazevich, who has looked out for me and is a constant, comforting fixture in ISR; Nick Frollini, who has done a fantastic job advocating for students and the department to the larger university in addition to taking care of us within the department; and Helen Higgins, Monika DeReno, and Jaime Lou Hagerty for taking good financial care of me and respectively my paychecks, funding, and reimbursements. Thanks to Tom Pope for tech help and for 3d printing help, to Ryan Johnson and (from earlier in my degree) Chris Dalansky for help with all my technology orders, and to Joshua Quicksall for a fantastic headshot, design contributions to the department, and the design advice he has given me. And thanks to Jim Herbsleb, the director of Societal Computing, for being somebody I can rely on and go to for problems I encounter, for solid methodological training through his course, and for the opportunity to be a TA for his fantastic class, Ethics and Policy Issues in Computing.

I have only been part of CMU's Ubicomp lab for a short period of time, but much of what I have been able to do there is because of the help of lab members. Afsaneh Doryab in particular has been an amazing leader and colleague who I frequently rely on for making decisions, directing research, and being the last line of defense to prevent the failure of study infrastructures. Many thanks to Denzil Ferreira and Yuuki Nishiyama's for their continued work on developing the AWARE framework, and despite being new faculty, serving as de facto developers and tech support for all the bugs we encountered when using AWARE. It has also been wonderful to be around Sang Won Bae to discuss running studies, Nikola Banovic to discuss conceptual problems in statistical modeling, and Julian Andres Ramos Rojas for starting to work out better ways to approach and teach machine learning, in addition to the many other wonderful interns, research assistants, graduate students, postdocs, visitors and research scientists in the Ubicomp Lab and HCII.

CMU has also provided me with many fantastic faculty and teachers, and members of the community. Thanks to Carol Frieze for all her amazing work behind Bias Busters, I have been honored to support and be a part of the work. Thanks also to Diana Marculescu, and at Google Pittsburgh, Gerry Katilius, for letting me tag along in Bias Busters presentations on national stages, and thanks to all the facilitators with whom I did Bias Busters sessions. My slight involvement in service in the CMU community has helped me meet many amazing people: I have been honored to be around the amazing work of Vaasavi Unnava, Daniel Gingrich, Carolyn Commer, Beth Halayko, Jess Kaminsky, Gina Casalegno, Renee Camerlengo, Holly B. Hippensteel, Jamie Edwards, Jess Klein, Kristin Hughes, Amy Burkert, Suzie Laurich-McIntyre, Jaime Rossi, Geoff Kaufmann, Jonathan Reynolds, Lucia Gonzalez-Prier, and Ty Walton.

Thanks to Radu Marculescu, whose amazing enthusiasm, curiosity, and intellectual intrepidness has been a pleasure to interact with and learn from. Ann Lee and Valerie Ventura were fantastic statistics teachers, and their respective courses (and the time they took to answer my many questions) greatly helped me along. Alex Smola's machine learning class was fantastic, as was Larry Wasserman's legendary Intermediate Statistics. Ann Lee's probability and mathematical statistics class gave me the background to take all subsequent courses, and Valerie Ventura's class on regression analysis was the place where things finally clicked and I came to understand what it means to do statistical analysis. Thanks to Brian Junker, whose foray into networks gave me another fantastic fountain of statistical understanding, wisdom, and experience, and thanks for his patience in guiding me along through my statistical naïveté in his networks course and beyond. Thanks to Teddy Seidenfeld for some amazing thoughts, and opening up to me views about how diverse and strange statistics can be. Thanks to Dave Choi for agreeing to give me technical guidance, and for excellent advice pushing me to formalize my thinking. Through sessions with him, I grasped the counterfactual framework and other fundamentals of statistical modeling. Kate Anderson has also been a fantastic informant for social networks within economics, and I am grateful for her encouragement and providing an audience for my work on explaining statistical models for social networks.

A very special thanks to David Krackhardt, who is an absolute treasure. Both in the two courses of his I took and in my general interactions, he has imparted amazing wisdom from a deep understanding of what it is to express insights and understandings of the social world in terms of network models, and how to manipulate network models as a guide to thinking about the social world (and not a replacement for it). His amazing research stories have also filled a critical humanistic role in showing me process of exploration that true research is.

Thanks also to Nynke M. D. Niezink, who graciously agreed to serve on the committee for my Secondary Master's in Machine Learning; I regret that I am leaving CMU just as she has joined, such that I won't benefit much from her extensive expertise with SIENA and to take up her advances, but I am very grateful to have had the brief benefit of her advice.

I am also grateful to the many amazing colleagues I have at CMU. Foremost is my labmate and frequent co-author Hemank Lamba, with whom I gave grown as a researcher and thinker and have enjoyed a fantastic synergy. Thanks to Maria De Arteaga and Ben Boecking for always reaching out, and to them and to Connor Walsh, Maria Cueller, Petar Stojanov, Wenting Yu, Shayak Sen, Jakob Bauer, Naji Shajarisales, Antonis Manousis, Yasmine Kotturi, and my other colleagues with whom I have worked on problem sets and class projects, and discussed statistics and machine learning and methodology. Thanks to the fantastic TAs I have had, and particularly to Manzil Zaheer. Thanks to Jacqui Stimson at the Eberly Center for Teaching Excellence & Educational Innovation for helping me construct a teaching statement and course syllabus, and to the staff and instructors of the Eberly center for providing excellent resources that have helped me become a better educator. Thanks to Seth Flaxman, my undergraduate classmate who, in the brief time we overlapped at CMU, became more of a mentor than colleague around RKHS and using machine learning for social science. Thanks to the people with whom I worked on networks: Philipp Burkhardt and Alex Loewi, with whom Hemank and I had an informal networks meetup that helped along much of my thinking and understanding. Alessandro Iorio has also been a fantastic colleague in networks, as has his advisor Brandy Aven and her partner Jeremy Koster. Thanks to my longest-term officemate, Geoffrey P. Morgan, for entertaining my barrage of skeptical questions about simulation modeling; his answers, along with some fascinating virtual and then finally in-person interactions with Edmund Chattoe-Brown, to whom I am also grateful, helped me appreciate what simulation modeling can offer and sharpened my arguments against its use. Thanks to my many wonderful colleagues in ISR, and particularly to Kenny Joseph and Peter Landwehr for early technical help, to Hanan Hibshi for her work to create a community, and to Ghita Mezzour for her incredible generosity. Thanks to David Widder, my most recent office-mate, for fantastic discussions. Thanks to Diyi Yang for organizing a Social Computing reading group, to Maggie Oats and Reuben Aronson for organizing an STS reading group, and Lee Richardson for organizing a History of Statistics reading group. All of these have exposed me to fantastic work, and led to fantastic conversations. Many thanks to the stat-networks group in the Statistics Department (Neil Spencer, Nicolas Kim, Alden Green, and alumni Beau Dobbs, Samrachana Adhikari, and Lawrence Wang, under Cosma Shalizi and Brian Junker): being able to interact with statistics research into networks, and even present my own work on the applied and pedagogical side, has been enormously beneficial in my intellectual development and for this I am forever grateful. Thanks also to Lizzie Silver and João Martins for being leaders and models for serving the Pittsburgh community in which we live, and for establishing a path in which I strive to follow.

Thanks also to my various co-authors. Thanks to Jason B. Colditz, Kar-Hai Chu, and my collaborators in public health at Pitt for spearheading fantastic work and allowing me to contribute and shape the project. Thanks to Mike Merrill for help in making possible my sensors work, and for being a fantastic researcher remarkably early in his career. Thanks also to Yuvraj Agarwal, who helped me and Jürgen when we first started our investigations into sensor data for social networks in 2013. Thanks to Constantine Nakos, whose deft coding skills made possible my first first-authored paper, and whose generosity during our Washington trip made it both comfortable and memorable. Thanks also to Gabriel Coutinho Ferreira, and to Claus Hunsen, Sofie Kemper, Thomas Bock and Sven Apel for giving me opportunities to apply my new knowledge of statistics to new and fascinating questions around software development networks. Thanks to the members of the Health & Human Performance Lab in the psychology department for amazing ongoing work around sensor data. For Erowid.org, thanks to Sylvia Thyssen and Matthew Baggott for replying to my queries, being open to collaboration, and guiding me and giving feedback on my proposed analysis plan, even if the work around this site will wait until after this thesis.

I had a fantastic summer in Lisbon in 2017 as part of the first class of Data Science for Social Good Europe fellows. It was everything I could have hoped for and more: I was able to see how service modeling plays out in practice, addressing real needs in social services and local government. Thanks first and foremost to my fantastic project team, Io Flament and Cristina Lozano, with whom I both had a fantastic time and learned so much, about software engineering and tenacity and life. I will forever treasure memories of sitting in magical places in Lisbon, sipping a galão or some other wondrous drink and looking out on the setting sun over the ocean. Thanks to our project manager Laura Szczuczak, who took time off from deeply worthwhile work to do I hope equally worthwhile work. João Fernandes was a fantastic manager and guide to Portugal's geography and bureaucracy. Qiwei Han was a fantastic technical mentor and de facto sysadmin, helping us with all our big data, Postgres, and AWS issues. To Euro Beinat and his team, including Dario Bertocchi, Bartosz Hawelka, and Pavlos Kazakopoulos, I am grateful for both technical help, for being the connection through which we were even able to get the data for the project. And to Euro, I am also grateful for getting a fantastic tour of the Doumo, and to Florence overall. Our collaborators in Florence, as well, were wonderful: Alberto Peruzzini and Laura Morelli at Toscana Promozione Turistica for being fantastic partners as well as fantastic hosts for our team's site visit, and to all the others who met with us during out visit (including struggling through language barriers). I have made lifelong friends among the rest of the DSSG cohort, from mentors Jane Zanzig, Gizem Bacaksizlar, Paul van der Boor and Anna Delgado, and Lauren Haynes, and the other fellows Iván Higuera-Mendieta, Helga Holmestad, Sefi Mintzer, Victor Vicente, Kathy Yang, Will Grimes, Tarciso Oliveira, Shubham Tomar, Hui Xiang Chua, Manas Gaur, Iñigo Martinez de Rituerto de Troya, and Mengxin Ji.

Thanks especially to Rayid Ghani, who is not only responsible for the program, but who has given me fantastic advice and guidance about how to approach machine learning, how to actually do real machine learning with real data (something I have found nowhere else, and still hope to see teaching materials about someday), and how to think about the kind of knowledge infrastructures needed to proceed with using modeling in positive ways and towards egalitarian ends. And, eternal thanks to our organizer Leid Zejnilovic, who is an amazing human being on so many levels, whose mode and model of life contains remarkable wisdom, and who always took fantastic care of us particularly around organizing getting us to and from amazing beaches.

Lastly, thanks to all my old friends from college and before: Kris Bartkus, Daniel Fried, Jonathan M. Estey, Tatiana H. Chaterji, Sejal Babaria, Justin Grosslight, Vivek Rudrapatna, Yiyi Deng, Sashka Kjuchukova, and Vasanthi Viswanathan. Thanks to Stephen C. Walker for being my reliable informant about how everything we think we know about intellectual history is wrong (and how things are so much more amazing than we know), and for being a musical companion and fellow conspirator. Thanks to Lewis Z. Liu, an amazing friend throughout the years who has supported me in ways both overt and imperceptible. I am immensely grateful that he and Andrea Snavely are in my life. From after college, thanks to Nishant Shah for giving me a view into the depths of the humanities, and into the process of shaping new worlds, and for vicarious views into the life of a humanities scholar. A very special thanks to Sophie-Jung Kim for being such a close friend and confidant as we forge our ways through our respective areas in academia, and I look forward to many more years of friendship and intellectual growth. And eternal thanks to Maya Randolph not only for keeping me intellectually and emotionally honest, but for showing me that I haven't even begun to grasp what it is to understand theory and the social world.

Bibliography

Abbott, Andrew (1988). "Transcending general linear reality". In: *Sociological Theory* 6 (2), pp. 169–186. DOI: 10.2307/202114.

– (2004). *Methods of discovery: Heuristics for the social sciences*. Contemporary Societies. New York, NY: W. W. Norton & Company.

adams, jimi (2010). "Distant friends, close strangers? Inferring friendships from behavior". In: *Proceedings of the National Academy of Sciences* 107 (9), E29–E30. DOI: 10.1073/pnas.0911195107.

Agre, Philip E. (1997). "Towards a critical technical practice: Lessons learned from trying to reform AI". In: *Social science, technical systems, and cooperative work: Beyond the great divide*. Ed. by Geoffrey C. Bowker, Susan Leigh Star, Will Turner, and Les Gasser. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 131–158.

Aharony, Nadav, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland (2011). "Social fMRI: Investigating and shaping social mechanisms in the real world". In: *Pervasive and Mobile Computing* 7 (6), pp. 643–659. DOI: 10.1016/j.pmcj.2011.09.004.

Ahmetovic, Dragan, Cole Gleason, Chengxiong Ruan, Kris Kitani, Hironobu Takagi, and Chieko Asakawa (2016). "NavCog: A navigational cognitive assistant for the blind". In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '16, pp. 90–99. DOI: 10.1145/2935334.2935361.

Allem, Jon-Patrick, Kar-Hai Chu, Tess Boley Cruz, and Jennifer B. Unger (2017). "Waterpipe promotion and use on Instagram: #Hookah". In: *Nicotine & Tobacco Research* 19 (10), pp. 1248–1252. DOI: 10.1093/ntr/ntw329.

Allem, Jon-Patrick, Patricia Escobedo, Kar-Hai Chu, Daniel W. Soto, Tess Boley Cruz, and Jennifer B. Unger (2017). "Campaigns and counter campaigns: Reactions on Twitter to e-cigarette education". In: *Tobacco Control* 26 (2), pp. 226–229. DOI: 10.1136/tobaccocontrol-2015-052757.

Angelopoulos, Constantinos Marios, Christofoulos Mouskos, and Sotiris Nikoletseas (2011). "Social signal processing: Detecting human interactions using wireless sensor networks". In: *Proceedings of the 9th ACM International Symposium on Mobility Management and Wireless Access*. MobiWac '11, pp. 171–174. DOI: 10.1145/2069131.2069163.

Angrist, Joshua D. and Jörn-Steffen Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Anonymous Author (2015). "Without scars: Domestic violence, abuse and the tech pipeline". In: *Model View Culture* (27). URL: https://modelviewculture.com/pieces/without-scars-domestic-violence-abuse-and-the-tech-pipeline.

Anselin, Luc (2002). “Under the hood: Issues in the specification and interpretation of spatial regression models”. In: *Agricultural Economics* 27 (3), pp. 247–267. doi: 10.1111/j.1574-0862.2002.tb00120.x.

Anselin, Luc and Serge Rey (1991). “Properties of tests for spatial dependence in linear regression models”. In: *Geographical Analysis* 23 (2), pp. 112–131. doi: 10.1111/j.1538-4632.1991.tb00228.x.

Anselin, Luc, Sanjeev Sridharan, and Susan Gholston (2007). “Using exploratory spatial data analysis to leverage social indicator databases: The discovery of interesting patterns”. In: *Social Indicators Research* 82 (2), pp. 287–309. doi: 10.1007/s11205-006-9034-x.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan (2009). “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks”. In: *Proceedings of the National Academy of Sciences* 106 (51), pp. 21544–21549. doi: 10.1073/pnas.0908800106.

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green (2010). “A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark”. In: *Sociological Methods & Research* 39 (2), pp. 256–282. doi: 10.1177/0049124110378098.

Arlot, Sylvain and Alain Celisse (2010). “A survey of cross-validation procedures for model selection”. In: *Statistics Surveys* 4, pp. 40–79. doi: 10.1214/09-SS054.

Athey, Susan (2017). “Beyond prediction: Using big data for policy problems”. In: *Science* 355 (6324), pp. 483–485. doi: 10.1126/science.aal4321.

Babbie, Earl (2010). *The practice of social research*. 12th ed. Belmont, CA: Wadsworth, Cengage Learning.

Baeza-Yates, Ricardo (2018). “Bias on the Web”. In: *Communications of the ACM* 61 (6), pp. 54–61. doi: 10.1145/3209581.

