Measuring Public Opinion with Social Media Data

Marko Klašnja
Georgetown University
marko.klasnja@georgetown.edu

Pablo Barberá
University of Southern California
pbarbera@usc.edu

Nicholas Beauchamp
Northeastern University
n.beauchamp@neu.edu

Jonathan Nagler
New York University
jonathan.nagler@nyu.edu

Joshua Tucker
New York University
joshua.tucker@nyu.edu

September 30, 2016

1 Social media and public opinion: opportunities and challenges

Social media sites such as Facebook or Twitter are playing an increasingly central role in politics. As Kreiss (2014) shows, the 2012 Obama and Romney presidential election campaigns relied heavily on social media to appeal to their supporters and influence the agendas and frames of citizens and journalists. In 2016, the role of social media has only accelerated, with Twitter, for instance, becoming a central pillar of the Trump campaign. Social media sites have also been essential for disseminating information and organizing during many recent episodes of mass protest, from the pro-democracy revolutions during the Arab Spring to Euromaidan to the recent wave of pro-civil rights demonstrations in the United States (see e.g. Tufekci and Wilson, 2012; Tucker et al., 2016). The influence of social media has also become pervasive in traditional news outlets: Twitter is commonly used as a source of information about breaking news events: journalists and traditional media often solicit feedback from their viewers through social media, and political actors can rely on social media rather than press releases to reach the public. Most fundamentally, for numerous political organizations and millions of users, social media has become the primary means of acquiring, sharing, and discussing political information (Kwak et al., 2010; Neuman et al., 2014).

In this chapter we examine to what extent one can aggregate political messages published on social networking sites to obtain a measure of public opinion that is comparable or better than those obtained through surveys. It is well known that public opinion surveys are facing growing difficulties in reaching and persuading reluctant respondents (De Leeuw and De Heer, 2002). According to the Pew Research Center, the typical contact rates have dropped from 90% to 62% from 1997 to 2012, with response rates dropping from about 40% to 9% (Pew Research Center, 2012).[1] One important reason behind these trends is the falling rate of landline phone use, coupled with the fact that federal regulations prohibit the use of automated dialers for all unsolicited calls to cell phones (but not landline phones). According to one estimate, the share of cell phone-only households in the U.S. has grown by 70% in four years to reach 44% of all households in 2014.[2] While the relationship between non-response rates and non-response bias—arising when those who answer are different from those who do not—is complex (Groves, 2006; Groves and Peytcheva, 2008), survey responders tend to be more likely to vote, contact a public official, or volunteer than the survey non-responders (e.g. Pew Research Center, 2012). The responders' answers tend to exhibit less measurement error and lower social desirability bias (Abraham, Helms and Presser, 2009; Tourangeau, Groves and Redline, 2010). The cell phone-only respondents can differ in political preferences than those with landline phones; for example, they were significantly more likely to support Obama in 2008, especially among *older* voters (Mokrzycki, Keeter and Kennedy, 2009).

[1] While the response rates for the "gold standard" surveys such as the General Social Survey, the American National Election Study, or the National Household Education Survey are higher, they too have been falling markedly (Brick and Williams, 2013; Hillygus, 2011).

[2] See for example: `http://www.businesswire.com/news/home/20150402005790/en#.VR2B1JOPoyS`.

These trends have raised questions about the reliability and precision of representative surveys, and increased the costs of fielding high-quality polls, at the same time as funding available for a number of established large-scale surveys has been threatened.[3]

These factors are increasing the incentives for using social media to measure public opinion. First and foremost, social media provides an opportunity for us to examine the opinions of the public without any prompting or framing effects from analysts. Rather than measure what someone thinks about politics in the artificial environments of a front porch, dinnertime phone-call, or survey webpage, we can observe how people spontaneously speak about politics in the course of their daily lives. And instead of depending on the analyst's view of which topics are important at any time, we can observe the topics that the public chooses to raise without our prompting.

The second major appeal of social media data is its reach: over time, across individuals, cross-nationally and within small geographical regions. Due to the fine-grained nature of Twitter and Facebook data, for example, it should be possible to measure changes in opinion on a daily or even hourly basis. Similarly, due to the fact that hundreds of millions of people use Twitter and Facebook regularly, the scope of opinion that can be measured goes far beyond anything we could previously have attempted. And since social media can be found throughout the world, it provides a convenient platform for sampling opinion in many countries where it would otherwise be difficult or impossible for survey researchers to work. In fact, it is likely that the Twitter archive is already the largest cross-national time-series data set of individual public opinion available to the mass public.[4]

The third appeal of using social media to measure public opinion is its cost and practicality. With a little programming and a decent-sized hard drive, anyone can capture every comment made about a presidential debate, for instance, in real time and for free. To the extent that we care about public opinion because we think it helps to hold rulers more accountable and to make policy more responsive to the mass citizenry, the potential to dramatically reduce the cost of studying public opinion may be perhaps the most exciting opportunity afforded by social media of all.[5]

[3] The newspaper industry, a major source of public opinion polls, has shrunk 43% from 2000 to 2012 (see: http://www.stateofthemedia.org/2012/overview-4/). The declining public support to higher education due to the financial crisis has led to the closing of some university-based survey research centers (Keeter, 2012), and there have been increasing political pressures to defund such initiatives as the American Community Survey and the Economic Census. Overall interest in polls, however, has only grown, with the total number of active pollsters (with at least 10 polls per campaign) having risen since 2000: in presidential years, this has increased from appoximately 10 to 20 over the last two decades, and from approximately 5 to 10 in midterm elections (based on our analysis of data from http://projects.fivethirtyeight.com/pollster-ratings/).

[4] The Twitter archive is of course dwarfed by the Facebook archive, but this is not yet available to the public. And to be clear, by "available" we mean available for purchase; collecting relatively large amounts of Twitter data is free in real time, but it is not free to retrieve tweets with a broad backward looking search.

[5] It is also one that raises all sorts of new questions for social scientists, who will find themselves in the future wanting to work with huge private companies, such as Facebook or Twitter, much in the way that natural scientists have had to learn how to work with big pharma. Although beyond the scope of the current chapter, this too will likely pose all sorts of new challenges for researchers, the likes of which we have previously rarely encountered.

Of course while social media has desirable properties that traditional public opinion surveys cannot match, truly developing tools to effectively harness its potential will involve facing enormous challenges, which we discuss in the next section. Each of the strengths discussed above also constitutes a challenge – both theoretical and technical – for measuring opinion in the ways we are used to from traditional surveys. First, identifying emergent topics and sentiments is hugely challenging, not just computationally but theoretically, as we strive to understand machine- or human-generated summaries and reconcile them with previous survey measures and research agendas. Second, the breadth and scale of social media use is counter-balanced by the opacity of its user population, and the steps needed to reweigh this entirely unrepresentative "survey" in order to measure any population of interest remain difficult and uncertain. And third, the technical challenges of collecting and aggregating the data are non-trivial, particularly given the diffident and often opaque cooperation of private social media providers like Twitter and Facebook.

We believe that many of these challenges involved in using social media data to study public opinion can be overcome, and that the potential payoff certainly justifies the effort. But we also believe it is crucial to be upfront about these challenges moving forward, and therefore one of the goals of this chapter is to lay these out explicitly. In section 2, we discuss these three main challenges in greater detail, and how they have arisen in past social media research. In section 3, we discuss some of the strategies for overcoming many of these challenges, both drawing upon past work that suggests various successful strategies, and suggesting new ones. And in section 4, we discuss in greater detail some of the novel uses for social media, ones that have fewer direct analogues in traditional survey work. We conclude in section 5 with a series of recommendations for a research agenda in using social media for public opinion work, as well as providing a list describing *how* social media data was collected that we suggest scholars and practitioners use when reporting any results based on social media data, but especially when reporting results claiming to be representative of public opinion.

We focus here on Twitter data because it is widely used, mainly public, and relatively easy to collect; for these reasons, it has been the focus of the majority of recent social media research. But of course all of these concepts apply more generally, and the difficulties and solutions we propose here will likely continue well into a future where social media platforms that do not exist yet may dominate the landscape .

2 Challenges in the measurement of public opinion with social media data

In the study of public opinion, a survey is commonly defined as a systematic method for gathering information from a sample of individuals for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the individuals are members (see e.g. Groves et al., 2011). This information is commonly gathered by asking people questions. The three core components of a survey are thus: a standardized questionnaire, a population frame from which individuals are sampled using a probability sampling method, and a method to aggregate individual responses to estimate a quantity of interest.

Without any adjustment, treating social media data as a survey fails to meet any of these three criteria: the opinions expressed by individuals are unprompted and unstructured; the probability that an individual is included in the sample varies in systematic but opaque ways; and the collection and aggregation of data into quantities of interest is problematic due to uncertainties in the data generating and collection processes. In this section we describe each of these difficulties and why they are critical for the measurement of public opinion with social media data.

We note that when we speak of trying to measure 'public opinion' we will be primarily concerned with the traditional notion of who 'the public' is: adults in a particular polity (or set of polities). However, one of the benefits of social media is that there is no such constraint on whose opinion is uttered on social media. We are potentially able to measure sub-populations of interest within a polity, such as ethnic groups, ideological groups, or speakers of particular languages (Metzger et al., 2016). On the other hand, this also extends to populations such as children, political activists, persecuted minorities, and other subpopulations that desire, expect, or deserve privacy in their online activities. Fully tackling the myriad ethical issues entailed in using social media to measure public opinion would require an entire separate handbook chapter, but we should be aware that such issues permeate every stage discussed below. Some of these issues are common to any sort of collection of publicly available data, including the issue of consent regarding data that has been made public but which may be used in ways not anticipated by the participant, and the collection of data from minors and others not able to give consent themselves. Others are common to data collection and storage more generally, including data protection and anonymization, and issues specific to sharing data, particularly for replication purposes. Other questions are more specific to social media data, including how to deal with posts that were deleted by users after the data was collected; the potential privacy violations inherent in using sophisticated machine learning methods to infer demographic and other characteristics that had not been publicly revealed; and the question of whether the results of these analyses could put exposed users at risk of political or other forms of retaliation (Flicker, Haans and Skinner, 2004; Tuunainen, Pitkänen and Hovi, 2009; Zimmer, 2010; Solberg, 2010; Bruns et al., 2014). The scope of ethical considerations in social media studies is rapidly growing, but for the present purposes we will focus here on the technical challenges of measuring public opinion.

2.1 Identifying political opinion

If we seek to fit social media into the framework of existing public opinion measurement, we may consider social media posts as something like unstructured and entirely voluntary responses to external stimuli analogous to public opinion questions. In this sense, like a survey question, the stimuli set (or affect) the topics, and our job is to identify these topics and turn the unstructured responses into something like sentiment, approval levels, feeling thermometers, or the like. In both traditional surveys and unstructured social media, we have something like a subject (the question, or a post's topic) and a predicate (the numeric response, or a post's sentiment), and we seek to turn the raw data of the unstructured social media text and metadata into something more like the structured survey responses we are familiar with. This analogy is often latent in research using social media to measure public opinion, but making it more explicit clarifies a number of issues in putting social media to such a use. This distinction has also been referred to as the distinction between 'Designed Data' and 'Organic Data' (Groves, 2011). Whereas traditional collections of public opinion data, or data on economic behavior based on survey responses, are curated and created by the designer with intent in mind, many datasets now available are based on data that exist simply because much human behavior occurs online - and is recorded.

First, the questions are not directly asked of people; instead, people give their opinions in response to events and discussions. How do we define what topics to examine and which tweets are relevant for a given topic? For example, if we want to measure users' sentiment toward the candidates in the 2016 Presidential election, how do we identify a corpus of relevant tweets? The vast majority of studies focus on tweets mentioning candidate names, without discussing the possibility of systematic selection bias in determining the search criteria in this way (but see King, Lam and Roberts, 2014). For example, focusing only on tweets that mention Hillary Clinton or Donald Trump may miss a number of social media messages that also relate to the 2016 election but do not mention candidate names (He and Rothschild, 2014). Should tweets that refer to either candidate *without* using the candidate's name tend to be either more positive or negative than tweets that do explicitly mention the candidate's name, then obviously selecting tweets based on the use of the name will generate a large amount of selection bias. And that's just bias relative to the full corpus of tweets on the candidates, even apart from bias relative to the population of interest: it may also be that only persons with some particular characteristic use particular terms. If that is the case, and we omit terms used by that group, we will fail to measure group opinion accurately. Even without generating bias by collecting based on candidate names, collections based on names may include substantial noise, or miss substantial numbers of tweets. Tweets containing "Hillary" in 2016 may be predominantly about Hillary Clinton, but tweets containing "Trump" or "Cruz" may not be about either Donald Trump or Ted Cruz, thus adding noise to the corpus. While filtering on tweets containing "Donald Trump" or "Ted Cruz" may miss many tweets actually focused on the candidates. In general, the choice of the relevant corpus of tweets is almost invariably ad hoc, in part because the analyst cannot be omniscient about what constitutes the set of tweets related to a given topic.