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau (2013). “Summary report of the AAPOR task force on non-probability sampling”. In: *Journal of Survey Statistics and Methodology* 1 (2), pp. 90–143. doi: 10.1093/jssam/smt008.

Bangu, Sorin (2016). “On ‘The unreasonable effectiveness of mathematics in the natural sciences’”. In: *Models and inferences in science*. Ed. by Emiliano Ippoliti, Fabio Sterpetti, and Thomas Nickles, pp. 11–29. doi: 10.1007/978-3-319-28163-6_2.

Barrat, Alain and Ciro Cattuto (2013). “Temporal networks of face-to-face human interactions”. In: *Temporal networks*. Ed. by Petter Holme and Jari Saramäki. Berlin, Heidelberg: Springer, pp. 191–216. doi: 10.1007/978-3-642-36461-7_10.

Barrat, Alain, Ciro Cattuto, Vittoria Colizza, Francesco Gesualdo, Lorenzo Isella, Elisabetta Pandolfi, Jean-François Pinton, Lucilla Ravà, Caterina Rizzo, Mariateresa Romano, Juliette Stehlé, Alberto Eugenio Tozzi, and Wouter van den Broeck (2013). “Empirical temporal networks of face-to-face human interactions”. In: *The European Physical Journal Special Topics* 222 (6), pp. 1295–1309. doi: 10.1140/epjst/e2013-01927-7.

Barrat, Alain, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Caterina Rizzo, Alberto Eugenio Tozzi, and Wouter van den Broeck (2012). “Wearable sensor networks for measuring face-to-face contact patterns in healthcare settings”. In: *Revised Selected Papers from the Third International Conference on Electronic Healthcare*. eHealth 2010, pp. 192–195. doi: 10.1007/978-3-642-23635-8_24.

Barrat, Alain, Ciro Cattuto, Vittoria Colizza, Jean-François Pinton, Wouter Van den Broeck, and Alessandro Vespignani (2008). “High resolution dynamical mapping of social interactions with active RFID”. eprint: https://arxiv.org/abs/0811.4170.

Barrat, Alain, Ciro Cattuto, Alberto Eugenio Tozzi, Philippe Vanhems, and Nicolas Voirin (2014). "Measuring contact patterns with wearable sensors: Methods, data characteristics and applications to data-driven simulations of infectious diseases". In: *Clinical Microbiology and Infection* 20 (1), pp. 10–16. DOI: 10.1111/1469-0691.12472.

Basu, Sumit (2002). "Conversational scene analysis". PhD thesis. MIT Department of Electrical Engineering and Computer Science.

Benevenuto, Fabrício, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida (2012). "Characterizing user navigation and interactions in online social networks". In: *Information Sciences* 195, pp. 1–24. DOI: 10.1016/j.ins.2011.12.009.

Bergmeir, Christoph and José M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191, pp. 192–213. DOI: 10.1016/j.ins.2011.12.028.

Berk, Richard A. and David A. Freedman (2003). "Statistical assumptions as empirical commitments". In: *Law, punishment, and social control: Essays in honor of Sheldon Messinger*. Ed. by Thomas G. Blomberg and Stanley Cohen. Transaction Publishers, pp. 235–254.

Bernard, H. Russell and Peter D. Killworth (1977). "Informant accuracy in social network data II". In: *Human Communication Research* 4 (1), pp. 3–18. DOI: 10.1111/j.1468-2958.1977.tb00591.x.

Bernard, H. Russell, Peter D. Killworth, and Lee Sailer (1979). "Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data". In: *Social Networks* 2 (3), pp. 191–218. DOI: 10.1016/0378-8733(79)90014-5.

– (1982). "Informant accuracy in social-network data V: An experimental attempt to predict actual communication from recall data". In: *Social Science Research* 11 (1), pp. 30–66. DOI: 10.1016/0049-089X(82)90006-0.

Bernstein, Michael S., Eytan Bakshy, Moira Burke, and Brian Karrer (2013). "Quantifying the invisible audience in social networks". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13, pp. 21–30. DOI: 10.1145/2470654.2470658.

Bivand, Roger S., Jan Hauke, and Tomasz Kossowski (2013). "Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods". In: *Geographical Analysis* 45 (2), pp. 150–179. DOI: 10.1111/gean.12008.

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio (2013). *Applied spatial data analysis with R*. 2nd ed. Springer, NY. URL: http://www.asdar-book.org/.

Bivand, Roger S. and Gianfranco Piras (2015). "Comparing implementations of estimation methods for spatial econometrics". In: *Journal of Statistical Software* 63 (18), pp. 1–36. DOI: 10.18637/jss.v063.i18.

Blank, Grant (2016). "The digital divide among Twitter users and its implications for social research". In: *Social Science Computer Review* 35 (6), pp. 679–697. DOI: 10.1177/0894439316671698.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2 (1), pp. 1–8. DOI: https://doi.org/10.1016/j.jocs.2010.12.007.

Bopp, Chris, Ellie Harmon, and Amy Voida (2017). "Disempowered by data: Nonprofits, social enterprises, and the consequences of data-driven work". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17, pp. 3608–3619. DOI: 10.1145/3025453.3025694.

Borgatti, Stephen P. and Rob Cross (2003). "A relational view of information seeking and learning in social networks". In: *Management Science* 49 (4), pp. 432–445. DOI: 10.1287/mnsc.49.4.432.14428.

Borgatti, Stephen P., Martin G. Everett, and Jeffrey C. Johnson (2013). *Analyzing social networks*. London: SAGE.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca (2009). "Network analysis in the social sciences". In: *Science* 323 (5916), pp. 892–895. DOI: 10.1126/science.1165821.

Boudreau, Marie-Claude and Daniel Robey (2005). "Enacting integrated information technology: A human agency perspective". In: *Organization Science* 16 (1), pp. 3–18. DOI: 10.1287/orsc.1040.0103.

Boughorbel, Sabri, Fethi Jarray, and Mohammed El-Anbari (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PLOS ONE* 12 (6), pp. 1–17. DOI: 10.1371/journal.pone.0177678.

Box, George E. P. (1979). *Robustness in the strategy of scientific model building*. Tech. rep. #1954. University of Madison-Wisconsin, Mathematics Research Center.

boyd, danah m. (2008). "Why youth ♥ social network sites: The role of networked publics in teenage social life". In: *Youth, Identity, and Digital Media*. Ed. by David Buckingham. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, pp. 119–142. DOI: 10.1162/dmal.9780262524834.119.

boyd, danah m. and Kate Crawford (2012). "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon". In: *Information, Communication & Society* 15 (5), pp. 662–679. DOI: 10.1080/1369118X.2012.678878.

boyd, danah m., Scott Golder, and Gilad Lotan (2010). "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter". In: *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*. HICSS '10, pp. 1–10. DOI: 10.1109/HICSS.2010.412.

Breiman, Leo (2001). "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16 (3), pp. 199–231. DOI: 10.1214/ss/1009213726.

Brennan, Emily, Laura A. Gibson, Ani Kybert-Momjian, Jiaying Liu, and Robert C. Hornik (2017). "Promising themes for antismoking campaigns targeting youth and young adults". In: *Tobacco Regulatory Science* 3 (1), pp. 29–46. DOI: 10.18001/trs.3.1.4. URL: https://doi.org/10.18001/trs.3.1.4.

Brogueira, Gaspar, Fernando Batista, and Joao Paulo Carvalho (2016). "Using geolocated tweets for characterization of Twitter in Portugal and the Portuguese administrative regions". In: *Social Network Analysis and Mining* 6 (1). DOI: 10.1007/s13278-016-0347-8.

Buchanan, Alanna (2008). "A racial justice perspective on monitoring domestic violence offenders using GPS systems conversation: GPS monitoring of domestic violence offenders". In: *Harvard Civil Rights-Civil Liberties Law Review* 43, pp. 271–275.

Bucher, Taina (2012). "Want to be on the top? Algorithmic power and the threat of invisibility on Facebook". In: *New Media & Society* 14 (7), pp. 1164–1180. DOI: 10.1177/1461444812440159.

– (2017). "The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms". In: *Information, Communication & Society* 20 (1), pp. 30–44. DOI: 10.1080/1369118X.2016.1154086.

Buja, Andreas, Richard Berk, Lawrence Brown, Edward George, Emil Pitkin, Mikhail Traskin, Linda Zhao, and Kai Zhang (2016). "Models as approximations—A conspiracy of random regressors and model deviations against classical inference in regression". In: *Statistical Science*. URL: http://www-stat.wharton.upenn.edu/~buja/PAPERS/Buja_et_al_Conspiracy-v4.pdf.

Bump, Philip (2013). "The NSA admits it analyzes more people's data than previously revealed". In: *The Atlantic* (17 July 2013). URL: https://www.theatlantic.com/politics/archive/2013/07/nsa-admits-it-analyzes-more-peoples-data-previously-revealed/313220/.

Burke, Moira and Robert E. Kraut (2014). “Growing closer on Facebook: Changes in tie strength through social network site use”. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. CHI ’14, pp. 4187–4196. DOI: 10.1145/2556288.2557094.

Burke, Moira, Robert E. Kraut, and Cameron Marlow (2011). “Social capital on Facebook: Differentiating uses and users”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11, pp. 571–580. DOI: 10.1145/1978942.1979023.

Burke, Moira, Cameron Marlow, and Thomas Lento (2009). “Feed me: Motivating newcomer contribution in social network sites”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’09, pp. 945–954. DOI: 10.1145/1518701.1518847.

Butts, Carter T. (2008). “A relational event framework for social action”. In: *Sociological Methodology* 38 (1), pp. 155–200. DOI: 10.1111/j.1467-9531.2008.00203.x.

Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell (1983). “Machine learning: A historical and methodological analysis”. In: *AI Magazine* 4 (3). DOI: 10.1609/aimag.v4i3.406.

Carr, Nicholas (2014). “The limits of social engineering”. In: *MIT Technology Review* May/June 2014. URL: https://www.technologyreview.com/s/526561/the-limits-of-social-engineering/.

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina (2008). “An empirical evaluation of supervised learning in high dimensions”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. DOI: 10.1145/1390156.1390169.

Cattuto, Ciro, Marco Quaggiotto, André Panisson, and Alex Averbuch (2013). “Time-varying social networks in a graph database: A Neo4J use case”. In: *Proceedings of the First International Workshop on Graph Data Management Experiences and Systems*. GRADES ’13, 11:1–11:6. DOI: 10.1145/2484425.2484442.

Cattuto, Ciro, Wouter van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani (2010). “Dynamics of person-to-person interactions from distributed RFID sensor networks”. In: *PLOS ONE* 5 (7), e11596. DOI: 10.1371/journal.pone.0011596.

CDC (2016). “Hookahs”. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/tobacco_industry/hookahs/index.htm.

– (2018). “Tips from former smokers®”. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/tobacco/campaign/tips/index.html.

Centola, Damon (2010). “The spread of behavior in an online social network experiment”. In: *Science* 329 (1194), pp. 1194–1197. DOI: 10.1126/science.1185231.

Chan, Andrew and Susan Murin (2011). “Up in smoke”. In: *Chest* 139 (4), pp. 737–738. DOI: 10.1378/chest.10-2985.

Chang, Jonathan, Itamar Rosenn, Lars Backstrom, and Cameron Marlow (2010). “ePluribus: Ethnicity on social networks”. In: *Proceedings of the Fourth International Conference on Weblogs and Social Media*. ICWSM-10, pp. 18–25. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1534.

Chen, Kehui and Jing Lei (2018). “Network cross-validation for determining the number of communities in network data”. In: *Journal of the American Statistical Association* 113 (521), pp. 241–251. DOI: 10.1080/01621459.2016.1246365.

Chen, Yiming, Sheng Zhao, and Jay A. Farrell (2015). “Computationally efficient carrier integer ambiguity resolution in multiepoch GPS/INS: A common-position-shift approach”. In: *IEEE Transactions on Control Systems Technology* 24 (5), pp. 1541–1556. DOI: 10.1109/TCST.2015.2501352.

Cherry, Frances (1995). “One man’s social psychology is another woman’s social history”. In: *The stubborn particulars of social psychology: Essays on the research process*. London: Routledge, pp. 68–83.

Chin, Alvin, Bin Xu, Hao Wang, Lele Chang, Hao Wang, and Lijun Zhu (2013). “Connecting people through physical proximity and physical resources at a conference”. In: *ACM Transactions on Intelligent System Technologies* 4 (3), 50:1–50:21. DOI: 10.1145/2483669.2483683.

Chin, Alvin, Bin Xu, Hao Wang, and Xia Wang (2012). “Linking people through physical proximity in a conference”. In: *Proceedings of the 3rd International Workshop on Modeling Social Media*. MSM ’12, pp. 13–20. DOI: 10.1145/2310057.2310061.

Cho, Eunjoon, Seth A. Myers, and Jure Leskovec (2011). “Friendship and mobility: User movement in location-based social networks”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD ’11, pp. 1082–1090. DOI: 10.1145/2020408.2020579.

Choudhury, Tanzeem (2004). “Sensing and modeling human networks”. PhD thesis. Massachusetts Institute of Technology.

Choudhury, Tanzeem and Alex Pentland (2002). “The sociometer: A wearable device for understanding human networks”. In: *Proceedings of the Workshop on Ad hoc Communications and Collaboration in Ubiquitous Computing Environments, Computer Supported Cooperative Work*.

Chris, Chatfield (2002). “Confessions of a pragmatic statistician”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 51 (1), pp. 1–20. DOI: 10.1111/1467-9884.00294.

Christ, Maximilian, Andreas W. Kempa-Liehr, and Michael Feindt (2016). “Distributed and parallel time series feature extraction for industrial big data applications”. eprint: https://arxiv.org/abs/1610.07717.

Christakis, Nicholas A. and James H. Fowler (2010). “Social network sensors for early detection of contagious outbreaks”. In: *PLOS ONE* 5 (9), pp. 1–8. DOI: 10.1371/journal.pone.0012948.

Chronis, Iolanthe, Anmol Madan, and Alex Pentland (2009). “SocialCircuits: The art of using mobile phones for modeling personal interactions”. In: *Proceedings of the ICMI-MLMI ’09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing*. ICMI-MLMI ’09, 1:1–1:4. DOI: 10.1145/1641389.1641390.

Chu, Kar-Hai, Jennifer B. Unger, Jon-Patrick Allem, Monica Pattarroyo, Daniel Soto, Tess Boley Cruz, Haodong Yang, Ling Jiang, and Christopher C. Yang (2015). “Diffusion of messages from an electronic cigarette brand to potential users through Twitter”. In: *PLOS ONE* 10 (12), e0145387. DOI: 10.1371/journal.pone.0145387.

Chung, Jae Eun (2016). “A smoking cessation campaign on Twitter: Understanding the use of Twitter and identifying major players in a health campaign”. In: *Journal of Health Communication* 21 (5), pp. 517–526. DOI: 10.1080/10810730.2015.1103332.

Ciavarella, Constanze, Laura Fumanelli, Stefano Merler, Ciro Cattuto, and Marco Ajelli (2016). “School closure policies at municipality level for mitigating influenza spread: a model-based evaluation”. In: *BMC Infectious Diseases* 16 (576), pp. 1–11. DOI: 10.1186/s12879-016-1918-z.

Clark, Meredith D. (2014). “To tweet our own cause: A mixed-methods study of the online phenomenon ‘Black Twitter’”. PhD thesis. The University of North Carolina at Chapel Hill, School of Journalism and Mass Communication. URL: http://search.proquest.com/docview/1648168732.

Clauset, Aaron and Nathan Eagle (2012). “Persistence and periodicity in a dynamic proximity network”. eprint: https://arxiv.org/abs/1211.7343.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E. J. Newman (2009). “Power-law distributions in empirical data”. In: *SIAM Review* 51 (4), pp. 661–703. DOI: 10.1137/070710111.

Cohen, Raviv and Derek Ruths (2013). “Classifying political orientation on Twitter: It’s not easy!” In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. ICWSM-13, pp. 91–99. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128.

Cohen-Cole, Ethan and Jason M. Fletcher (2008). “Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic”. In: *Journal of Health Economics* 27 (5), pp. 1382–1387. DOI: 10.1016/j.jhealeco.2008.04.005.

Cole, David (2014). “‘We kill people based on metadata’”. In: *New York Review of Books Daily* (10 May 2014). URL: http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/.

Cole-Lewis, Heather, Arun Varghese, Amy Sanders, Mary Schwarz, Jillian Pugatch, and Erik Augustson (2015). “Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning”. In: *Journal of Medical Internet Research* 17 (8), e208. DOI: 10.2196/jmir.4392.

Conte, Rosaria, Nigel G. Gilbert, Giulia Bonelli, Claudio A. Cioffi-Revilla, Guillaume Deffuant, János Kertész, Vittorio Loreto, Susannah Helen Moat, Jean-Pierre Nadal, Angel Sanchez, Andrzej Nowak, Andreas Flache, Maxi San Miguel, and Dirk Helbing (2012). “Manifesto of computational social science”. In: *The European Physical Journal Special Topics* 214 (1), pp. 325–346. DOI: 10.1140/epjst/e2012-01697-8.

Corten, Rense (2012). “Composition and structure of a large online social network in the Netherlands”. In: *PLOS ONE* 7 (4), e34760. DOI: 10.1371/journal.pone.0034760.

Cottica, Alberto, Guy Melançon, and Benjamin Renoust (2017). “Online community management as social network design: Testing for the signature of management activities in online communities”. In: *Applied Network Science* 2 (1). DOI: 10.1007/s41109-017-0049-9.

Cox, D. R. (1990). “Role of models in statistical analysis”. In: *Statistical Science* (5), pp. 169–174. DOI: 10.1214/ss/1177012165.

Crampton, Jeremy W., Mark Graham, Ate Poorthuis, Taylor Shelton, Monica Stephens, Matthew W. Wilson, and Matthew Zook (2013). “Beyond the geotag: Situating ‘big data’ and leveraging the potential of the geoweb”. In: *Cartography and Geographic Information Science* 40 (2), pp. 130–139. DOI: 10.1080/15230406.2013.777137.

Cranshaw, Justin, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh (2010). “Bridging the gap between physical location and online social networks”. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. Ubicomp ’10, pp. 119–128. DOI: 10.1145/1864349.1864380.

Csárdi, Gábor and Tamás Nepusz (2006). “The igraph software package for complex network research”. In: *InterJournal* Complex Systems (1695). URL: http://igraph.org.

Dabbs, Beau and Brian Junker (2016). “Comparison of cross-validation methods for stochastic block models”. eprint: https://arxiv.org/abs/1612.04717.

Dalton, Craig and Jim Thatcher (2014). “What does a critical data studies look like, and why do we care?” In: *Society and Space* (12 May 2014). URL: http://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher.

de Freitas, Adrian A. and Anind K. Dey (2015). “Using multiple contexts to detect and form opportunistic groups”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’15, pp. 1612–1621. DOI: 10.1145/2675133.2675213.

Depue, Jacob B., Brian G. Southwell, Anne E. Betzner, and Barbara M. Walsh (2015). “Encoded exposure to tobacco use in social media predicts subsequent smoking behavior”. In: *American Journal of Health Promotion* 29 (4), pp. 259–261. DOI: 10.4278/ajhp.130214-arb-69.

DeVellis, Robert F. (2017). *Scale development: Theory and applications*. 4th ed. Los Angeles, CA: SAGE.

Dey, Anind K., Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos (2011). “Getting closer: An empirical investigation of the proximity of user to their smart phones”. In: *Proceedings of the 13th International Conference on Ubiquitous Computing*. UbiComp ’11, pp. 163–172. DOI: 10.1145/2030112.2030135.

Dey, Ratan, Zubin Jelveh, and Keith Ross (2012). “Facebook users have become much more private: A large-scale study”. In: *Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. PERCOM Workshops, pp. 346–352. DOI: 10.1109/PerComW.2012.6197508.

Dhar, Vasant (2013). “Data science and prediction”. In: *Communications of the ACM* 56 (12), pp. 64–73. DOI: 10.1145/2500499.

Diakopoulos, Nicholas (2014). “Algorithmic accountability reporting: On the investigation of black boxes”. Tow Center for Digital Journalism, Columbia Journalism School. URL: http://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2/.

– (2015). “Algorithmic accountability: Journalistic investigation of computational power structures”. In: *Digital Journalism* 3 (3), pp. 398–415. DOI: 10.1080/21670811.2014.976411.

Dillman, Don A., Jolene D. Smyth, and Leah Melani Christian (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. 4th ed. Hoboken, NJ: John Wiley & Sons, Inc.

Dixon, Kitsy (2014). “Feminist online identity: Analyzing the presence of hashtag feminism”. In: *Journal of Arts and Humanities* 3 (7), pp. 34–40. URL: https://www.theartsjournal.org/index.php/site/article/view/509.

Do, Trinh Minh Tri and Daniel Gatica-Perez (2011). “GroupUs: Smartphone proximity data and human interaction type mining”. In: *Proceedings of the 15th Annual International Symposium on Wearable Computers*. ISWC 2011, pp. 21–28. DOI: 10.1109/ISWC.2011.28.

Donath, Judith (2007). “Signals in social supernets”. In: *Journal of Computer-Mediated Communication* 13 (1), pp. 231–251. DOI: 10.1111/j.1083-6101.2007.00394.x.

Donath, Judith and danah m. boyd (2004). “Public displays of connection”. In: *BT Technology Journal* 22 (4), pp. 71–82. DOI: 10.1023/B:BTTJ.0000047585.06264.cc.

Dong, Wen, Bruno Lepri, and Alex Pentland (2011). “Modeling the co-evolution of behaviors and social relationships using mobile phone data”. In: *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*. MUM ’11, pp. 134–143. DOI: 10.1145/2107596.2107613.

Doom, Jenalee R., Colleen M. Doyle, and Megan R. Gunnar (2017). “Social stress buffering by friends in childhood and adolescence: Effects on HPA and oxytocin activity”. In: *Social Neuroscience* 12 (1), pp. 8–21. DOI: 10.1080/17470919.2016.1149095.

Doran, Derek, Swapna Gokhale, and Aldo Dagnino (2013). “Human sensing for smart cities”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’13, pp. 1323–1330. DOI: 10.1145/2492517.2500240.

Dourish, Paul (2016). “Algorithms and their others: Algorithmic culture in context”. In: *Big Data & Society* 3 (2). DOI: 10.1177/2053951716665128.

Drucker, Karin (2017). “Electronic monitoring: Punishment and liberty in the age of GPS”. Harvard Civil Rights-Civil Liberties Law Review Amicus Blog (23 October 2017). URL: http://harvardcrcl.org/electronic-monitoring-punishment-and-liberty-in-the-age-of-gps-2/.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden (2015). “Demographics of key social networking platforms”. Pew Research Center: Internet, Science & Tech. URL: http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/.

Duijn, Marijtje A. J. van, Evelien P. H. Zeggelink, Mark Huisman, Frans N. Stokman, and Frans W. Wasseur (2003). “Evolution of sociology freshmen into a friendship network”. In: *The Journal of Mathematical Sociology* 27 (2-3), pp. 153–191. DOI: 10.1080/00222500305889.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth (2015). “Preserving statistical validity in adaptive data analysis”. In: *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*. STOC ’15, pp. 117–126. DOI: 10.1145/2746539.2746580.

Eagle, Nathan (2005). “Machine perception and learning of complex social systems”. PhD thesis. Massachusetts Institute of Technology.

Eagle, Nathan, Aaron Clauset, Alex Pentland, and David Lazer (2010). “Reply to adams: Multi-dimensional edge inference”. In: *Proceedings of the National Academy of Sciences* 107 (9), E31. DOI: 10.1073/pnas.0913678107.

Eagle, Nathan and Alex Pentland (2006). “Reality mining: Sensing complex social systems”. In: *Personal Ubiquitous Computing* 10 (4), pp. 255–268. DOI: 10.1007/s00779-005-0046-3.

Eagle, Nathan, Alex Pentland, and David Lazer (2009). “Inferring friendship network structure by using mobile phone data”. In: *Proceedings of the National Academy of Sciences* 106 (36), pp. 15274–15278. DOI: 10.1073/pnas.0900282106.

Eberle, Julia, Karsten Stegmann, Frank Fischer, Alain Barrat, and Kristine Lund (2017). “Finding collaboration partners in a scientific community: The role of cognitive group awareness, career level, and disciplinary background collaboration and integration of newcomers in scientific communities”. In: *Proceedings of the 12th International Conference on Computer Supported Collaborative Learning*. CSCL 2017, pp. 519–526.

Efron, Bradley and Carl Morris (1977). “Stein’s paradox in statistics”. In: *Scientific American* 236 (5), pp. 119–127. DOI: 10.1038/scientificamerican0577-119.

Efstathiades, Hariton, Demetris Antoniades, George Pallis, Marios D. Dikaiakos, Zoltán Szlávik, and Robert-Jan Sips (2016). “Online social network evolution: Revisiting the Twitter graph”. In: *Proceedings of the 2016 IEEE International Conference on Big Data*. IEEE Big Data 2016, pp. 626–635. DOI: 10.1109/BigData.2016.7840655.

Eisenstein, Jacob, Brendan O’Connor, Noah A. Smith, and Eric P. Xing (2010). “A latent variable model for geographic lexical variation”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’10, pp. 1277–1287. URL: http://dl.acm.org/citation.cfm?id=1870658.1870782.

Elmer, Timor, K. Chaitanya, P. Purwar, and Christoph Stadtfeld (Under review). “Testing and improving the validity of RFID tags measuring face-to-face interactions”.

Erikson, Emily (2013). “Formalist and relationalist theory in social network analysis”. In: *Sociological Theory* 31 (3), pp. 219–242. DOI: 10.1177/0735275113501998.

Eubanks, Virginia (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Farrelly, Matthew C., Cheryl G. Healton, Kevin C. Davis, Peter Messeri, James C. Hersey, and M. Lyndon Haviland (2002). "Getting to the truth: Evaluating national tobacco countermarketing campaigns". In: *American Journal of Public Health* 92 (6), pp. 901–907. DOI: 10.2105/ajph.92.6.901.

Fehr, Beverley (1996). *Friendship processes*. Thousand Oaks, California: SAGE.

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015). "Certifying and removing disparate impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15, pp. 259–268. DOI: 10.1145/2783258.2783311.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *Journal of Machine Learning Research* 15 (1), pp. 3133–3181. URL: http://jmlr.org/papers/v15/delgado14a.html.

Ferreira, Denzil, Vassilis Kostakos, and Anind K. Dey (2015). "AWARE: Mobile context instrumentation framework". In: *Frontiers in ICT* 2 (6), pp. 1–9. DOI: 10.3389/fict.2015.00006.

Festinger, Leon, Kurt W. Back, and Stanley Schachter (1950). *Social pressure in informal groups: A study of human factors in housing*. Stanford, CA: Stanford University Press.

Fiorio, Lee, Guy Abel, Jixuan Cai, Emilio Zagheni, Ingmar Weber, and Guillermo Vinué (2017). "Using Twitter data to estimate the relationship between short-term mobility and long-term migration". In: *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17, pp. 103–110. DOI: 10.1145/3091478.3091496.

Fisher, Ronald A. (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 222, pp. 309–368. DOI: 10.1098/rsta.1922.0009.

Florini, Sarah (2014). "Tweets, tweeps, and signifyin': Communication and cultural performance on 'Black Twitter'". In: *Television & New Media* 15 (3), pp. 223–237. DOI: 10.1177/1527476413480247.

Förster, Anna, Kamini Garg, Hoang Anh Nguyen, and Silvia Giordano (2012). "On context awareness and social distance in human mobility traces". In: *Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks*. MobiOpp '12, pp. 5–12. DOI: 10.1145/2159576.2159581.

Forster, Malcolm and Elliott Sober (1994). "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions". In: *The British Journal for the Philosophy of Science* 45 (1), pp. 1–35. DOI: doi.org/10.1093/bjps/45.1.1.

Foucault Welles, Brooke (2014). "On minorities and outliers: The case for making Big Data small". In: *Big Data & Society* 1 (1), pp. 1–2. DOI: 10.1177/2053951714540613.

Fournet, Julie and Alain Barrat (2014). "Contact patterns among high school students". In: *PLOS ONE* 9 (9), pp. 1–17. DOI: 10.1371/journal.pone.0107878.

– (2016). "Epidemic risk from friendship network data: an equivalence with a non-uniform sampling of contact networks". In: *Scientific Reports* 6 (1). DOI: 10.1038/srep24593.

– (2017). "Estimating the epidemic risk using non-uniformly sampled contact data". In: *Scientific Reports* 7 (1). DOI: 10.1038/s41598-017-10340-y.

Frank, Eibe, Mark A. Hall, and Ian H. Witten (2016). "The WEKA workbench". In: *Data mining: Practical machine learning tools and techniques (online appendix)*. 4th ed. Morgan Kaufmann.

Freed, Diana, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell (2018). "'A stalker's paradise': How intimate partner abusers exploit technology". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18, 667:1–667:13. DOI: 10.1145/3173574.3174241.

Freedman, David A. (1991). "Statistical models and shoe leather". In: *Sociological Methodology* 21, pp. 291–313. DOI: 10.2307/270939.

– (1997). "Some issues in the foundation of statistics". In: *Topics in the foundation of statistics*. Ed. by Bas C. van Fraassen. Dordrecht: Springer Netherlands, pp. 19–39. DOI: 10.1007/978-94-015-8816-4_4.

– (2004). "Graphical models for causation, and the identification problem". In: *Evaluation Review* 28 (4), pp. 267–293. DOI: 10.1177/0193841X04266432.

Freedman, David A., Stephen P. Klein, Michael Ostland, and Michael R. Roberts (2009a). "On 'solutions' to the ecological inference problem". In: *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press, pp. 83–96.

– (2009b). "Rejoinder to King". In: *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press, pp. 97–104.

Freeman, Linton C., A. Kimball Romney, and Sue C. Freeman (1987). "Cognitive structure and informant accuracy". In: *American Anthropologist* 89 (2), pp. 310–325. DOI: 10.1525/aa.1987.89.2.02a00020.

Frias-Martinez, Vanessa, Victor Soto, Heath Hohwald, and Enrique Frias-Martinez (2012). "Characterizing urban landscapes using geolocated tweets". In: *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*. SOCIALCOM-PASSAT '12, pp. 239–248. DOI: 10.1109/SocialCom-PASSAT.2012.19.

Friggeri, Adrien, Guillaume Chelius, Eric Fleury, Antoine Fraboulet, France Mentré, and Jean-Christophe Lucet (2011). "Reconstructing social interactions using an unreliable wireless sensor network". In: *Computer Communications* 34 (5), pp. 609–618. DOI: 10.1016/j.comcom.2010.06.005.

Gable, Robert S. (2017). "On their last legs: Smartphones should replace GPS ankle bracelets for monitoring offenders". In: *IEEE Spectrum* 54 (8), pp. 44–49. DOI: 10.1109/MSPEC.2017.8000290.

Gaetan, Carlo and Xavier Guyon (2012). *Spatial statistics and modeling*. Springer Series in Statistics. New York: Springer. DOI: 10.1007/978-0-387-92257-7.

Gaffney, Devin and Cornelius Puschmann (2014). "Data collection on Twitter". In: *Twitter and society*. Ed. by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. New York: Peter Lang, pp. 55–67. DOI: 10.3726/978-1-4539-1170-9.

Garnett, Roman, Michael A. Osborne, Steven Reece, Alex Rogers, and Stephen J. Roberts (2010). "Sequential Bayesian prediction in the presence of changepoints and faults". In: *The Computer Journal* 53 (9), pp. 1430–1446. DOI: 10.1093/comjnl/bxq003.

Gayo-Avello, Daniel (2011). "Don't turn social media into another 'Literary Digest' poll". In: *Communications of the ACM* 54 (10), pp. 121–128. DOI: 10.1145/2001269.2001297.

– (2012). "No, you cannot predict elections with Twitter". In: *IEEE Internet Computing* 16 (6), pp. 91–94. DOI: 10.1109/MIC.2012.137.

– (2013). "A meta-analysis of state-of-the-art electoral prediction From Twitter data". In: *Social Science Computer Review* 31 (6), pp. 649–679. DOI: 10.1177/0894439313493979.