In addition to defining the topics and content that shape the collected data, measuring the topics in individual tweets, particular when topics may be rapidly changing over time or responding to major events, remains both a technical and theoretical challenge. Are people responding to changing questions on the same topic, or are the topics themselves changing, or do we need a complex hierarchical and temporal structure of all our content before we can begin to quantify public opinion in a systematic way? For example, during presidential debates, there was considerably more commentary and discussion among users than at other times, when information-sharing tweets (with high frequency of URLs within tweets) were more common (Diaz et al., 2014). Similarly, with politically charged events, such as the Wisconsin Labor strikes of 2011, many social media users seemed to have been particularly focused on tweeting non-mainstream news and alternative narratives of the protest, unlike during less contentious events (Veenstra et al., 2014). The same occurs during mass protest events, since regime elites can respond strategically to protest and try to shift the focus of the discussion (Munger, 2015; King, Pan and Roberts, 2016). And topics themselves can change because the comments on social media may represent a change in public opinion: either the development of a new issue that was previously not part of political discourse, or the disapearance of an issue from public concern. The set of topics dominating political discussion in 2000 would be very different than the set of topics dominating political discussion in 2016. And just as refusal to answer surveys may not be random, but may vary systematically with the likely response, discussion on any issue on social media may vary with context. During a period of 'good news' for a candidate, we may see more tweets by the candidate's supporters, and vice-versa. Thus the population, topics, and sentiments may all be continually shifting in ways that are very challenging to measure.

Even assuming we are able to resolve the subject – the topics – what of the predicate: the sentiment, approval, enthusiasm, etc? What exactly is the quantity of interest? Simple counts of mentions of political parties or issues is a method that in some cases has produced meaningful results. For example, Tumasjan et al. (2010) and Skoric et al. (2012) showed that mentions of parties on Twitter were correlated with election results. However, that is often not the case (Metaxas, Mustafaraj and Gayo-Avello, 2011; Bermingham and Smeaton, 2011). In fact, Gayo-Avello (2011) showed that tweet-counting methods perform worse than a random classifier assigning vote intentions based on the proportion of votes from a subset of users who directly revealed their election-day intentions to the researcher. Similarly, Jungherr, Jürgens and Schoen (2012) criticize the tweet-counting method used by Tumasjan et al. (2010) to predict German elections for focusing only on tweets mentioning the largest parties. They show that if tweets mentioning a new party – the Pirate Party – were counted as well, the results differed considerably and mispredicted the election outcome, as the Pirate Party was a clearly predicted election winner, whereas in fact it won only 2 percent of the vote (see also Jungherr et al., 2016).

One common alternative to counting methods is the use of sentiment analysis, which aims at measuring not the volume of tweets on a particular topic, but the valence of their content. This method often relies on existing dictionaries of positive and negative words, where the ratio of positive to negative words that co-occur with a topic on a given day (e.g.) is taken as a measure of the overall public sentiment on that topic on that day. For example, O'Connor et al. (2010) shows that Twitter sentiment over time in economics-related tweets are correlated with consumer confidence measures in the U.S. The downside of this approach is that its performance can vary in unpredictable ways. The approach depends on potentially ad-hoc dictionaries and often exhibits low out-of-sample accuracy (González-Bailón and Paltoglou, 2015) and even significant differences in its performance across different applications within a similar context. For example, Gayo-Avello (2011) finds that the performance of a lexicon-based classifier was considerably more reliable for tweets about Barack Obama than John McCain during the 2008 election campaign.

Finally, even if we have a good method for measuring topics and (e.g.) sentiments, it is not at all clear that what we are measuring is necessarily an honest expression of opinion. It remains unknown to what degree the (semi-)public nature of social media could induce stronger social desirability bias than in the context of traditional survey responses. On the one hand, given potential social stigma, users may be even less likely to reveal attitudes on sensitive topics than in standard surveys (Newman et al., 2011; Pavalanathan and De Choudhury, 2015), and individuals can control their content after it is posted, with changes and edits potentially inducing selection bias in the type of content that remains (Marwick and Boyd, 2011). On the other hand, Twitter in particular does allow users a certain degree of anonymity (though perhaps less than they think), and thus may allow individuals to express their true preferences and attitudes more honestly than in many traditional surveys (Joinson, 1999; Richman et al., 1999). However, to our knowledge this potential issue has not been examined systematically in the context of measuring public opinion on political (particularly sensitive) topics.

2.2 Representativeness of social media users

One crucial advantage we lose with social media relative to traditional surveys is the opportunity to control our sampling frame. Traditional surveys attempt to guarantee a known probability of any individual in the population being asked a survey question. Where those surveys fail is due to both high and non-random non-response, and non-random item non-response. With social media, since control of the sampling frame is lost, we can neither know the likelihood that someone is asked a "question", nor know the likelihood of a response. In a traditional survey in order to generalize to a target population we have to assume that non-responses are missing at random, or that they are missing at random conditioning on measured covariates. In the best of worlds, this is a strong assumption: it may be that people who choose not to reveal their preferences on something are systematically different than those who do reveal their preferences. On social media, where we do not ask the question but depend on the participant to reveal their opinions we might have more trouble. The set of people offering *unprompted* opinions on a topic may be more passionate or different in myriad other ways from the set of people who offer opinions on that topic when explicitly asked. This presumably makes our missing data problems far worse than those caused by traditional survey non-response.

It would of course be extremely unwise to generalize directly from Twitter behavior to any of the standard populations of interest in most surveys. A number of studies have demonstrated that Twitter users are not representative of national populations (Duggan and Brenner, 2015; Mislove et al., 2011; Malik et al., 2015). For example, in the U.S., most populous counties are overrepresented, and the user population is non-representative in terms of race (Mislove et al., 2011). Comparing geotagged tweets and census data, Malik et al. (2015) also demonstrate significant biases towards younger users and users of higher income. Differences in usage rates of social media platforms across countries is also an obstacle for the comparative study of public opinion (Mocanu et al., 2013). These differences are also present, although perhaps to a lesser extent, in the analysis of other social media platforms like Facebook (Duggan and Brenner, 2015).

For the purposes of the study of public opinion, however, it is more important whether and how representative the politically active Twitter users are relative to the general population. But here too, the evidence is consistent with Twitter users being highly non-representative. For example, women are the majority of Twitter users, but a much smaller minority among politically active Twitter users (Hampton et al., 2011); politically active Twitter users are more polarized than the general population (Barberá and Rivero, 2014); and they are typically younger, better educated, more interested in politics, and ideologically more left-wing than the population as a whole (Vaccari et al., 2013). Crucially, non-representativeness may even vary by topic analyzed, as different issues attract different users to debate it (Diaz et al., 2014).[6]

Evaluating the representativeness of Twitter users is not straightforward, given that unlike standard surveys, Twitter does not record precise demographic information. Instead, most studies try to infer these characteristics. While some approaches have been quite successful (see e.g. Al Zamal, Liu and Ruths, 2012*a*; Barberá and Rivero, 2014), these are still approximations. These difficulties can be compounded by the possibility of bots and spammers acting like humans (Nexgate, 2013), specially in the context of autocratic regimes (Sanovich, 2015). It becomes much harder to infer how representative tweets are of any given population if some tweets come from automated computer programs, not people. And even determining how many people are paying

[6] Note that all of the studies cited here are country-specific, we cannot really make these claims about the global set of Twitter users.

attention to discussions is problematic as fake accounts can be used to inflate common metrics of popularity. For example, one study found at least 20 sellers of followers on eBay, at an average price of $18 per 1,000 followers, demonstrating how fake accounts can rack up followers very easily (Barracuda Labs, 2012).[7] In addition, there may be important deviations from one-to-one correspondences between individual users and individual accounts, given the existence of duplicate and parody accounts, and accounts that represent institutions, companies or products, such as the White House, Walmart or Coca-Cola, for example.