Geer, Sara A. van de and Peter Bühlmann (2009). "On the conditions used to prove oracle results for the lasso". In: *Electronic Journal of Statistics* (3), pp. 1360–1392. DOI: 10.1214/09-EJS506.

Gehl, Robert W. (2014). *Reverse engineering social media: Software, culture, and political economy in new media capitalism*. Philadelphia, PA: Temple University Press.

Gelman, Andrew (2009). "A statistician's perspective on *Mostly harmless econometrics: An empiricist's companion*, by Joshua D. Angrist and Jörn-Steffen Pischke". In: *Stata Journal* 9 (2), pp. 315–320. URL: http://www.stata-journal.com/article.html?article=gn0046.

Gelman, Andrew and Cosma Rohilla Shalizi (2012). "Philosophy and the practice of Bayesian statistics in the social sciences". In: *British Journal of Mathematical and Statistical Psychology* 66 (1), pp. 8–38. DOI: 10.1111/j.2044-8317.2011.02037.x.

Gemmetto, Valerio, Alain Barrat, and Ciro Cattuto (2014). "Mitigation of infectious disease at school: Targeted class closure vs school closure". In: *BMC Infectious Diseases* 14 (695), pp. 1–10. DOI: 10.1186/s12879-014-0695-9.

Génois, Mathieu and Alain Barrat (2018). "Can co-location be used as a proxy for face-to-face contacts?" In: *EPJ Data Science* 7 (11), pp. 1–18. DOI: 10.1140/epjds/s13688-018-0140-1.

Génois, Mathieu, Christian L. Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat (2015). "Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers". In: *Network Science* 3 (3), pp. 326–347. DOI: 10.1017/nws.2015.10.

Ghose, Avik, Chirabrata Bhaumik, and Tapas Chakravarty (2013). "BlueEye: A system for proximity detection using Bluetooth on mobile phones". In: *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*. UbiComp '13 Adjunct, pp. 1135–1142. DOI: 10.1145/2494091.2499771.

Ghosh, Debarchana (Debs) and Rajarshi Guha (2013). "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System". In: *Cartography and Geographic Information Science* 40 (2), pp. 90–102. DOI: 10.1080/15230406.2013.776210.

Ghosh, Saptarshi, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi (2012). "Understanding and combating link farming in the Twitter social network". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12, pp. 61–70. DOI: 10.1145/2187836.2187846.

Gilbert, Eric and Karrie Karahalios (2009). "Predicting tie strength with social media". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09, pp. 211–220. DOI: 10.1145/1518701.1518736.

Gilbert, Nigel and Klaus G. Troitzsch (2005). *Simulation for the social scientist*. Berkshire, UK: Open University Press.

Gillespie, Tarleton (2014). "The relevance of algorithms". In: *Media technologies: Essays on communication, materiality, and society*. Ed. by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. DOI: 10.7551/mitpress/9780262525374.003.0009.

Gillespie, Tarleton and Nick Seaver (2016). "Critical algorithm studies: A reading list". Social Media Collective Research Blog (15 December 2016). URL: https://socialmediacollective.org/reading-lists/critical-algorithm-studies/.

Golder, Scott A. and Michael W. Macy (2014). "Digital footprints: Opportunities and challenges for online social research". In: *Annual Review of Sociology* 40 (1), pp. 129–152. DOI: 10.1146/annurev-soc-071913-043145.

Golder, Scott and Michael W. Macy (2012). "Social science with social media". In: *ASA footnotes* 40 (1). URL: http://www.asanet.org/footnotes/jan12/socialmedia_0112.html.

Goodchild, Michael F. (2007). “Citizens as sensors: The world of volunteered geography”. In: *GeoJournal* 69 (4), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

Gourarie, Chava (2016). “Investigating the algorithms that govern our lives”. In: *Columbia Journalism Review* (14 April 2016). URL: http://www.cjr.org/innovations/investigating_algorithms.php.

Graham, Mark, Scott A. Hale, and Devin Gaffney (2014). “Where in the world are you? Geolocation and language identification in Twitter”. In: *The Professional Geographer* 66 (4), pp. 568–578. DOI: 10.1080/00330124.2014.907699.

Grant, Roy (2017). “Public health professionals urgently need to develop more effective communications strategies”. In: *American Journal of Public Health* 107 (5), pp. 658–659. DOI: 10.2105/ajph.2017.303738.

Grebe, Nicholas M., Andreas Aarseth Kristoffersen, Trond Viggo Grøntvedt, Melissa Emery Thompson, Leif Edward Ottesen Kennair, and Steven W. Gangestad (2017). “Oxytocin and vulnerable romantic relationships”. In: *Hormones and Behavior* 90, pp. 64–74. DOI: 10.1016/j.yhbeh.2017.02.009.

Greenwood, Daniel, Arkadiusz Stopczynski, Brian Sweatt, Thomas Hardjono, and Alex Pentland (2014). “The New Deal on Data: A framework for institutional controls”. In: *Privacy, big data, and the public good: Frameworks for engagement*. Ed. by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum. Cambridge University Press, pp. 192–210. DOI: 10.1017/CBO9781107590205.012.

Grothoff, Christian and J. M. Porup (2016). “The NSA’s SKYNET program may be killing thousands of innocent people: ‘Ridiculously optimistic’ machine learning algorithm is ‘completely bullshit,’ says expert”. In: *Ars Technica* (16 February 2016). URL: https://arstechnica.com/information-technology/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/.

Grove, William M. (2005). “Clinical versus statistical prediction: The contribution of Paul E. Meehl”. In: *Journal of Clinical Psychology* 61 (10), pp. 1233–1243. DOI: 10.1002/jclp.20179.

Grove, William M. and Paul E. Meehl (1996). “Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy”. In: *Psychology, Public Policy, and Law* 2 (2), pp. 293–323. DOI: 10.1037/1076-8971.2.2.293.

Guo, Diansheng and Chao Chen (2014). “Detecting non-personal and spam users on geo-tagged Twitter network”. In: *Transactions in GIS* 18 (3), pp. 370–384. DOI: 10.1111/tgis.12101.

Hacking, Ian (2000). *The social construction of what?* Cambridge, MA: Harvard University Press.

– (2007). “Kinds of people: Moving targets”. In: *Proceedings of the British Academy*. Vol. 151, pp. 285–318. DOI: 10.5871/bacad/9780197264249.003.0010.

Haddad, Linda, Debra Lynch Kelly, Linda S. Weglicki, Tracey E. Barnett, Anastasiya V. Ferrell, and Roula Ghadban (2016). “A systematic review of effects of waterpipe smoking on cardiovascular and respiratory health outcomes”. In: *Tobacco Use Insights* 9. DOI: 10.4137/tui.s39873.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001). “Identification and estimation of treatment effects with a regression-discontinuity design”. In: *Econometrica* 69 (1), pp. 201–209. DOI: 10.1111/1468-0262.00183.

Halevy, Alon, Peter Norvig, and Fernando Pereira (2009). “The unreasonable effectiveness of data”. In: *IEEE Intelligent Systems* 24 (2), pp. 8–12. DOI: 10.1109/MIS.2009.36.

Hall, Mark A. (1999). “Correlation-based feature selection for machine learning”. PhD thesis. Department of Computer Science, The University of Waikato.

Hall, Mark A., Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). "The WEKA data mining software: An update". In: *SIGKDD Exploration Newsletter* 11 (1), pp. 10–18. DOI: 10.1145/1656274.1656278.

Hammerla, Nils Y. and Thomas Plötz (2015). "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '15, pp. 1041–1051. DOI: 10.1145/2750858.2807551.

Hargittai, Eszter (2015). "Is bigger always better? Potential biases of big data derived from social network sites". In: *The ANNALS of the American Academy of Political and Social Science* 659 (1), pp. 63–76. DOI: 10.1177/0002716215570866.

– (2018). "Potential biases in big data: Omitted voices on social media". In: *Social Science Computer Review*. DOI: 10.1177/0894439318788322.

Hargittai, Eszter and Eden Litt (2011). "The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults". In: *New Media & Society* (13), pp. 824–842. DOI: 10.1177/1461444811405805.

Harper, David, Darren Ellis, and Ian Tucker (2014). "Surveillance". In: *Encyclopedia of critical psychology*. Ed. by Thomas Teo. Springer, pp. 1887–1892. DOI: 10.1007/978-1-4614-5583-7_305.

Harris, Jenine K., Sarah Moreland-Russell, Bechara Choucair, Raed Mansour, Mackenzie Staub, and Kendall Simmons (2014). "Tweeting for and against public health policy: Response to the Chicago Department of Public Health's electronic cigarette Twitter campaign". In: *Journal of Medical Internet Research* 16 (10), e238. DOI: 10.2196/jmir.3622.

Harris, Shane (2013). "The cowboy of the NSA: Inside Gen. Keith Alexander's all-out, barely-legal drive to build the ultimate spy machine". In: *Foreign Policy* (9 September 2013). URL: https://foreignpolicy.com/2013/09/09/the-cowboy-of-the-nsa/.

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti (2014). "Geo-located Twitter as proxy for global mobility patterns". In: *Cartography and Geographic Information Science* 41 (3), pp. 260–271. DOI: 10.1080/15230406.2014.890072.

Hayes, Alexander (2017). "Pervasive technology: Aboriginal communities and oppression [opinion]". In: *IEEE Technology and Society Magazine* 36 (4), pp. 25–54. DOI: 10.1109/MTS.2017.2763445.

Healy, Kieran (2015). "The performativity of networks". In: *European Journal of Sociology* 56 (2), pp. 175–205. DOI: 10.1017/S0003975615000107.

Hecht, Brent, Lichan Hong, Bongwon Suh, and Ed H. Chi (2011). "Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11, pp. 237–246. DOI: 10.1145/1978942.1978976.

Hecht, Brent and Monica Stephens (2014). "A tale of cities: Urban biases in volunteered geographic information". In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. ICWSM-14, pp. 197–205. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114.

Hempel, Jessi (2015). "Banks are now handing out loans to people they'd normally shun". In: *Wired* (20 January 2015). URL: https://www.wired.com/2015/01/banks-handing-loans-people-normally-shun/.

Hendeby, Gustaf, Fredrik Gustafsson, and Niklas Wahlström (2014). "Teaching sensor fusion and Kalman filtering using a smartphone". In: *IFAC Proceedings Volumes* 47 (3), pp. 10586–10591. DOI: 10.3182/20140824-6-ZA-1003.00967.

Hennig, Marina, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel (2013). *Studying social networks: A guide to empirical research*. Frankfurt, Germany: Campus Verlag.

Hertzberg, Vicki Stover, Jason Baumgardner, C. Christina Mehta, Lisa K. Elon, George Cotsonis, and Douglas W. Lowery-North (2017). "Contact networks in the emergency department: Effects of time, environment, patient characteristics, and staff role". In: *Social Networks* 48, pp. 181–191. DOI: 10.1016/j.socnet.2016.08.005.

Hess, Amanda (2014). "Why women aren't welcome on the Internet". In: *Pacific Standard Magazine* 7 (1). URL: https://psmag.com/social-justice/women-arent-welcome-internet-72170.

Hidalgo, César A. (2016). "Disconnected, fragmented, or united? A trans-disciplinary review of network science". In: *Applied Network Science* 1 (6), pp. 1–19. DOI: 10.1007/s41109-016-0010-3.

Hildner, Laura (2006). "Defusing the threat of RFID: Protecting consumer privacy through technology-specific legislation at the state level". In: *Harvard Civil Rights-Civil Liberties Law Review* (41), pp. 133–176.

Hindman, Matthew (2015). "Building better models: Prediction, replication, and machine learning in the social sciences". In: *The ANNALS of the American Academy of Political and Social Science* 659 (1), pp. 48–62. DOI: 10.1177/0002716215570279.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts (2017). "Prediction and explanation in social systems". In: *Science* 355 (6324), pp. 486–488. DOI: 10.1126/science.aal3856.

Hogan, Bernie, Juan Antonio Carrasco, and Barry Wellman (2007). "Visualizing personal networks: Working with participant-aided sociograms". In: *Field Methods* 19 (2), pp. 116–144. DOI: 10.1177/1525822X06298589.

Honeycutt, Courtenay and Susan C. Herring (2009). "Beyond microblogging: Conversation and collaboration via Twitter". In: *Proceedings of the 42nd Hawaii International Conference on System Sciences*. HICSS-42, pp. 1–10. DOI: 10.1109/HICSS.2009.602.

Hong, Liangjie, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis (2012). "Discovering geographical topics in the Twitter stream". In: WWW '12, pp. 769–778. DOI: 10.1145/2187836.2187940.

Hossain, A. K. M. Mahtab and Wee-Seng Soh (2007). "A comprehensive study of Bluetooth signal parameters for localization". In: *Proceedings of the 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. PIMRC'07, pp. 1–5. DOI: 10.1109/PIMRC.2007.4394215.

Hossain, Nabil, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry A. Kautz (2016). "Precise localization of homes and activities: Detecting drinking-while-tweeting patterns in communities". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 587–590. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13118.

Hristova, Desislava, Matthew J. Williams, Mirco Musolesi, Pietro Panzarasa, and Cecilia Mascolo (2016). "Measuring urban social diversity using interconnected geo-social networks". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16, pp. 21–30. DOI: 10.1145/2872427.2883065.

Hsieh, Jeng-Cheng, Chih-Ming Chen, and Hsiao-Fang Lin (2010). "Social interaction mining based on wireless sensor networks for promoting cooperative learning performance in classroom learning environment". In: *Proceedings of the 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education*. WMUTE 2010, pp. 219–221. DOI: 10.1109/WMUTE.2010.22.

Huang, Jeff, Katherine M. Thornton, and Efthimis N. Efthimiadis (2010). "Conversational tagging in Twitter". In: *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. HT '10, pp. 173–178. DOI: 10.1145/1810617.1810647.

Huberman, Bernardo A. (2012). "Sociology of science: Big data deserve a bigger audience". In: *Nature* 482 (7385), p. 308. DOI: 10.1038/482308d.

Hung, Hayley, Gwenn Englebienne, and Jeroen Kools (2013). "Classifying social actions with a single accelerometer". In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '13, pp. 207–210. DOI: 10.1145/2493432.2493513.

Hyndman, Rob J., Anne B. Koehler, Ralph D. Snyder, and Simone Grose (2002). "A state space framework for automatic forecasting using exponential smoothing methods". In: *International Journal of Forecasting* 18 (3), pp. 439–454. DOI: 10.1016/S0169-2070(01)00110-8.

Iliadis, Andrew and Federica Russo (2016). "Critical data studies: An introduction". In: *Big Data & Society* 3 (2). DOI: 10.1177/2053951716674238.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica* 62 (2), pp. 467–475. DOI: 10.2307/2951620.

Imbens, Guido W. and Thomas Lemieux (2008). "Regression discontinuity designs: A guide to practice". In: *Journal of Econometrics* 142 (2), pp. 615–635. DOI: 10.1016/j.jeconom.2007.05.001.

Isella, Lorenzo, Mariateresa Romano, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Wouter Van den Broeck, Francesco Gesualdo, Elisabetta Pandolfi, Lucilla Ravà, Caterina Rizzo, and Alberto Eugenio Tozzi (2011). "Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors". In: *PLOS ONE* 6 (2), pp. 1–10. DOI: 10.1371/journal.pone.0017144.

Isella, Lorenzo, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter van den Broeck (2011). "What's in a crowd? Analysis of face-to-face behavioral networks". In: *Journal of Theoretical Biology* 271 (1), pp. 166–180. DOI: 10.1016/j.jtbi.2010.11.033.

Ishida, Kazunari (2012). "Geographical bias on social media and geo-local contents system with mobile devices". In: *Proceedings of the 45th Hawaii International Conference on System Sciences*. HICSS '12, pp. 1790–1796. DOI: 10.1109/HICSS.2012.292.

Iyengar, Shanto and Sean J. Westwood (2014). "Fear and loathing across party lines: New evidence on group polarization". In: *American Journal of Political Science* 59 (3), pp. 690–707. DOI: 10.1111/ajps.12152.

Jacobs, Abigail Z. (2017). "Comparative, population-level analysis of social networks in organizations". PhD thesis. University of Colorado at Boulder. URL: https://scholar.colorado.edu/csci_gradetds/147.