Moreover, the demographic composition of users can change over time, particularly in response to important events, such as presidential debates or primaries. These changes may be quite unpredictable. For example, during the presidential debates in the 2012 Presidential election in the U.S., the male over-representation among political tweeters dropped significantly, whereas the geographic distribution of tweets (by region) became considerably less representative (Diaz et al., 2014). In the Spanish context, Barberá and Rivero (2014) find that important events during the 2011 legislative election, such as party conferences and the televised debates, increase the inequality on Twitter by increasing the rate of participation of the most active and most polarized users. It is important to keep these shifts in mind, since raw aggregates of public opinion may be due to these shifts in demographic composition rather than any shifts in actual opinion (see e.g. Wang et al., 2014).

2.3 Aggregating from individual responses to public opinion

There are a number of other platform-specific issues that also affect researchers' ability to aggregate individual social media messages. At present, access to 100% of tweets is only available through third-party companies like Gnip (recently bought by Twitter) at prices often beyond what most researchers can afford. Instead, researchers rely on Twitter's Streaming API, which only provides content in real-time, not historical data. That means most researchers have to anticipate in advance the period of study they will focus on. Results can change significantly when using different time windows (Jungherr, 2014), which can lead to ad-hoc choices of period of coverage, and a non-negligible likelihood of missing key events.

Most importantly, Morstatter et al. (2013) and González-Bailón et al. (2014) found significant differences between the full population of tweets (the so-called Twitter "firehose") and the samples obtained through Twitter's Streaming API, the most popular source of data used by researchers. In particular, it appears the rate of coverage (the share of relevant content provided by the Streaming

[7] Such concerns could be particularly pernicious if politicians are buying bots precisely for the *purpose* of manipulating measures of public opinion. Although we do not have evidence of this yet occurring, it does not seem to be a large leap to imagine politicians moving from simply buying followers to buying accounts that will deliver positive sentiment about themselves (or negative sentiment about opponents) in an attempt to manipulate reports in the media about online popularity.

API relative to all content) varies considerably over time; the topics extracted through text analysis from the Streaming API can significantly differ from those extracted from the Firehose data; that users who participate less frequently are more likely to be excluded from the sample; and that top hashtags from the Streaming API data can deviate significantly from the full data when focusing on a small number of hashtags.

In those cases in which researchers are interested in aggregating data from social media to specific geographic units such as a state or congressional district, they face the problem that only a small proportion of tweets are annotated with exact coordinates (Leetaru et al., 2013). Geolocated tweets are highly precise, but are not a representative subset of all tweets (Malik et al., 2015). An alternative is to parse the text in the "location" field of users' profiles. While this increases the degree of coverage, it is not a perfect solution either, as Hecht et al. (2011) found that up to a third of Twitter users do not provide any sort of valid geographic information into this field.

Finally, one important issue often overlooked in social media studies is that, given Twitter's opt-in nature, tweets often cannot be treated as independent because many individuals tweet multiple times. It is often the case that only a minority of unique individuals dominates the discussion in terms of tweet and retweet volume, making over-sampling of most active users very likely (Barberá and Rivero, 2014; Gruzd and Haythornthwaite, 2013; Mustafaraj et al., 2011). For example, in the run-up to the 2012 presidential election, 70% of tweets came from the top 10% of users, with 40% of the tweets coming from the top 1% of users (Barberá and Rivero, 2014). This problem is exacerbated by practices such as astroturfing – coordinated messaging from multiple centrally-controlled accounts disguised as spontaneous behavior (Castillo, Mendoza and Poblete, 2011; Morris et al., 2012). Importantly, politically-motivated actors use astroturf-like strategies to influence the opinion of their candidates during electoral campaigns (Kreiss, 2014; Mustafaraj and Metaxas, 2010). The more influential are the attempts to characterize the behavior of online users, the greater may be the incentive to manipulate such behavior (Lazer et al., 2014).

3 How should it be done? Potential solutions and areas for future research

The three concerns about using social media data to measure public opinion outlined in the previous section – measuring opinion; assessing representativeness; and overcoming technical challenges in aggregation – are, we argue, the main challenges to overcome in this field. Each of these stages has its analog in traditional survey methodology, but each presents unique challenges when using social media. In this section we describe current efforts by previous studies to address these issues, and potential solutions that could be implemented in future research.

3.1 Better methods for identifying political opinion

In choosing the corpus of tweets that will be included in the analysis, previous studies often define a set of ad-hoc search criteria, such as a list of hashtags related to an event, or the names of political actors. This is partially driven by the limitations imposed by Twitter's Streaming API and researchers' inability to collect historic data freely. We claim that it is necessary to establish more systematic criteria to select what set of tweets will be included in the sample.

One approach that has yielded promising results is the development of automated selection of keywords. He and Rothschild (2014) apply such a method in their study of the 2012 Senate elections: they start with a corpus drawn based on candidate names, and then iteratively expand it by identifying the most likely entities related to each candidate. Their final corpus is 3.2 times larger, which gives an indication of the magnitude of the potential biases associated with simple keyword selection methods. For example, they find that the aggregate sentiment of tweets mentioning only candidate names is different from that of the extended corpus after applying their selection method. King, Lam and Roberts (2014) also propose a similar method that adds human supervision in the selection of new keywords to resolve linguistic ambiguities and reduce the proportion of false positives.

An alternative solution is to abandon keyword filtering altogether and instead sample at the user level. As Lin et al. (2013) demonstrate, tracking opinion shifts within a carefully selected group of Twitter users can overcome some of the limitations mentioned above by learning from users' prior behavior to detect their bias, and control for it in any analysis.[8] These "computational focus groups" could be further improved if they are combined with surveys of Twitter users that contain questions about sociodemographic and political variables (Vaccari et al., 2013).

In addition to topics, the other half of assessing opinion is the predicate side, such as the estimation of sentiment about those topics. One of the most successful examples of sentiment analysis applied to election prediction, the Voices from the Blogs project (Ceron et al., 2014; Ceron, Curini and Iacus, 2015), combines supervised learning methods with human supervision in the creation of datasets of labeled tweets that are *specific to each example*. González-Bailón and Paltoglou (2015) conducted a systematic comparison of dictionary and machine learning methods finding similar results: classifiers trained with a random sample of the dataset to be used for prediction purposes outperform dictionary methods, which are in many cases no better than random. One possible refinement of application-specific methods is the combination of topic models and sentiment analysis (Fang et al., 2015), which could leverage differences in words' usage across different topics to improve the performance of these techniques.

[8] See below for a discussion of working with a randomly chosen set of users.

3.2 Increasing representativeness

The majority of studies using Twitter data to date, particularly those estimating voting preferences and predicting election outcomes, do not attempt to address the non-representativeness of (politically-active) Twitter users (the exceptions include Gayo-Avello, 2011; Choy et al., 2011; 2012). In fact, many of these studies do not clearly specify the target population, which in the case of electoral predictions should be the voting population. The implicit assumption is that the size of the data, the diversity of Twitter users, and the decentralized nature of social media may compensate for any potential bias in the sample. Of course as we know in cases where it has been studied, the set of Twitter users is *not* representative of typical target populations such as voters or eligible voters (see e.g. Duggan and Brenner, 2015).

We need considerably more work to examine the plausibility of these assumptions. On the one hand, for predictive purposes, the skew in the sample may not be problematic if politically active users on Twitter act as opinion leaders who can influence the behavior of media outlets (Ampofo, Anstead and O'Loughlin, 2011; Farrell and Drezner, 2008; Kreiss, 2014) or a wider audience (Vaccari et al., 2013). On the other hand, as we discussed in the previous section, the non-representativeness of these users relative to the general population may be quite severe, suggesting that the biases may not balance out unless addressed by reweighting.

One potentially promising method is multi-level modeling and post-stratification (MRP), particularly because it relies on post-stratification adjustments to correct for known differences between the sample and the target population (Little, 1993; other potential weighting approaches can be found in AAPOR, 2010). Somewhat like traditional weighting in telephone or online polls, this approach partitions the target population into cells based on combinations of certain demographic characteristics, estimates via multi-level modeling the variable of interest in the sample within each cell (e.g. average presidential approval for white females, ages 18–29, etc.), and then aggregates the cell-level estimates up to the population level by weighting each cell by the proportion in the target population (Park, Gelman and Bafumi, 2004; Lax and Phillips, 2009). This approach has been fruitfully used to generate quite accurate election predictions from highly non-representative samples, such as XBox users (Wang et al., 2014).

The main challenge with this approach is of course to obtain the detailed sample demographics needed for post-stratification. Twitter does not collect or provide data on demographics. And unlike some other platforms such as Facebook, Twitter metadata and profile feeds contain limited information to directly classify users. There are two ways to address this concern: one, to consider demographic variables as latent traits to be estimated, and two, to augment Twitter data with other types of data, such as voter registration records or surveys.

Pennacchiotti and Popescu (2011) and Rao et al. (2010) provide proofs of concept that demonstrate that coarse categories of age, political orientation, ethnicity, and location can be estimated by applying a variety of supervised machine-learning algorithms to user profiles, tweets, and social networks. Al Zamal, Liu and Ruths (2012*b*) demonstrate that users' networks (i.e. who they follow and their followers) can be particularly informative about their age and gender. However, these studies often rely on small convenience samples of labeled users, and it is still an open question whether these methods can scale up to the large samples researchers often work with.

One of the key variables in MRP applications has been party ID (Park, Gelman and Bafumi, 2004; Lax and Phillips, 2009). Thus it is extremely useful to be able to infer ideological orientation and partisanship, in addition to gender, ethnicity, age, and geographic location. There are several promising approaches in this direction. Barberá (2015) shows that Twitter users' ideology can be accurately estimated by observing what political actors they decide to follow. Other studies estimate political ideology or partisan identification using different sources of information, such as the structure of retweet interactions, follower networks, or similarity in word use with respect to political elites (Boutet, Kim and Yoneki, 2013; Cohen and Ruths, 2013; Conover et al., 2011; Golbeck and Hansen, 2011; Wong et al., 2013). One limitation of these approaches is that ideology, as well as the other demographic variables, often cannot be estimated for the entire sample of users, or at least with the same degree of accuracy, especially if they rely on usage of specific hashtags, which can vary significantly across users.

An alternative solution to this problem is to augment Twitter data with demographic information from other sources. For example, Bode and Dalrymple (2014) and Vaccari et al. (2013) conducted surveys of Twitter users by sampling and directly contacting respondents through this platform, achieving relatively high response and completion rates. By asking respondents to provide their Twitter user name, they were able to learn key characteristics of a set of twitter users directly from survey responses provided by those users. Matching Twitter profiles with voting registration files, publicly available in the United States, can also provide researchers with additional covariates, such as party affiliation, gender, and age (see e.g. Barberá, Jost, Nagler, Tucker and Bonneau, 2015). The subset of users for which this information is available could then be used as a training dataset for a supervised learning classifier that infers these sociodemographic characteristics for all Twitter users.[9] These matching approaches could also be conducted at the zipcode or county level with census data to control for aggregate-level income or education levels (see e.g. Eichstaedt et al., 2015).

[9] As an additional challenge, social media users and their demographic distributions are presumably constantly evolving, so these models will have to be frequently updated to keep up with this rapidly shifting landscape.

3.3 Improving aggregation

It is perhaps the last step – aggregating from tweets to a measure of public opinion – on which most attention has been placed in previous studies. We now have a good understanding of the biases induced by how Twitter samples the data that will be made available through the API (Morstatter et al., 2013; González-Bailón et al., 2014), the power-law distribution of users' Twitter activity (Barberá and Rivero, 2014; Wu et al., 2011), and that very few tweets contain enough information to locate their geographic origin (Leetaru et al., 2013; Compton, Jurgens and Allen, 2014). Researchers need to be aware of these limitations and address them in their analyses. For example, if the purpose of a study is to measure public opinion about a topic, then the analysis should add weights at the user level to control for their different levels of participation in the conversation. When such a solution is not possible, the study should include a discussion of the direction and magnitude of the potential biases introduced by these limitations.