Jacobs, Abigail Z., Samuel F. Way, Johan Ugander, and Aaron Clauset (2015). "Assembling Thefacebook: Using heterogeneity to understand online social network assembly". In: *Proceedings of the ACM Web Science Conference*. WebSci '15, 18:1–18:10. DOI: 10.1145/2786451.2786477.

Jamal, Ahmed, Andrea Gentzke, S. Sean Hu, Karen A. Cullen, Benjamin J. Apelberg, David M. Homa, and Brian A. King (2017). "Tobacco use among middle and high school students — United States, 2011–2016". In: *Morbidity and Mortality Weekly Report (MMWR)* 66 (23), pp. 597–603. DOI: 10.15585/mmwr.mm6623a1.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher (2015). *AAPOR report on Big Data*. Oakbrook Terrace, IL: American Association of Public Opinion Researchers. URL: http://www.aapor.org/Education-Resources/Reports/Big-Data.aspx.

Jasanoff, Sheila and Sang-Hyun Kim, eds. (2015). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. Chicago, IL: The University of Chicago Press.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng (2007). “Why we Twitter: Understanding microblogging usage and communities”. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pp. 56–65. DOI: 10.1145/1348549.1348556.

Jawad, Mohammed, Jooman Abass, Ahmad Hariri, and Elie A. Akl (2015). “Social media use for public health campaigning in a low resource setting: The case of waterpipe tobacco smoking”. In: *BioMed Research International*, pp. 1–4. DOI: 10.1155/2015/562586.

Jayarajah, Kasthuri, Archan Misra, Xiao-Wen Ruan, and Ee-Peng Lim (2015). “Event detection: Exploiting socio-physical interactions in physical spaces”. In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM 2015, pp. 508–513. DOI: 10.1145/2808797.2809387.

Johnson, Graham R., Luke D. Knibbs, Timothy J. Kidd, Claire E. Wainwright, Michelle E. Wood, Kay A. Ramsay, Scott C. Bell, and Lidia Morawska (2016). “A novel method and its application to measuring pathogen decay in bioaerosols from patients with respiratory disease”. In: *PLOS ONE* 11 (7), pp. 1–20. DOI: 10.1371/journal.pone.0158763.

Johnson, Isaac L., Subhasree Sengupta, Johannes Schöning, and Brent Hecht (2016). “The geography and importance of localness in geotagged social media”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16, pp. 515–526. DOI: 10.1145/2858036.2858122.

Jones, Jason J., Jaime E. Settle, Robert M. Bond, Christopher J. Fariss, Cameron Marlow, and James H. Fowler (2013). “Inferring tie strength from online directed behavior”. In: *PLOS ONE* (8), e52168. DOI: 10.1371/journal.pone.0052168.

Junqué de Fortuny, Enric, David Martens, and Foster Provost (2013). “Predictive modeling with big data: Is bigger really better?” In: *Big Data* 1 (4), pp. 215–226. DOI: 10.1089/big.2013.0037.

Jurafsky, Dan, Rajesh Ranganath, and Daniel A. McFarland (2009). “Extracting social meaning: Identifying interactional style in spoken conversation”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL ’09, pp. 638–646. URL: http://dl.acm.org/citation.cfm?id=1620754.1620847.

Kadhum, Murtaza, Abed Sweidan, Ali Emad Jaffery, Adam Al-Saadi, and Brendan Madden (2015). “A review of the health effects of smoking shisha”. In: *Clinical Medicine* 15 (3), pp. 263–266. DOI: 10.7861/clinmedicine.15-3-263.

Kadushin, Charles (2012). *Understanding social networks: Theories, concepts, and findings*. Oxford, UK: Oxford University Press.

Kamath, Krishna Y., James Caverlee, Kyumin Lee, and Zhiyuan Cheng (2013). “Spatio-temporal dynamics of online memes: A study of geo-tagged tweets”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13, pp. 667–678. DOI: 10.1145/2488388.2488447.

Kane, Gerald C., Maryam Alavi, Giuseppe (Joe) Labianca, and Stephen P. Borgatti (2014). “What’s different about social media networks? A framework and research agenda”. In: *MIS Quarterly* 38 (1), pp. 274–304.

Kass, Robert E. (2011). “Statistical inference: The big picture”. In: *Statistical Science* 26 (1), pp. 1–9. DOI: 10.1214/10-STS337.

Ketay, Sarah, Keith M. Welker, and Richard B. Slatcher (2017). “The roles of testosterone and cortisol in friendship formation”. In: *Psychoneuroendocrinology* 76, pp. 88–96. DOI: 10.1016/j.psyneuen.2016.11.022.

Killworth, Peter D. and H. Russell Bernard (1976). “Informant accuracy in social network data”. In: *Human Organization* 35 (3), pp. 269–286. DOI: 10.17730/humo.35.3.10215j2m359266n2.

Killworth, Peter D. and H. Russell Bernard (1979). "Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data". In: *Social Networks* 2 (1), pp. 19–46. DOI: 10.1016/0378-8733(79)90009-1.

King, Gary (2011). "Ensuring the data-rich future of the social sciences". In: *Science* 331 (6018), pp. 719–721. DOI: 10.1126/science.1197872.

King, Gary, Ori Rosen, and Martin A. Tanner (2004). *Ecological inference: New methodological strategies*. Cambridge, UK: Cambridge University Press.

Kinsella, Sheila, Vanessa Murdock, and Neil O'Hare (2011). "'I'm eating a sandwich in Glasgow': Modeling locations with tweets". In: *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*. SMUC '11, pp. 61–68. DOI: 10.1145/2065023.2065039.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017). "The flash crash: High-frequency trading in an electronic market". In: *The Journal of Finance* 72 (3), pp. 967–998. DOI: 10.1111/jofi.12498.

Kiti, Moses C., Michele Tizzoni, Timothy M. Kinyanjui, Dorothy C. Koech, Patrick K. Munywoki, Milosch Meriac, Luca Cappa, André Panisson, Alain Barrat, Ciro Cattuto, and D. James Nokes (2016). "Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors". In: *EPJ Data Science* 5 (21), pp. 1–21. DOI: 10.1140/epjds/s13688-016-0084-2.

Kitts, James A. and Eric Quintane (forthcoming). "Rethinking social networks in the era of computational social science". In: *The Oxford handbook of social network analysis*. Ed. by Ryan Light and James Moody. Oxford University Press.

Kiukkonen, Niko, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila (2010). "Towards rich mobile phone datasets: Lausanne data collection campaign". In: *Proceedings of the 7th ACM International Conference on Pervasive Services*. ICPS '10.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015). "Prediction policy problems". In: *American Economic Review* 105 (5), pp. 491–95. DOI: 10.1257/aer.p20151023.

Koenker, Roger (2005). *Quantile regression*. Econometric Society Monographs. Cambridge, UK: Cambridge University Press.

Kolaczyk, Eric D. and Gábor Csárdi (2014). *Statistical analysis of network data with R*. Use R! New York, NY: Springer. DOI: 10.1007/978-1-4939-0983-4.

Kooti, Farshad, Haeryun Yang, Meeyoung Cha, Krishna Gummadi, and Winter Mason (2012). "The emergence of conventions in online social networks". In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ICWSM-12, pp. 194–201. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4661/4983.

Koren, Yehuda (2009a). "Collaborative filtering with temporal dynamics". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09, pp. 447–456. DOI: 10.1145/1557019.1557072.

– (2009b). "The BellKor solution to the Netflix Grand Prize". URL: http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.

Kornienko, Olga, Katherine H. Clemans, Dorothée Out, and Douglas A. Granger (2014). "Hormones, behavior, and social network analysis: Exploring associations between cortisol, testosterone, and network structure". In: *Hormones and Behavior* 66 (3), pp. 534–544. DOI: 10.1016/j.yhbeh.2014.07.009.

Kostakos, Vassilis and Eamonn O'Neill (2008). "Cityware: Urban computing to bridge online and real-world social networks". In: *Handbook of research on urban informatics: The practice and promise of the* *real-time city*. Ed. by Marcus Foth. Hershey, PA: Information Science Reference. DOI: 10.4018/978-1-60566-152-0.ch013.

Krackhardt, David (1987). "Cognitive social structures". In: *Social Networks* 9 (2), pp. 109–134. DOI: 10.1016/0378-8733(87)90009-8.

– (1996). "Social networks and the liability of newness for managers". In: *Trends in Organizational Behavior*. Ed. by Cary L. Cooper and Denise M. Rousseau. John Wiley & Sons, Inc., pp. 159–173.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences* 111 (24), pp. 8788–8790. DOI: 10.1073/pnas.1320040111.

Krasnova, Hanna, Thomas Hildebrand, Oliver Guenther, Alexander Kovrigin, and Aneta Nowobilska (2008). "Why participate in an online social network? An empirical analysis". In: *Proceedings of the 2008 European Conference on Information Systems*. ECIS 2008. URL: http://aisel.aisnet.org/ecis2008/33.

Krauss, Melissa J., Shaina J. Sowles, Megan Moreno, Kidist Zewdie, Richard A. Grucza, Laura J. Bierut, and Patricia A. Cavazos-Rehg (2015). "Hookah-related Twitter chatter: A content analysis". In: *Preventing Chronic Disease* 12. DOI: 10.5888/pcd12.150140.

Kuhn, Thomas S. (1962). *The structure of scientific revolutions*. 1st ed. Chicago, IL: University of Chicago Press.

Kumar, Shamanth, Xia Hu, and Huan Liu (2014). "A behavior analytics approach to identifying tweets from crisis regions". In: *Proceedings of the 25th ACM conference on Hypertext and social media*. HT '14, pp. 255–260. DOI: 10.1145/2631775.2631814.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon (2010). "What is Twitter, a social network or a news media". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10, pp. 591–600. DOI: 10.1145/1772690.1772751.

Kwan, Grace Chi En and Marko M. Skoric (2013). "Facebook bullying: An extension of battles in school". In: *Computers in Human Behavior* 29 (1), pp. 16–25. DOI: 10.1016/j.chb.2012.07.014.

Laibowitz, Mathew, Jonathan Gips, Ryan Aylward, Alex Pentland, and Joseph A. Paradiso (2006). "A sensor network for social dynamics". In: *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks*. IPSN 2006, pp. 483–491. DOI: 10.1109/IPSN.2006.243937.

Lampe, Cliff A. C., Nicole Ellison, and Charles Steinfield (2007). "A familiar face(book): Profile elements as signals in an online social network". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07, pp. 435–444. DOI: 10.1145/1240624.1240695.

Latané, Bibb, James H. Liu, Andrzej Nowak, Michael Bonevento, and Long Zheng (1995). "Distance matters: Physical space and social impact". In: *Personality and Social Psychology Bulletin* 21 (8), pp. 795–805. DOI: 10.1177/0146167295218002.

Laumann, Edward O. (1973). *Bonds of pluralism: the form and substance of urban social networks*. New York: John Wiley & Sons, Inc.

Laumann, Edward O., Peter V. Marsden, and David Prensky (1983). "The boundary specification problem in network analysis". In: *Applied network analysis: A methodological introduction*. Ed. by Ron S. Burt and Michael J. Minor. Beverly Hills, CA: SAGE, pp. 18–34.

Lawrence, Jamie, Terry R. Payne, and David De Roure (2006). "Co-presence communities: Using pervasive computing to support weak social networks". In: *Proceedings of the 15th IEEE International Workshops* *on Enabling Technologies: Infrastructure for Collaborative Enterprises*. WETICE '06, pp. 149–156. DOI: 10.1109/WETICE.2006.24.

Lazer, David (2014). "Mistaken analysis". In: *MIT Technology Review* (May/June). URL: https://www.technologyreview.com/s/526416/mistaken-analysis/.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014). "The parable of Google Flu: Traps in big data analysis". In: *Science* 343 (6176), pp. 1203–1205. DOI: 10.1126/science.1248506.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne (2009). "Computational social science". In: *Science* 323 (5915), pp. 721–723. DOI: 10.1126/science.1167742.

Lazer, David and Jason Radford (2017). "Data ex machina: Introduction to Big Data". In: *Annual Review of Sociology* 43 (1), pp. 19–39. DOI: 10.1146/annurev-soc-060116-053457.

Lee, David S. and Thomas Lemieux (2010). "Regression discontinuity designs in economics". In: *Journal of Economic Literature* 48 (2), pp. 281–355. DOI: 10.1257/jel.48.2.281.

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook (2013). "Mapping the global Twitter heartbeat: The geography of Twitter". In: *First Monday* 18 (5). URL: http://firstmonday.org/ojs/index.php/fm/article/view/4366.

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis (2008). "Tastes, ties, and time: A new social network dataset using Facebook.com". In: *Social Networks* 30 (4), pp. 330–342. DOI: 10.1016/j.socnet.2008.07.002.

LexisNexis® Risk Solutions (2014). "Survey of law enforcement personnel and their use of social media". URL: https://www.lexisnexis.com/risk/downloads/whitepaper/2014-social-media-use-in-law-enforcement.pdf.

Li, Minshu, Haipeng Wang, Bin Guo, and Zhiwen Yu (2012). "Extraction of human social behavior from mobile phone sensing". In: *Proceedings of the 8th International Conference on Active Media Technology*. AMT 2012, pp. 63–72. DOI: 10.1007/978-3-642-35236-2_7.

Li, Xitong (2013). "How does online reputation affect social media endorsements and product sales? Evidence from regression discontinuity design". In: *The 24th Workshop on Information Systems Economics*. WISE 2013.

Liere, Diederik van (2010). "How far does a tweet travel? Information brokers in the Twitterverse". In: *Proceedings of the International Workshop on Modeling Social Media*. MSM '10, 6:1–6:4. DOI: 10.1145/1835980.1835986.

Lin, Jimmy (2015). "On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences". In: *The ANNALS of the American Academy of Political and Social Science* 659 (1), pp. 33–47. DOI: 10.1177/0002716215569174.

Lin, Yu-Ru and Drew Margolin (2014). "The ripple of fear, sympathy and solidarity during the Boston bombings". In: *EPJ Data Science* 3 (31). DOI: 10.1140/epjds/s13688-014-0031-z.

Litt, Eden (2012). "Knock, knock. Who's there? The imagined audience". In: *Journal of Broadcasting & Electronic Media* 56 (3), pp. 330–345. DOI: 10.1080/08838151.2012.705195.

Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove (2014). "The tweets they are a-changin': Evolution of Twitter users and behavior". In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. ICWSM-14, pp. 305–314. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043.

Longley, Paul A., Muhammad Adnan, and Guy Lansley (2015). "The geotemporal demographics of Twitter usage". In: *Environment and Planning A* 47 (2), pp. 465–484. DOI: 10.1068/a130122p. URL: http://www.envplan.com/abstract.cgi?id=a130122p.

Lotan, Gilad (2015). "Apple's App Charts: 2015 Data and Trends ... or how much harder it is to get into the top charts". In: *i ♥ data* (15 December 2015). URL: https://medium.com/i-data/apple-s-app-charts-2015-data-and-trends-abb95300df57.

Lowly Worm (2012a). "Derwent closes shop". Buy the Hype blog (22 June 2012). URL: http://sellthenews.tumblr.com/post/25682131606/derwent-closes.

– (2012b). "The junk science behind the 'Twitter Hedge Fund'". Buy the Hype blog (14 April 2012). URL: http://sellthenews.tumblr.com/post/21067996377/noitdoesnot.

– (2013). "No limits to garbatrage". Buy the Hype blog (29 August 2013). URL: http://sellthenews.tumblr.com/post/59720892780/no-limits-to-garbatrage.

Luker, Kristin (2010). *Salsa dancing into the social sciences: Research in an age of info-glut*. Cambridge, MA: Harvard University Press.

Lunt, Christopher (2016). "Authorization and authentication based on an individual's social network". United States Patent No. 9,432,351. URL: http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=9432351.

MacDonald, Blake, Pritam Ranjan, and Hugh Chipman (2015). "GPfit: An R package for fitting a Gaussian process model to deterministic simulator outputs". In: *Journal of Statistical Software* 64 (12), pp. 1–23. DOI: 10.18637/jss.v064.i12.

MacKinlay, A. Craig (1997). "Event studies in economics and finance". In: *Journal of Economic Literature* 35, pp. 13–39.