Finally, regardless of the approaches to aggregation, weighting, or opinion measurement that we chose, an important step in any analysis should be the removal of spam messages and accounts (or *bots*), which in some cases can represent a large share of the dataset (King, Pan and Roberts, 2016). One option is to apply simple filters to remove users who are not active or exhibit suspicious behavior pattern. For example, in their study of political communication on Twitter, Barberá, Jost, Nagler, Tucker and Bonneau (2015) only consider users who sent tweets related to at least two different topics, which should filter spam bots that "hijack" a specific trending topic or hashtag (Thomas, Grier and Paxson, 2012). Ratkiewicz et al. (2011) and Castillo, Mendoza and Poblete (2011) implement more sophisticated methods that rely on supervised learning to find accounts that are intentionally spreading misinformation. Their study shows that spam users often leave a distinct footprint, such as a low number of connections to other users, high retweet count between a limited set of strongly connected (and likely fake) users, and a string of very similar URLs (e.g. differing only in mechanically created suffixes). Therefore, it appears possible and therefore potentially warranted to invest more effort in pre-processing the data by removing the suspect content, or at least inspecting the sensitivity of the results to the presence of bot accounts.

3.4 Validation

Once we have specified our data collection and aggregation strategies, our population of interest and weighting strategies, and our opinion measurement methods, it is essential to validate these purported measures against trusted ground truths, or at least previously established measures. The success of these approaches must be examined relative to clear benchmarks, such as previous election results, existing surveys, public records, manually labeled data, etc. (Metaxas, Mustafaraj and Gayo-Avello, 2011; Beauchamp, 2016). This validation should be conducted with out-of-sample data, ideally forward in time, and should be measured statistically, by computing the predicted accuracy. Depending on the application, other forms of validity should be considered, such as convergent construct validity (the extent to which the measure matches other measures of the same variable) or, in the case of topic-specific measures, semantic validity (the extent to which each topic has a coherent meaning).[10]

Conversely, rather than engaging in demographics-based weighting and topic/sentiment estimation in order to predict public opinion, it may also be possible to reverse the validation process, and instead train machine-learning models to sift through thousands of raw features (such as word counts) to find those that directly correlate with variations in the quantities of interest (such as past polling measures of vote intention) (Beauchamp, 2016). In this way, one could potentially go directly from word counts and other meta-data (such as retweets, urls, or network data) to opinion tracking with no worry about demographics, topics, or sentiments – although potentially at the cost of interpretability and generalizability to other regions, times, and political circumstances.

4 New directions for measuring public opinion: going beyond survey replication

As we have said, each of the challenges to using social media data to measure public opinion also reveals how social media can also be taken well beyond existing survey methods.

Weighting and demographics aside, the sheer size of social media data makes it theoretically possible to study sub-populations that would not be possible with traditional survey data, including sub-populations defined by demographic, geographic, or even temporal characteristics (Aragón et al., 2016; Barberá, Wang, Bonneau, Jost, Nagler, Tucker and González-Bailón, 2015). Social media also enables us to measure opinion across national borders. While Twitter penetration varies in different countries (Poblete et al., 2011), as long as we know something about the characteristics of who in a country is on Twitter we can try to generalize from tweets to a measure of mass opinion in a country.

Because of its organic nature, social media data is generated continuously, and thus we can track changes over time at very fine-grained temporal units (for example, Golder and Macy 2011 track changes in mood across the world over a course of one day). This means we can aggregate the data by any temporal unit we choose, and simulate designed data for tracking opinion change over time, and for using in traditional time-series analysis. Moreover, because social media data comes with individual identifiers, it also constitutes panel data. We (often) have repeated observations from the same informant. This high frequency means that social media can reveal public

[10] See Quinn et al. (2010) for a more extensive discussion of different types of validity.

opinion changes over time about issues that are not polled very frequently by traditional surveys. Reasonably dense time-series survey data exists for some issues, such as presidential approval or consumer sentiment, but social media data offers the opportunity to put together dense time series of public opinion on a host of specific issues that are rarely or infrequently polled.[11] And by taking advantage of information identifying characteristics of informants, those time series could be evaluated for distinct sub-groups of populations.

The temporal nature of social media also lets us observe the emergence of public opinion. This is perhaps social media's greatest strength: it does not depend on the analyst to ask a preconceived question. So while at some point almost no survey firm would think to invest in asking respondents whether or not they thought gay marriage or marijuana should be legal, by collecting sufficiently large collections of social media posts in real time it should be possible to observe when new issues emerge, and indeed to identify these newly emerging issues *before* we even know we should be looking for them. While the free-form nature of opinion revelation on social media can be a barrier to measuring what *everyone* is thinking about an issue, it may give us a way to measure not just sentiment, but intensity of sentiment via content and retweeting, as well as richer measures of sentiment along as many potential dimensions as there are topics.[12]

Social media also allows us to measure not just mass opinion, but especially that of political activists and other elites. Legislators, political parties, interest groups, world leaders, and many other political elites tweet. And while these are public revelations and not necessarily truthful, we do see what these actors choose to reveal. And we are able to see this contemporaneously with mass opinion revelation. And again by taking advantage of the fine-grained temporal nature of social media data, we can observe how elite opinion responds to mass opinion and vice-versa (Barberá et al., 2014; Franco, Grimmer and Lee, 2016), and how both groups respond to exogenous events. While of course we need to be sensitive to the fact that elites know that what they are posting on Twitter or Facebook is intended for public consumption and thus may not reflect genuine "opinion" in the sense that we are trying to measure mass opinion, the social media record of elite expression may nevertheless prove extremely valuable for studying both elite communication strategy generally as well as *changes* in the issues that elites are emphasizing. Thus while we may not know whether a particular politician genuinely believes gun control laws need to be changed, social media can easily help us measure whether that politician is emphasizing gun control more at time t than at time $t-1$.

[11] For instance, such topics might include intermittently polled issues in the US like gun control or immigration; government approval measures in less well-polled nations; public opinion about specific foreign or domestic policies (eg, Syria or the ACA) or factual questions (eg, climate change or GMOs); and more local issues, such as opinion on the policies or services in specific cities.

[12] In addition to issues with representativeness, the public nature of social media means that these sentiments are presumably also affected by social desirability bias. It may be that in these more polarized times, mean sentiment will remain representative even as both sides are driven to extremes by social pressures, but it will nevertheless be important to measure and correct for these effects using existing polling measures as ground-truth tests.

Social media also comes with a natural set of contextual data about revealed opinion: the social network of the individual informant (Larson et al., 2016). Attempts to measure this with survey questions are notoriously difficult as people have not proven capable of stating the size of their network, much less providing information necessary for contacting the network members. Yet social media provides us not only with the size of the social networks of informants, but a means to measure the opinion of network members. Traditional surveys have tried to measure the connection of network ties primarily by depending on self-response, which has proven to be notoriously unreliable.

Social media data also provides the potential to link online information directly to opinion data. Many social media users reveal enough information about themselves to make it possible to link them to public records such as voter files. Social media data can also be directly supplemented with survey data obtained by contacting social media users directly through Twitter, via "replies" (as in Vaccari et al., 2013) or promoted tweets targeted to a specific list of users. However, it remains challenging to construct matched samples using these more direct methods that are sufficiently large and representative, or which can be reweighted to ensure representativeness.