Madan, Anmol, Manuel Cebrian, David Lazer, and Alex Pentland (2010). "Social sensing for epidemiological behavior change". In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. UbiComp '10, pp. 291–300. DOI: 10.1145/1864349.1864394.

Madan, Anmol, Katayoun Farrahi, Daniel Gatica-Perez, and Alex Pentland (2011). "Pervasive sensing to model political opinions in face-to-face networks". In: *Proceedings of the 9th International Conference on Pervasive Computing*. Pervasive 2011, pp. 214–231. DOI: 10.1007/978-3-642-21726-5_14.

Madan, Anmol, Sai T. Moturu, David Lazer, and Alex Pentland (2010a). "Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks". In: *Proceedings of Wireless Health 2010*. WH '10, pp. 104–110. DOI: 10.1145/1921081.1921094.

– (2010b). "Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks". In: *Wireless Health 2010*. WH '10, pp. 104–110. DOI: 10.1145/1921081.1921094.

Madianou, Mirca (2015a). "Digital inequality and second-order disasters: Social media in the Typhoon Haiyan recovery". In: *Social Media + Society* 1 (2), pp. 1–11. DOI: 10.1177/2056305115603386.

– (2015b). "Polymedia and ethnography: Understanding the social in social media". In: *Social Media + Society* 1 (1), pp. 1–3. DOI: 10.1177/2056305115578675.

Madianou, Mirca, Liezel Longboan, and Jonathan Ong (2015). "Finding a voice through humanitarian technologies? Communication technologies and participation in disaster recovery". In: *International Journal of Communication* 9, pp. 3020–3038. URL: http://ijoc.org/index.php/ijoc/article/view/4142.

Madianou, Mirca, Jonathan Corpus Ong, Liezel Longboan, Jayeel Cornelio, and Nicole Curato (2015). "Humanitarian technologies: Understanding the role of digital media in disaster recovery". Humanitarian

Technologies Project. URL: http://humanitariantechnologies.net/wp-content/uploads/2015/11/Humanitarian-Technologies-executive-summary-FINAL.pdf.

Maher, A. Carol, K. Lucy Lewis, Katia Ferrar, Simon Marshall, Ilse De Bourdeaudhuij, and Corneel Vandelanotte (2014). "Are health behavior change interventions that use online social networks effective? A systematic review". In: *Journal of Medical Internet Research* 16 (2), e40. DOI: 10.2196/jmir.2952.

Mahrt, Merja and Michael Scharkow (2013). "The value of Big Data in digital media research". In: *Journal of Broadcasting & Electronic Media* 57 (1), pp. 20–33. DOI: 10.1080/08838151.2012.761700.

Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer (2015). "Population bias in geotagged tweets". In: *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*. ICWSM-15 SPSM, pp. 18–27. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662.

Malik, Momin M. and Jürgen Pfeffer (2016a). "A macroscopic analysis of news content in Twitter". In: *Digital Journalism* 8 (8), pp. 955–979. DOI: 10.1080/21670811.2015.1133249.

– (2016b). "Identifying platform effects in social media data". In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 241–249.

Margolin, Drew, Devon Brewer, and David Lazer (2014). "Opportunities and limitations for research using call data records to extract personal networks". Presentation given at the Sunbelt XXXIV International Social Network Conference, session on Egocentric Networks.

Marin, Alexandra and Barry Wellman (2011). "Social network analysis: An introduction". In: *The SAGE handbook of social network analysis*. Ed. by John Scott and Peter J. Carrington. London: SAGE, pp. 11–25. DOI: 10.4135/9781446294413.n2.

Marsden, Peter V. (2011). "Survey methods for network data". In: *The SAGE handbook of social network analysis*. Ed. by John Scott and Peter J. Carrington. London: SAGE, pp. 370–388. DOI: 10.4135/9781446294413.n25.

Marwick, Alice E. and danah m. boyd (2010). "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience". In: *New Media & Society* 13 (1), pp. 114–133. DOI: 10.1177/1461444810365313.

Mast, Marianne Schmid, Daniel Gatica-Perez, Denise Frauendorfer, Laurent Nguyen, and Tanzeem Choudhury (2015). "Social sensing for psychology: Automated interpersonal behavior assessment". In: *Current Directions in Psychological Science* 24 (2), pp. 154–160. DOI: 10.1177/0963721414560811.

Mastrandrea, Rossana, Julie Fournet, and Alain Barrat (2015). "Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys". In: *PLOS ONE* 10 (9), pp. 1–26. DOI: 10.1371/journal.pone.0136497.

Mastrandrea, Rossana, Alberto Soto-Aladro, Philippe Brouqui, and Alain Barrat (2015). "Enhancing the evaluation of pathogen transmission risk in a hospital by merging hand-hygiene compliance and contact data: a proof-of-concept study". In: *BMC Research Notes* 8 (426), pp. 1–11. DOI: 10.1186/s13104-015-1409-0.

Matias, J. Nathan, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, and Charlie DeTar (2015). "Reporting, reviewing, and responding to harassment on Twitter". Women, Action, and the Media. URL: http://womenactionmedia.org/twitter-report/.

Matthews, Brian W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405 (2), pp. 442–451. DOI: 10.1016/0005-2795(75)90109-9.

McCulloh, Ian, Helen Armstrong, and Anthony Johnson (2013). *Social network analysis with applications*. Wiley.

McDermott, Drew (1976). "Artificial intelligence meets natural stupidity". In: *SIGART Bulletin* 57, pp. 4–9. DOI: 10.1145/1045339.1045340.

McDowall, David, Richard McCleary, Errol E. Meidinger, and Richard A. Hay Jr. (1980). *Interrupted time series analysis*. Quantitative applications in the social sciences. Thousand Oaks, CA: SAGE. DOI: 10.4135/9781412984607.

McNeill, Graham, Jonathan Bright, and Scott A. Hale (2016). "Estimating local commuting patterns from geolocated Twitter data". eprint: https://arxiv.org/abs/1612.01785.

McQuillan, Dan (2016). "Algorithmic paranoia and the convivial alternative". In: *Big Data & Society* 3 (2). DOI: 10.1177/2053951716671340.

Meehl, Paul E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota.

– (1996). "Preface to the 1996 printing". In: *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Lanham, MD: Jason Aronson Inc., pp. ii–ix.

Megarry, Jessica (2014). "Online incivility or sexual harassment? Conceptualising women's experiences in the digital age". In: *Women's Studies International Forum* 47 (Part A), pp. 46–55. DOI: 10.1016/j.wsif.2014.07.012.

Mehta, Pranjal H., Nicole M. Lawless DesJardins, Mark van Vugt, and Robert A. Josephs (2017). "Hormonal underpinnings of status conflict: Testosterone and cortisol are related to decisions and satisfaction in the hawk-dove game". In: *Hormones and Behavior* 92, pp. 141–154. DOI: 10.1016/j.yhbeh.2017.03.009.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4), pp. 417–473. DOI: 10.1111/j.1467-9868.2010.00740.x.

Mejova, Yelena, Ingmar Weber, and Michael W. Macy, eds. (2015). *Twitter: A digital socioscope*. Cambridge University Press. DOI: 10.1017/CBO9781316182635.

Mellon, Jonathan and Christopher Prosser (2017). "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users". In: *Research & Politics* 4 (3). DOI: 10.1177/2053168017720008.

Meyer, Robert and Michel Cukier (2006). "Assessing the attack threat due to IRC channels". In: *Proceedings of the International Conference on Dependable Systems and Networks*. DSN '06, pp. 467–472. DOI: 10.1109/DSN.2006.12.

Meyer, Robinson (2015). "Could a bank deny your loan based on your Facebook friends? A recent patent from the company judges your own creditworthiness by your friends". In: *The Atlantic* (25 September 2015). URL: https://www.theatlantic.com/technology/archive/2015/09/facebooks-new-patent-and-digital-redlining/407287/.

Miller, Claire Can (2015). "When algorithms discriminate". In: *The New York Times* (9 July 2015). URL: https://nyti.ms/2kelrzy.

Min, Jun-Ki, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong (2014). "Toss 'n' turn: Smartphone as sleep and sleep quality detector". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14, pp. 477–486. DOI: 10.1145/2556288.2557220.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Rosenquist (2011). "Understanding the demographics of Twitter users". In: *Proceedings of the Fifth International AAAI Conference* *on Weblogs and Social Media*. ICWSM-11, pp. 554–557. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816.

Mislove, Alan, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee (2007). "Measurement and analysis of online social networks". In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. IMC '07, pp. 29–42. DOI: 10.1145/1298306.1298311.

Mitchell, Lewis, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth (2013). "The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place". In: *PLOS ONE* 8 (5), e64417. DOI: 10.1371/journal.pone.0064417.

Mitchell, Tom M. (1997). *Machine learning*. McGraw-Hill.

Mollgaard, Anders, Ingo Zettler, Jesper Dammeyer, Mogens H. Jensen, Sune Lehmann, and Joachim Mathiesen (2016). "Measure of node similarity in multilayer networks". In: *PLOS ONE* 11 (6), pp. 1–10. DOI: 10.1371/journal.pone.0157436.

Mones, Enys, Arkadiusz Stopczynski, and Sune Lehmann (2017). "Contact activity and dynamics of the social core". In: *EPJ Data Science* 6 (1). DOI: 10.1140/epjds/s13688-017-0103-y.

Monge, Peter R. and Noshir Contractor (2003). *Theories of communication networks*. Oxford University Press.

Montasser, Omar and Daniel Kifer (2017). "Predicting demographics of high-resolution geographies with geotagged tweets". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI-17, pp. 1460–1466. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14818.

Montjoye, Yves-Alexandre de, Arkadiusz Stopczynski, Erez Shmueli, Alex Pentland, and Sune Lehmann (2014). "The strength of the strongest ties in collaborative problem solving". In: *Scientific Reports* 4 (5277). DOI: 10.1038/srep05277.

Moreno, Jacob L. (1934). *Who shall survive? A new approach to the problem of human interrelations*. Washington, D.C.: Nervous and Mental Disease Publishing Co.

Morozov, Evgeny (2013a). *To save everything, click here: The folly of technological solutionism*. New York: PublicAffairs.

– (2013b). "Your social networking credit score: 'Big data' can help determine who really deserves a loan. But there are dangers". Future Tense: Arizona State University, the New America Foundation, and Slate. URL: http://www.slate.com/articles/technology/future_tense/2013/01/wonga_lenddo_lendup_big_data_and_social_networking_banking.html.

Morstatter, Fred, Nichola Lubold, Heather Pon-Barry, Jürgen Pfeffer, and Huan Liu (2014). "Finding eyewitness tweets during crises". In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. ACL LACSS 2014. Baltimore, MD, USA: Association for Computational Linguistics, pp. 23–27. URL: http://www.aclweb.org/anthology/W14-2509.

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu (2014). "When is it biased? Assessing the representativeness of Twitter's Streaming API". In: *Companion to the Proceedings of the 23rd International Conference on World Wide Web*. WWW Companion '14, pp. 555–556. DOI: 10.1145/2567948.2576952.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen Carley (2013). *Is the sample good enough? Ccomparing data from Twitter's streaming API with Twitter's firehose*. URL: http://aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071/6379.

Moturu, Sai T., Inas Khayal, Nadav Aharony, Wei Pan, and Alex Pentland (2011). “Using social sensing to understand the links between sleep, mood, and sociability”. In: *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*. PASSAT/SocialCom 2011, pp. 208–214. DOI: 10.1109/PASSAT/SocialCom.2011.200.

Mowery, Jared (2016). “Twitter influenza surveillance: Quantifying seasonal misdiagnosis patterns”. In: *Online Journal of Public Health Informatics* 8 (3). DOI: 10.5210/ojphi.v8i3.7011.

Mullainathan, Sendhil and Jann Spiess (2017). “Machine learning: An applied econometric approach”. In: *Journal of Economic Perspectives* 31 (2), pp. 87–106. DOI: 10.1257/jep.31.2.87.

Muñoz, Cecilia, Megan Smith, and DJ Patil (2016). “Big data: A report on algorithmic systems, opportunity, and civil rights”. Executive Office of the President.

Myslín, Mark, Shu-Hong Zhu, Wendy Chapman, and Mike Conway (2013). “Using Twitter to examine smoking behavior and perceptions of emerging tobacco products”. In: *Journal of Medical Internet Research* 15 (8), e174. DOI: 10.2196/jmir.2534.

Nagar, Ruchit, Qingyu Yuan, C. Clark Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and S. John Brownstein (2014). “A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives”. In: *Journal of Medical Internet Research* 16 (10), e236. DOI: 10.2196/jmir.3416.

Newcomb, Theodore Mead (1961). *The acquaintance process*. New York, NY: Holt, Reinhard & Winston.

Newell, Edward, Stefan Dimitrov, Andrew Piper, and Derek Ruths (2016). “To buy or to read: How a platform shapes reviewing behavior”. In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 643–646. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13135/12818.

Newell, Edward, David Jurgens, Haji Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths (2016). “User migration in online social networks: A case study on Reddit during a period of community unrest”. In: *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. ICWSM-16, pp. 279–288. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13137/12729.

Nitins, Tanya and Jean Burgess (2014). “Twitter, brands, and user engagement”. In: *Twitter and society*. Ed. by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. New York, NY: Peter Lang, pp. 293–304. DOI: 10.3726/978-1-4539-1170-9.

Nordlie, Peter G. (1958). “A longitudinal study of interpersonal attraction in a natural group setting”. PhD thesis. University of Michigan.

O’Donnell, Matthew Brook and Emily B. Falk (2015). “Linking neuroimaging with functional linguistic analysis to understand processes of successful communication”. In: *Communication Methods and Measures* 9 (1–2), pp. 55–77. DOI: 10.1080/19312458.2014.999751.

Oktay, Hüseyin, Brian J. Taylor, and David D. Jensen (2010). “Causal discovery in social media using quasi-experimental designs”. In: *Proceedings of the First Workshop on Social Media Analytics*. SOMA ’10, pp. 1–9. DOI: 10.1145/1964858.1964859.

Olguín, Daniel Olguín, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland (2009). “Sensible organizations: technology and methodology for automatically measuring organizational behavior”. In: *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* 39 (1), pp. 43–55. DOI: 10.1109/TSMCB.2008.2006638.

Oloritun, Rahman O., Anmol Madan, Alex Pentland, and Inas Khayal (2012). "Evolution of social encounters in ad-hoc mobile face-to-face interaction networks". In: *Proceedings of the 2012 International Conference on Social Informatics*, pp. 192–198. DOI: 10.1109/SocialInformatics.2012.101.

– (2013). "Identifying close friendships in a sensed social network". In: *Procedia - Social and Behavioral Sciences* 79, pp. 18–26. DOI: https://doi.org/10.1016/j.sbspro.2013.05.054.

O'Loughlin, John (2000). "Can King's ecological inference method answer a social scientific puzzle: Who voted for the Nazi Party in Weimar Germany?" In: *Annals of the Association of American Geographers* 90 (3), pp. 592–601. DOI: 10.1111/0004-5608.00213.

O'Neil, Cathy (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown.

Ong, Jonathan Corpus (2015). "Witnessing distant and proximal suffering within a zone of danger: Lay moralities of media audiences in the Philippines". In: *International Communication Gazette* 77 (7), pp. 607–621. DOI: 10.1177/1748048515601555.

Ozer, Nicole (2016). "Police use of social media surveillance software is escalating, and activists are in the digital crosshairs". In: *@ACLU_NorCal* (23 September 2016). URL: https://medium.com/@ACLU_NorCal/police-use-of-social-media-surveillance-software-is-escalating-and-activists-are-in-the-digital-d29d8f89c48.

Pablo, Aragón, Gómez Vicenç, and Kaltenbrunner Andreas (2017). "Detecting platform effects in online discussions". In: *Policy & Internet* 9 (4), pp. 420–443. DOI: 10.1002/poi3.158.

Pachucki, Mark C., Emily J. Ozer, Alain Barrat, and Ciro Cattuto (2015). "Mental health and social networks in early adolescence: A dynamic study of objectively-measured social interaction behaviors". In: *Social Science & Medicine* 125, pp. 40–50. DOI: 10.1016/j.socscimed.2014.04.015.

Palaghias, Niklas, Seyed Amir Hoseinitabatabaei, Michele Nati, Alexander Gluhak, and Klaus Moessner (2016). "A survey on mobile social signal processing". In: *ACM Computing Surveys* 48 (4), 57:1–57:52. DOI: 10.1145/2893487.

Pan, Wei, Nadav Aharony, and Alex Pentland (2011). "Fortune monitor or fortune teller: Understanding the connection between interaction patterns and financial status". In: *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*. PASSAT/SocialCom 2011, pp. 200–207. DOI: 10.1109/PASSAT/SocialCom.2011.163.

Panger, Galen (2016). "Reassessing the Facebook experiment: critical thinking about the validity of Big Data research". In: *Information, Communication & Society* 19 (8), pp. 1108–1126. DOI: 10.1080/1369118X.2015.1093525.

Panisson, André, Laetitia Gauvin, Alain Barrat, and Ciro Cattuto (2013). "Fingerprinting temporal networks of close-range human proximity". In: *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshop on the Impact of Human Mobility in Pervasive Systems and Applications*. PERCOM '13 Workshops, pp. 261–266. DOI: 10.1109/PerComW.2013.6529492.

Paradiso, Joseph A., Jonathan Gips, Mathew Laibowitz, Sajid Sadi, David Merrill, Ryan Aylward, Pattie Maes, and Alex Pentland (2010). "Identifying and facilitating social interaction with a wearable wireless sensor network". In: *Personal and Ubiquitous Computing* 14 (2), pp. 137–152. DOI: 10.1007/s00779-009-0239-2.

Parkinson, Carolyn, Adam M. Kleinbaum, and Thalia Wheatley (2018). "Similar neural responses predict friendship". In: *Nature Communications* 9 (332). DOI: 10.1038/s41467-017-02722-7.

Pasquale, Frank (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Patel, Shwetak N., Julie A. Kientz, Gillian R. Hayes, Sooraj Bhat, and Gregory D. Abowd (2006). "Farther than you may think: An empirical investigation of the proximity of users to their mobile phones". In: *Proceedings of the 8th International Conference on Ubiquitous Computing*. Ubicomp 2006, pp. 123–140. DOI: 10.1007/11853565_8.

Patton, Michael Quinn (2015). *Qualitative research & evaluation methods: Integrating theory and practice*. 4th ed. Thousand Oaks, CA: SAGE.

Paul, Michael J. and Mark Dredze (2011). "You are what you tweet: Analyzing twitter for public health". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*. ICWSM-11, pp. 265–272. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880.

Pentland, Alex (2009). "Reality Mining of mobile communications: Toward a New Deal on Data". In: *The Global Information Technology Report 2008-2009: Mobility in a networked world*. Ed. by Soumitra Dutta and Irene Mia. World Economic Forum, pp. 75–80.

– (2014). *Social physics: How good ideas spread—the lessons from a new science*. The Penguin Press.

Pfeffer, Jürgen and Momin M. Malik (2017). "Simulating the dynamics of socio-economic systems". In: *Networked governance: New research perspectives*. Ed. by Betina Hollstein, Wenzel Matiaske, and Kai-Uwe Schnapp. Cham, Switzerland: Springer International Publishing, pp. 143–161. DOI: 10.1007/978-3-319-50386-8_9.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes (2011). "Do all birds tweet the same? Characterizing Twitter around the world". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM '11, pp. 1025–1030. DOI: 10.1145/2063576.2063724.

Porter, Jack and Ping Yu (2015). "Regression discontinuity designs with unknown discontinuity points: Testing and estimation". In: *Journal of Econometrics* 189 (1), pp. 132–147. DOI: 10.1016/j.jeconom.2015.06.002.

Prell, Christina (2011). *Social network analysis: History, theory and methodology*. London: SAGE.

Primack, Brian A., Mary V. Carroll, Patricia M. Weiss, Alan L. Shihadeh, Ariel Shensa, Steven T. Farley, Michael J. Fine, Thomas Eissenberg, and Smita Nayak (2016). "Systematic review and meta-analysis of inhaled toxicants from waterpipe and cigarette smoking". In: *Public Health Reports* 131 (1), pp. 76–85. DOI: 10.1177/003335491613100114.

Quercia, Daniele, Licia Capra, and Jon Crowcroft (2012). "The social world of Twitter: Topics, geography, and emotions". In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ICWSM-12, pp. 298–305. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4612.

Raacke, John and Jennifer Bonds-Raacke (2008). "MySpace and Facebook: Applying the uses and gratifications theory to exploring friend-networking sites". In: *CyberPsychology & Behavior* 11 (2), pp. 169–174. DOI: 10.1089/cpb.2007.0056.

Rachuri, Kiran K., Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas (2010). "EmotionSense: A mobile phones based adaptive platform for experimental social psychology research". In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. UbiComp '10, pp. 281–290. DOI: 10.1145/1864349.1864393.

Racine, Jeff (2000). “Consistent cross-validatory model-selection for dependent data: *hv*-block cross-validation”. In: *Journal of Econometrics* 99 (1), pp. 39–61. DOI: 10.1016/S0304-4076(00)00030-0.

Rahman, Md. Mahbubur, Amin Ahsan Ali, Kurt Plarre, Mustafa al’Absi, Emre Ertin, and Santosh Kumar (2011). “mConverse: Inferring conversation episodes from respiratory measurements collected in the field”. In: *Proceedings of the 2nd Conference on Wireless Health*. WH ’11, 10:1–10:10. DOI: 10.1145/2077546.2077557.

Ramsay, Stephen (2011). *Reading machines: Toward an algorithmic criticism*. Urbana, IL: University of Illinois Press.

Ranganath, Rajesh, Dan Jurafsky, and Daniel A. McFarland (2009). “It’s not you, it’s me: Detecting flirting and its misperception in speed-dates”. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. EMNLP ’09, pp. 334–342. URL: http://dl.acm.org/citation.cfm?id=1699510.1699554.

– (2013). “Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates”. In: *Computer Speech & Language* 27 (1), pp. 89–115. DOI: 10.1016/j.csl.2012.01.005.

Rasmussen, Carl Edward and Christopher K. I. Williams (2005). *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press.

Ren, Yuqing, Robert E. Kraut, and Sara Kiesler (2007). “Applying common identity and bond theory to design of online communities”. In: *Organization Studies* 28 (3), pp. 377–408. DOI: 10.1177/0170840607076007.

Ripley, Ruth M., Tom A. B. Snijders, Zsófia Boda, András Vörös, and Paulina Preciado (2017). “Manual for RSiena”. University of Oxford: Department of Statistics; Nuffield College, and University of Groningen: Department of Sociology. URL: http://www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf.

Robbins, Martin (2016). “Has a rampaging AI algorithm really killed thousands in Pakistan? A killer machine-learning algorithm guiding the U.S. drone program has killed thousands of innocent people according to some reports. What’s the truth?” In: *The Guardian* (16 February, 2016). URL: https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan.

Roberts, S., M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain (2012). “Gaussian processes for time-series modelling”. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371 (1984). DOI: 10.1098/rsta.2011.0550.

Robins, Gary (2015). *Doing social network research: Network-based research design for social scientists*. SAGE.

Roggero, M. (2012). “Discontinuity detection and removal from data time series”. In: *VII Hotine-Marussi Symposium on Mathematical Geodesy*. Ed. by Nico Sneeuw, Pavel Novák, Mattia Crespi, and Fernando Sansò. International Association of Geodesy Symposia. Berlin, Heidelberg: Springer, pp. 135–140. DOI: 10.1007/978-3-642-22078-4_20.

Romero, Daniel, Brendan Meeder, Vladimir Barash, and Jon Kleinberg (2011). “Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness”. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM-11, pp. 606–609. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2860/3253.

Rose, Shyanika W., Catherine L. Jo, Steven Binns, Melissa Buenger, Sherry Emery, and Kurt M. Ribisl (2017). “Perceptions of menthol cigarettes among Twitter users: Content and sentiment analysis”. In: *Journal of Medical Internet Research* 19 (2), e56. DOI: 10.2196/jmir.5694.

Ruths, Derek and Jürgen Pfeffer (2014). “Social media for large studies of behavior”. In: *Science* 346 (6213), pp. 1063–1064. DOI: 10.1126/science.346.6213.1063.

Sadilek, Adam, Henry Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio (2016). “Deploying nEmesis: Preventing foodborne illness by data mining social media”. In: *Proceedings of the Twenty-Eighth Innovative Applications of Artificial Intelligence Conference*. IAAI-16, pp. 3982–3989. URL: https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/11823.

Sadilek, Adam, Henry Kautz, and Vincent Silenzio (2012). “Predicting disease transmission from geo-tagged micro-blog data”. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. AAAI-12, pp. 136–142. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4844.

Saeb, Sohrab, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording (2016). “Voodoo machine learning for clinical predictions”. eprint: http://biorxiv.org/content/early/2016/06/19/059774.full.pdf.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). “Earthquake shakes Twitter users: Real-time event detection by social sensors”. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW ’10, pp. 851–860. DOI: 10.1145/1772690.1772777.

Salah, Albert Ali, Bruno Lepri, Fabio Pianesi, and Alex Pentland (2011). “Human behavior understanding for inducing behavioral change: Application perspectives”. In: *Proceedings of the International Workshop on Human Behavior Understanding*. HBU 2011, pp. 1–15. DOI: 10.1007/978-3-642-25446-8_1.

Salant, Priscilla and Don A. Dillman (1994). *How to conduct your own survey*. New York, NY: John Wiley & Sons, Inc.

Sampson, Samuel Franklin (1968). “A novitiate in a period of change: An experimental and case study of social relationships”. PhD thesis. Cornell University.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort (2014). “Auditing algorithms: Research methods for detecting discrimination on Internet platforms”. Paper presented to ‘Data and Discrimination: Converting Critical Concerns into Productive Inquiry,’ a preconference at the 64th Annual Meeting of the International Communication Association, May 22, 2014, Seattle, WA, USA. URL: http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Sarukkai, Sundar (2005). “Revisiting the ‘unreasonable effectiveness’ of mathematics”. In: *Current Science* 88 (3), pp. 415–423. URL: http://www.jstor.org/stable/24110208.

Savage, Mike and Roger Burrows (2007). “The coming crisis of empirical sociology”. In: *Sociology* 41 (5), pp. 885–899. DOI: 10.1177/0038038507080443.

– (2009). “Some further reflections on the coming crisis of empirical sociology”. In: *Sociology* 43 (4), pp. 762–772. DOI: 10.1177/0038038509105420.

Schneider, Fabian, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger (2009). “Understanding online social network usage from a network perspective”. In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*. IMC ’09, pp. 35–48. DOI: 10.1145/1644893.1644899.

Schoenebeck, Grant (2013). “Potential networks, contagious communities, and understanding social network structure”. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW ’13, pp. 1123–1132. DOI: 10.1145/2488388.2488486.

Scott, John (2012). *Social network analysis*. London: SAGE.

Scott, John and Peter J. Carrington, eds. (2011). *The SAGE handbook of social network analysis*. SAGE. DOI: 10.4135/9781446294413.

Sekara, Vedran and Sune Lehmann (2014). "The strength of friendship ties in proximity sensor data". In: *PLOS ONE* 9 (7), pp. 1–8. DOI: 10.1371/journal.pone.0100915.

Sekara, Vedran, Arkadiusz Stopczynski, and Sune Lehmann (2016). "Fundamental structures of dynamic social networks". In: *Proceedings of the National Academy of Sciences* 113 (36), pp. 9977–9982. DOI: 10.1073/pnas.1602803113.

Selivanov, Dmitriy and Qing Wang (2018). *text2vec: Modern text mining framework for R*. R package version 0.5.1. URL: https://CRAN.R-project.org/package=text2vec.

Shah, Dhavan V., Joseph N. Cappella, and W. Russell Neuman (2015). "Big data, digital media, and computational social science: Possibilities and perils". In: *The ANNALS of the American Academy of Political and Social Science* 659 (1), pp. 6–13. DOI: 10.1177/0002716215572084.

Shalizi, Cosma Rohilla and Andrew C. Thomas (2011). "Homophily and contagion are generically confounded in observational social network studies". In: *Sociological Methods & Research* 40 (2), pp. 211–239. DOI: 10.1177/0049124111404820.

Sharma, Amit, Jake M. Hofman, and Duncan J. Watts (2015). "Estimating the causal impact of recommendation systems from observational data". In: *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. EC '15, pp. 453–470. DOI: 10.1145/2764468.2764488.

Sharma, Sanjay (2013). "Black Twitter? Racial hashtags, networks and contagion". In: *new formations: a journal of culture/theory/politics* 78 (1). URL: http://muse.jhu.edu/journals/new_formations/v078/78.sharma.html.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook (2014). "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'". In: *Geoforum* 52 (0), pp. 167–179. DOI: 10.1016/j.geoforum.2014.01.006.

Shmueli, Galit (2010). "To explain or to predict?" In: *Statistical Science* 25 (3), pp. 289–310. DOI: 10.1214/10-STS330.

Shumway, Robert H. and David S. Stoffer (2011). *Time series analysis and its applications with R examples*. 3rd ed. Springer Texts in Statistics. New York: Springer. DOI: 10.1007/978-1-4419-7865-3.

Sidani, Jaime E., Ariel Shensa, Saul Shiffman, Galen E. Switzer, and Brian A. Primack (2015). "Behavioral associations with waterpipe tobacco smoking dependence among US young adults". In: *Addiction* 111 (2), pp. 351–359. DOI: 10.1111/add.13163.

Siegel, Michael and Lois Biener (2010). "Evaluating the impact of statewide anti-tobacco campaigns: The Massachusetts and California tobacco control programs". In: *Journal of Social Issues* 53 (1), pp. 147–168. DOI: 10.1111/j.1540-4560.1997.tb02436.x.

Singh, Tushar, René A. Arrazola, Catherine G. Corey, Corinne G. Husten, Linda J. Neff, David M. Homa, and Brian A. King (2016). "Tobacco use among middle and high school students — United States, 2011–2015". In: *Morbidity and Mortality Weekly Report (MMWR)* 65 (14), pp. 361–367. DOI: 10.15585/mmwr.mm6514a1.

Sloan, Luke, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana (2013). "Knowing the tweeters: Deriving sociologically relevant demographics from Twitter". In: *Sociological Research Online* (18). DOI: 10.5153/sro.3001.

Smieszek, Timo, Stefanie Castell, Alain Barrat, Ciro Cattuto, Peter J. White, and Gérard Krause (2016). "Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes". In: *BMC Infectious Diseases* 16 (1), 341:1–341:14. DOI: 10.1186/s12879-016-1676-y.

Snijders, Tom A. B., Gerhard G. van de Bunt, and Christian E. G. Steglich (2010). "Introduction to stochastic actor-based models for network dynamics". In: *Social Networks* 32 (1), pp. 44–60. DOI: 10.1016/j.socnet.2009.02.004.

Snijders, Tom A. B. and Mark Pickup (2016). "Stochastic actor oriented models for network dynamics". In: *The Oxford handbook of political networks*. Ed. by Jennifer Nicoll Victor, Alexander H. Montgomery, and Mark Lubell, pp. 221–248. DOI: 10.1093/oxfordhb/9780190228217.013.10.

Spirtes, Peter (2010). "Introduction to causal inference". In: *Journal of Machine Learning Research* 11, pp. 1643–1662. URL: http://www.jmlr.org/papers/v11/spirtes10a.html.

Spirtes, Peter, Clark Glymour, and Richard Scheines (2001). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.

Spirtes, Peter and Kun Zhang (2016). "Causal discovery and inference: Concepts and recent methodological advances". In: *Applied Informatics* 3 (3), pp. 1–28. DOI: 10.1186/s40535-016-0018-x.

Stadtfeld, Christoph and Per Block (2017). "Interactions, actors, and time: Dynamic network actor models for relational events". In: *Sociological Science* 4 (14), pp. 318–352. DOI: 10.15195/v4.a14.

Stadtfeld, Christoph, James Hollway, and Per Block (2017). "Dynamic network actor models: Investigating coordination ties through time". In: *Sociological Methodology* 47 (1), pp. 1–40. DOI: 10.1177/0081175017709295.