Finally, social media significantly democratizes the study of public opinion. Researchers can potentially address novel research questions without having to field their own surveys, which are often much more costly. Access to large volumes of social media is free and immediate, unlike many existing surveys that may be embargoed or restricted. Moreover, this extends well beyond scholars of public opinion: Anyone – from campaigns selling political candidates or consumer goods, to regimes trying to understand public wants – can access social media data and see what people are saying about any given issue with minimal expense. Even in a world where the number of surveys and surveyed questions are proliferating, social media potentially offers a spectrum of topics and a temporal and geographic density than cannot be matched by existing survey methods.

5 A research agenda for public opinion and social media

Making better use of social media for measuring public opinion requires making progress on multiple fronts. Perhaps the issue that remains the most theoretically challenging is the measurement of topic and sentiment: the "question" and "response." The unstructured text is what is most unusual about social media – it is *not* an answer to a question specified by the researcher but rather free-form writing – and attempts to transform it into the traditional lingua-franca of opinion research – answers to specific questions – remains an open problem. We may even eventually discover that it is an obsolete one, as we move entirely beyond the constraints imposed by traditional surveys into the naturally high-dimensional world of free-form text. The tremendous opportunity presented by social media data makes the payoff for solving such problems worth the investment. Social media data can give us measures of public opinion on a scale both geographically, temporally, and in breadth of subject that is vastly beyond anything we can measure with other means.

There are of course major mistakes that can be made when analyzing social media data. One advantage we have pointed out about social media is that it democratizes measuring public opinion: anyone can do it. That means anyone can do it badly. And the endemic deficiency of ground-truths can make it hard to know when a measure is a bad measure. Reputable scholars or organizations reporting measures based on traditional polls have adopted standardized practices to increase the transparency of their reporting (items such as sample size, response rates, and so one). We thus conclude with some obvious standards in reporting that social media-based measures of opinion should adhere to in order to at least guarantee a minimum amount of transparency and allow readers, or users of the measures created, to better evaluate the measures. We describe and list standards with respect to analyzing data from Twitter, but these can be easily applied to other sources of social media with appropriate modifications. The points apply generally across organic public opinion data.

First, researchers need to be clear on the technical means of gathering data. Data gathered in real-time thru a rate-limited means could be incomplete, or differ in unpredictable ways from data purchased after the fact from an archive or collected thru the Firehose. Second, researchers need to very clearly report the limitations placed on the sampling frame. Data on social media can be gathered based on user-ids, or on content of text, and can be further filtered based on other meta-data of users or individual tweets (such as language, or time of day).

Second, researchers need to explain whether data was gathered based on the context of the text, the sender of the text, or some contextual information (such as time or place of tweet).

Third, researchers need to very precisely describe the criteria for inclusion in their sample, and how those criteria were arrived at. If a sample is based on keywords, researchers need to describe how the keywords were selected. If someone claims to be measuring opinion about gun control, they could state that: "we collected all tweets about gun control." But this claim could not be evaluated unless the full set of keywords used to gather tweets is provided. One could collect all tweets containing the expression "gun control", but that would omit many tweets using assorted relevant hashtags and phrases. Or, if an analyst were to try to measure public opinion about Barack Obama with all tweets containing 'Barack Omaba', the analyst would miss any tweets that are about Barack Obama, but do not use his name. If one were to omit all tweets that contain 'Barack Hussein Obama' rather than 'Barack Obama', then this could obviously cause significant measurement error. Thus precise description of what can be included in the corpus of text is important.

If the corpus is collected with a set of keywords, the analyst should explain how the set was generated. Did the analyst assume omniscience and create the list of keywords, or was it generated using some algorithim proceeding from a set of core terms, and finding co-occuring terms? And, no matter how the keywords were chosen, or how the topic was collected, the analyst should describe any testing done to confirm that the corpus was in fact about the chosen topic.

Researchers also need to note if filters or constraints were imposed on the collection that would exclude some tweets based on language or geography? If one is *only* measuring opinions expressed in English, that is important. Similarly, if one filters out all tweets that can not be identified as being from a particular geographic region that is important. And researchers needs to clearly explain how any such constraints were implemented. If a language filter was used, precisely what was the filter? If a geographic constraint was imposed, what was it? Were only geo-coded tweets considered? Or was metadata about the source of the tweet considered, and if so how?

If tweets are aggregated by topic, the analyst must explain the aggregation method used. Or, the analyst must explain how tweets were assigned to topics: whether by topic modelling, or by human assignment. If by topic modelling, the analyst should provide information on how robust the results are to variations in the number of topics selected. Information about the criteria for tweets being assigned to topics (such as top terms from LDA analysis) is essential. And the analyst should indicated whether the topics were validated against human judgements.

If a collection of tweets claiming to measure public opinion excludes tweets *by some individuals*, that information is crucial and needs to be provided. Such exclusions could be based on the individual's frequency of social media use, on whether they are part of a set of individuals following particular political (or non-political actors). Such exclusion could also be based on the individual's characteristics: such as their use of language, or their demographic characteristics, or political characteristics. And as such characteristics are often estimated or inferred from meta-data, the analyst must be precise and transparent on how characteristics for individuals are inferred. Precise data on *who* is eligible to be included in the dataset is essential for any attempt to draw a population inference.

If a measure of sentiment is given, the analyst must carefully explain how sentiment was calculated. If a dictionary was used, the analyst should explain how robust the measure was to variations in the dictionary - or across dictionaries. And the analyst should explain if sentiment measures were validated in any way against human judgements.

We provide a set of guidelines in list form below.

1. Describe the source of data for the corpus.
   (a) Was data taken from the twitter firehose, or from one of the rate-limited sources?
   (b) Was data retrieved using the rest API or streaming API?

2. Describe whether the criteria for inclusion in the corpus was based on the text, the sender of the text, or contextual information (such as the place or time of the tweet).

3. For corpora collected based on keywords or regular expressions in the text, describe the criteria for inclusion.
   (a) What were the criteria by which the keywords or regular expressions were selected?
   (b) Were keywords or regular expressions chosen based on:
      - the analyst's expertise or prior beliefs?
      - an extant document?
      - an algorithm used to generate keywords based on an initial seeding and further processing of the text?

   (c) Describe any testing done to confirm that tweets gathered were relevant for the intended topic for which the keywords, or regular expressions, were used?

4. For corpora collected using *topics* as the criteria for inclusion:
   (a) Describe how individual documents were determined to be relevant to the chosen topic (i.e., what were the technical requirements for inclusion in the corpus).
   (b) Describe any testing done to estimate, or determine exactly, the number of documents in the corpora that were germane to their assigned topic(s).