Staiano, Jacopo, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland (2012). "Friends don't lie: Inferring personality traits from social network structure". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. UbiComp '12, pp. 321–330. DOI: 10.1145/2370216.2370266.

Starnini, Michele, Andrea Baronchelli, and Romualdo Pastor-Satorras (2016). "Model reproduces individual, group and collective dynamics of human contact networks". In: *Social Networks* 47, pp. 130–137. DOI: 10.1016/j.socnet.2016.06.002.

Starnini, Michele, Anna Machens, Ciro Cattuto, Alain Barrat, and Romualdo Pastor-Satorras (2013). "Immunization strategies for epidemic processes in time-varying contact networks". In: *Journal of Theoretical Biology* 337, pp. 89–100. DOI: 10.1016/j.jtbi.2013.07.004.

Steglich, Christian E. G., Tom A. B. Snijders, and Michael Pearson (2010). "Dynamics networks and behavior: Separating selection from influence". In: *Sociological Methodology* 40 (1), pp. 329–393. DOI: 10.1111/j.1467-9531.2010.01225.x.

Stehlé, Juliette, François Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat (2013). "Gender homophily from spatial behavior in a primary school: A sociometric study". In: *Social Networks* 35 (4), pp. 604–613. DOI: 10.1016/j.socnet.2013.08.003.

Stehlé, Juliette, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Corinne Régis, Jean-François Pinton, Nagham Khanafer, Wouter Van den Broeck, and Philippe Vanhems (2011). "Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees". In: *BMC Medicine* 9 (87), pp. 1–15. DOI: 10.1186/1741-7015-9-87.

Stehlé, Juliette, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina, and Philippe Vanhems (2011). "High-resolution measurements of face-to-face contact patterns in a primary school". In: *PLOS ONE* 6 (8), pp. 1–13. DOI: 10.1371/journal.pone.0023176.

Stephens, Monica and Ate Poorthuis (2014). “Follow thy neighbor: Connecting the social and the spatial networks on Twitter”. In: *Computers, Environment and Urban Systems* 53, pp. 87–95. DOI: 10.1016/j.compenvurbsys.2014.07.002.

Stevenson, Amanda Jean (2014). “Finding the Twitter users who stood with Wendy”. In: *Contraception* (90), pp. 502–507. DOI: 10.1016/j.contraception.2014.07.007.

Stieglitz, Stefan and Nina Krüger (2014). “Public enterprise-related communication and its impact on social media issue management”. In: *Twitter and society*. Ed. by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. Digital Formations. New York: Peter Lang, pp. 281–292. DOI: 10.3726/978-1-4539-1170-9.

Stinchcombe, Arthur L. (2001). *When formality works: Authority and abstraction in law and organizations*. Chicago, IL: University of Chicago Press.

Stopczynski, Arkadiusz, Piotr Sapiezynski, Alex Pentland, and Sune Lehmann (2015). “Temporal fidelity in dynamic social networks”. In: *The European Physical Journal B* 88 (10), p. 249. DOI: 10.1140/epjb/e2015-60549-7.

Stopczynski, Arkadiusz, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann (2014). “Measuring large-scale social networks with high resolution”. In: *PLOS ONE* 9 (4), pp. 1–24. DOI: 10.1371/journal.pone.0095978.

Su, Jessica, Aneesh Sharma, and Sharad Goel (2016). “The effect of recommendations on network structure”. In: *Proceedings of the 25th International Conference on World Wide Web*. WWW ’16, pp. 1157–1167. DOI: 10.1145/2872427.2883040.

Sylvester, Jared, John Healey, Chen Wang, and William M. Rand (2014). “Space, time, and hurricanes: Investigating the spatiotemporal relationship among social media use, donations, and disasters”. Research Paper No. RHS 2441314, Robert H. Smith School. DOI: 10.2139/ssrn.2441314.

Szekely, Francisc (2011). “Unreliable observers, flawed instruments, ‘disciplined viewings’: Handling specimens in early modern microscopy”. In: *Parergon* 28 (1), pp. 155–176. DOI: 10.1353/pgn.2011.0032.

Szomszor, Martin, Patty Kostkova, Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, and Harith Alani (2011). “Providing enhanced social interaction services for industry exhibitors at large medical conferences”. In: *Proceedings of the 3rd International Conference on Developments in eSystems Engineering*. DeSE 2010, pp. 42–45. DOI: 10.1109/DeSE.2011.113.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman (2012). “Geography of Twitter networks”. In: *Social Networks* 34 (1), pp. 73–81. DOI: 10.1016/j.socnet.2011.05.006.

Taljaard, Monica, Joanne E. McKenzie, Craig R. Ramsay, and Jeremy M. Grimshaw (2014). “The use of segmented regression in analysing interrupted time series studies: An example in pre-hospital ambulance care”. In: *Implementation Science* 9 (77). DOI: 10.1186/1748-5908-9-77.

Tasse, Dan, Zichen Liu, Alex Sciuto, and Jason I. Hong (2017). “State of the geotags: Motivations and recent changes”. In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. ICWSM 2017, pp. 250–259. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15588.

Taylor, James W. and Derek W. Bunn (1999). “A quantile regression approach to generating prediction intervals”. In: *Management Science* 45 (2), pp. 225–237. DOI: 10.1287/mnsc.45.2.225.

Thackeray, Rosemary, Brad L. Neiger, and Heidi Keller (2012). “Integrating social media and social marketing: A four-step process”. In: *Health Promotion Practice* 13 (2), pp. 165–168. DOI: 10.1177/1524839911432009.

Thebault-Spieker, Jacob, Loren Terveen, and Brent Hecht (2017). “Toward a geographic understanding of the sharing economy: Systemic biases in UberX and TaskRabbit”. In: *ACM Transactions on Computer-Human Interaction* 24 (3), 21:1–21:40. DOI: 10.1145/3058499.

Therneau, Terry and Beth Atkinson (2018). *rpart: Recursive partitioning and regression trees*. R package version 4.1-13. URL: https://CRAN.R-project.org/package=rpart.

Thomas, Kurt, Chris Grier, Dawn Song, and Vern Paxson (2011). “Suspended accounts in retrospect: An analysis of Twitter spam”. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. IMC ’11, pp. 243–258. DOI: 10.1145/2068816.2068840.

Thomas, Kurt, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson (2013). “Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse”. In: *Proceedings of the 22nd USENIX Security Symposium*. USENIX Security ’13, pp. 195–210. URL: https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/thomas.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1), pp. 267–288. URL: http://www.jstor.org/stable/2346178.

– (2011). “Regression shrinkage and selection via the lasso: A retrospective”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (3), pp. 273–282. DOI: 10.1111/j.1467-9868.2011.00771.x.

Tufekci, Zeynep (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*. ICWSM-14, pp. 505–514. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

– (2015). “Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency”. In: *Colorado Technology Law Journal* 13, pp. 203–218.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow (2011). “The anatomy of the Facebook social graph”. eprint: https://arxiv.org/abs/1111.4503.

U.S. Census Bureau (2008). *A compass for understanding and using American Community Survey data: What general data users need to know*. Washington, DC: U.S. Government Printing Office. URL: https://www.census.gov/library/publications/2008/acs/general.html.

Van den Broeck, Wouter, Ciro Cattuto, Alain Barrat, Martin Szomszor, Gianluca Correndo, and Harith Alani (2010). “The Live Social Semantics application: A platform for integrating face-to-face presence with on-line social networking”. In: *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications Workshops, First International Workshop on Communication, Collaboration and Social Networking*. PERCOM ’10 Workshops, pp. 226–231. DOI: 10.1109/PERCOMW.2010.5470665.

Van den Broeck, Wouter, Marco Quaggiotto, Lorenzo Isella, Alain Barrat, and Ciro Cattuto (2012). “The making of sixty-nine days of close encounters at the science gallery”. In: *Leonardo* 45 (3), pp. 285–285. DOI: 10.1162/LEON_a_00377.

van den Heerik, Romy A. M., Charlotte M. J. van Hooijdonk, Christian Burgers, and Gerard J. Steen (2017). “‘Smoking is sóóó... sandals and white socks’: Co-creation of a Dutch anti-smoking campaign to change social norms”. In: *Health Communication* 32 (5), pp. 621–628. DOI: 10.1080/10410236.2016.1168000.

van Dijck, José (2013). *The culture of connectivity: A critical history of social media*. New York, NY: Oxford University Press.

Vanhems, Philippe, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin (2013). "Estimating potential infection transmission routes in hospital wards using wearable proximity sensors". In: *PLOS ONE* 8 (9), pp. 1–9. DOI: 10.1371/journal.pone.0073970.

Veale, Michael, Max van Kleek, and Reuben Binns (2018). "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18, 440:1–440:14. DOI: 10.1145/3173574.3174014.

Viswanath, Bimal, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi (2009). "On the evolution of user interaction in Facebook". In: *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks*. WOSN '09, pp. 37–42. DOI: 10.1145/1592665.1592675.

Voirin, Nicolas, Céécile Payet, Alain Barrat, Ciro Cattuto, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, Jean-Sébastien Casalegno, Bruno Lina, and Philippe Vanhems (2015). "Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital". In: *Infection Control & Hospital Epidemiology* 36 (3), pp. 254–260. DOI: 10.1017/ice.2014.53.

Wagner, A. K., S. B. Soumerai, F. Zhang, and D. Ross-Degnan (2002). "Segmented regression analysis of interrupted time series studies in medication use research". In: *Journal of Clinical Pharmacy and Therapeutics* 27 (4), pp. 299–309. DOI: 10.1046/j.1365-2710.2002.00430.x.

Wallach, Hanna (2018). "Computational social science ≠ computer science + social data". In: *Communications of the ACM* 61 (3), pp. 42–44. DOI: 10.1145/3132698.

Wang, Yi-Chia, Moira Burke, and Robert E. Kraut (2013). "Gender, topic, and audience response: An analysis of user-generated content on Facebook". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13, pp. 31–34. DOI: 10.1145/2470654.2470659.

Wang, Di, Ahmad Al-Rubaie, John Davies, and Sandra Stinĉić Clarke (2014). "Real time road traffic monitoring alert based on incremental learning from tweets". In: *Proceedings of the 2014 IEEE Symposium on Evolving and Autonomous Learning Systems*. EALS '14, pp. 50–57. DOI: 10.1109/EALS.2014.7009503.

Wang, Haizhou and Mingzhou Song (2011). "Ckmeans.1d.dp: Optimal *k*-means clustering in one dimension by dynamic programming". In: *The R Journal* 3 (2), pp. 29–33. URL: https://journal.r-project.org/archive/2011-2/RJournal_2011-2_Wang+Song.pdf.

Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell (2014). "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '14, pp. 3–14. DOI: 10.1145/2632048.2632054.

Wang, Tricia (2013). "Big Data needs Thick Data". Ethnography Matters blog (13 May 2013). URL: http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/.

Wasserman, Larry (2014). "Rise of the machines". In: *Past, Present and Future of Statistical Science*. Ed. by Xihong Lin, Christian Genest, David Banks, Geert Molenberghs, David Scott, and Jane?Ling Wang. Boca Raton, FL: CRC Press, pp. 525–536.

Wasserman, Stanley (2013). "Comments on 'Social contagion theory: examining dynamic social networks and human behavior' by Nicholas Christakis and James Fowler". In: *Statistics in Medicine* 32 (4), pp. 578–580. DOI: 10.1002/sim.5483.

Watts, Duncan J. (2014). "Stop complaining about the Facebook study. It's a golden age for research". In: *The Guardian* (7 July 2014). URL: https://www.theguardian.com/commentisfree/2014/jul/07/facebook-study-science-experiment-research.

Webber, Richard (2009). "Response to 'The coming crisis of empirical sociology': An outline of the research potential of administrative and transactional data". In: *Sociology* 43 (1), pp. 169–178. DOI: 10.1177/0038038508099104.

Weller, Katrin, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, eds. (2014). *Twitter and society*. New York: Peter Lang. DOI: 10.3726/978-1-4539-1170-9.

Wiese, Jason, Jun-Ki Min, Jason I. Hong, and John Zimmerman (2014). *Assessing call and SMS logs as an indication of tie strength*. Tech. rep. CMU-HCII-14-101. Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University. URL: reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-14-101.pdf.

– (2015). "'You never call, you never write': Call and SMS logs do not always indicate tie strength". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15, pp. 765–774. DOI: 10.1145/2675133.2675143.

Wigner, Eugene P. (1960). "The unreasonable effectiveness of mathematics in the natural sciences". In: *Communications on Pure and Applied Mathematics* 13, pp. 1–14. DOI: 10.1002/cpa.3160130102.

Wilson, Christo, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao (2012). "Beyond social graphs: User interactions in online social networks and their implications". In: *ACM Transactions on the Web* 6 (4), 17:1–17:31. DOI: 10.1145/2382616.2382620.

Wright, Marvin N. and Andreas Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77 (1), pp. 1–17. DOI: 10.18637/jss.v077.i01.

Wu, Shaohua, T. J. Harris, and K. B. McAuley (2007). "The use of simplified or misspecified models: Linear case". In: *The Canadian Journal of Chemical Engineering* 85 (4), pp. 386–398. DOI: 10.1002/cjce.5450850401.

Wyatt, Danny, Tanzeem Choudhury, Jeff Bilmes, and James A. Kitts (2011). "Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science". In: *ACM Transactions on Intelligent System Technologies* 2 (1), 7:1–7:41. DOI: 10.1145/1889681.1889688.

Yan, Zhixian, Jun Yang, and Emmanuel Munguia Tapia (2013). "Smartphone Bluetooth based social sensing". In: *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*. UbiComp '13 Adjunct, pp. 95–98. DOI: 10.1145/2494091.2494118.

Yang, Wenjing and Yuhong Yang (2016). "Toward an objective and reproducible model choice via variable selection deviation". In: *Biometrics* 73 (1), pp. 20–30. DOI: 10.1111/biom.12554.

Yeung, Karen (2017). "'Hypernudge': Big Data as a mode of regulation by design". In: *Information, Communication & Society* 20 (1), pp. 118–136. DOI: 10.1080/1369118X.2016.1186713.

Yoo, Woohyun, JungHwan Yang, and Eunji Cho (2016). "How social media influence college students' smoking attitudes and intentions". In: *Computers in Human Behavior* 64, pp. 173–182. DOI: 10.1016/j.chb.2016.06.061.

Yuan, Nicholas Jing, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie (2013). “We know how you live: Exploring the spectrum of urban lifestyles”. In: *Proceedings of the First ACM Conference on Online Social Networks*. COSN ’13, pp. 3–14. DOI: 10.1145/2512938.2512945.

Zachary, Wayne W. (1977). “An information flow model for conflict and fission in small groups”. In: *Journal of Anthropological Research* 33 (4), pp. 452–473. URL: http://www.jstor.org/stable/3629752.

Zagheni, Emilio and Ingmar Weber (2015). “Demographic research with non-representative internet data”. In: *International Journal of Manpower* 36 (1), pp. 13–25. DOI: 10.1108/IJM-12-2014-0261.

Zhao, Mingmin, Fadel Adib, and Dina Katabi (2016). “Emotion recognition using wireless signals”. In: *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. MobiCom ’16, pp. 95–108. DOI: 10.1145/2973750.2973762.

Zhao, Peng and Bin Yu (2006). “On model selection consistency of lasso”. In: *Journal of Machine Learning Research* 7 (Nov), pp. 2541–2563.

Zickuhr, Kathryn (2012). “Geosocial services”. Pew Internet and American Life Project, Pew Research Center. URL: http://www.pewinternet.org/2012/05/11/geosocial-services/.

– (2013). “Location-based services”. Pew Internet and American Life Project, Pew Research Center. URL: http://www.pewinternet.org/2013/09/12/location-based-services/.

Ziewitz, Malte (2016). “Governing algorithms: Myth, mess, methods”. In: *Science, Technology, & Human Values* 41 (1), pp. 3–16. DOI: 10.1177/0162243915608948.

Zignani, Matteo, Sabrina Gaito, Gian Paolo Rossi, Xiaohan Zhao, Haitao Zheng, and Ben Zhao (2014). “Link and triadic closure delay: Temporal metrics for social network dynamics”. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. ICWSM-14, pp. 564–573. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8042.