5. If the content of tweets was *aggregated* by topic, after some selection criteria into the corpus, how were topics generated?
   (a) Were topics hand-coded?
   (b) Was some form of automated topic generation method used?
   (c) How robust are the topics to variations in the sample or the number of topics selected?
   (d) Was inclusion of text into topics validated against human judgments, and if so how?

6. Describe any limitations placed on the sampling frame that could limit **whose** opinions could be in the data. State any limitations of the sampling frame based on meta-data provided by informants, either *directly* provided in meta-data or *inferred* from meta-data. If characteristics were inferred, explain the procedure for inference. This would include the following:
   (a) Exclusions based on language *of the informant*.
   (b) Exclusions based on geography of the informant.

    (c) Exclusions based on gender, age, or other demographic characteristics of the informant.

7. For corpora in which selection was based on the sender how were the senders chosen?
8. Describe any constraints (or filters) imposed on the collection that would exclude some tweets from being included based on characteristics of the tweet, such as constraints on geography or language.
    (a) Describe whether any geographic constraints were based on geo-coding, or on an algorithim used to infer geography from meta-data.
    (b) Describe how language was determined if language constraints were imposed.
9. If a sentiment measure, or a related measure, was applied, describe how the measure calculated?
    (a) If a dictionary was used, describe how robust the measurewas to variations in the dictionary or across dictionaries.
    (b) Were the sentiment measures validated against human judgments?
    (c) What information other than the text of the tweet, such as linked content or images, characteristics of the sender, or context, was used to determine sentiment?
10. Describe the aggregation method used to generate the quantity of interest.
    (a) Describe precisely the temporal units used, including reference to time-zone.
    (b) Describe how retweets are treated.

For anyone interested in studying public opinion, it would be foolish to ignore the information about public opinion revealed by social media data. However, it would also be foolish to treat measurement of social media data in the same manner one treats a well designed survey yielding something approximating a random sample of a population of interest. We have listed many of the reasons that is not a viable strategy. Either one accepts that one has a non-representative opt-in sample, which may or may not be a useful sample for some goal *other than* measuring mass public opinion, or one attempts to weight the sample. We think continued work on studying public opinion via social media is a fruitful endeavour. And we urge scholars and practioners to both work on improving our ability to measure mass public opinion via social media, and to follow solid guidelines for reporting results obtained via social media.

References

AAPOR. 2010. "AAPOR REPORT ON ONLINE PANELS." http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.pdf?ijkey=0w3WetMtGItMuXs&keytype=ref.

Abraham, Katharine G., Sara Helms and Stanley Presser. 2009. "How Social Processes Distort Measurement: The Impact of Survey Nonresponse on Estimates of Volunteer Work in the United States." *American Journal of Sociology* 114(4):1129–1165.

Al Zamal, Faiyaz, Wendy Liu and Derek Ruths. 2012*a*. "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors." *ICWSM* 270.

Al Zamal, Faiyaz, Wendy Liu and Derek Ruths. 2012*b*. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *ICWSM*.

Ampofo, Lawrence, Nick Anstead and Ben O'Loughlin. 2011. "Trust, Confidence, and Credibility: Citizen Responses on Twitter to Opinion Polls During the 2010 UK general Election." *Information, Communication & Society* 14(6):850–871.

Aragón, Pablo, Yana Volkovich, David Laniado and Andreas Kaltenbrunner. 2016. When a Movement Becomes a Party: Computational Assessment of New Forms of Political Organization in Social Media. In *Tenth International AAAI Conference on Web and Social Media*.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* 23(1):76–91.

Barberá, Pablo and Gonzalo Rivero. 2014. "Understanding the Political Representativeness of Twitter Users." *Social Science Computer Review* .

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua Tucker and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* .

Barberá, Pablo, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker and Sandra González-Bailón. 2015. "The Critical Periphery in the Growth of Social Protests." *PloS one* 10(11):e0143611.

Barberá, Pablo, Richard Bonneau, Patrick Egan, John T. Jost, Jonathan Nagler and Joshua Tucker. 2014. "Leaders or Followers? Measuring Political Responsiveness in the US Congress using Social Media Data." Paper presented at the 110th American Political Science Association Annual Meeting.

Barracuda Labs. 2012. "The Twitter Underground Economy: A Blooming Business." Internet Security Blog, https://www.barracuda.com/blogs/labsblog?bid=2989.

Beauchamp, Nick. 2016. "Predicting and Interpolating State-level Polls using Twitter Textual Data." *American Journal of Political Science, forthcoming* .

Bermingham, Adam and Alan F Smeaton. 2011. "On Using Twitter to Monitor Political Sentiment and Predict Election Results." *Sentiment Analysis where AI meets Psychology (SAAIP)* p. 2.

Bode, Leticia and Kajsa E Dalrymple. 2014. "Politics in 140 characters or less: campaign communication, network interaction, and political participation on Twitter." *Journal of Political Marketing* (just-accepted).

Boutet, Antoine, Hyoungshick Kim and Eiko Yoneki. 2013. "What's in Twitter, I know What Parties are Popular and Who You are Supporting Now!" *Social Network Analysis and Mining* 3(4):1379–1391.

Brick, J. Michael and Douglas Williams. 2013. "Explaining Rising Nonresponse Rates in Cross-sectional Surveys." *The Annals of the American Academy of Political and Social Science* 645(1):36–59.

Bruns, Axel, Dr Dr Katrin Weller, Michael Zimmer and Nicholas John Proferes. 2014. "A topology of Twitter research: Disciplines, methods, and ethics." *Aslib Journal of Information Management* 66(3):250–261.

Castillo, Carlos, Marcelo Mendoza and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. ACM pp. 675–684.

Ceron, Andrea, Luigi Curini and Stefano M Iacus. 2015. "Using Sentiment Analysis to Monitor Electoral Campaigns Method Matters—Evidence From the United States and Italy." *Social Science Computer Review* 33(1):3–20.

Ceron, Andrea, Luigi Curini, Stefano M Iacus and Giuseppe Porro. 2014. "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France." *New Media &amp; Society* 16(2):340–358.

Choy, Murphy, Michelle Cheong, Ma Nang Laik and Koo Ping Shung. 2012. "US Presidential Election 2012 Prediction using Census Corrected Twitter Model." *arXiv preprint arXiv:1211.0938* .

Choy, Murphy, Michelle LF Cheong, Ma Nang Laik and Koo Ping Shung. 2011. "A Sentiment Analysis of Singapore Presidential Election 2011 using Twitter Data with Census Correction." *arXiv preprint arXiv:1108.5520* .

Cohen, Raviv and Derek Ruths. 2013. Classifying Political Orientation on Twitter: It's Not Easy! In *ICWSM*.

Compton, Ryan, David Jurgens and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE pp. 393–401.

Conover, Michael D, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini and Filippo Menczer. 2011. Predicting the Political Alignment of Twitter Users. In *Privacy, security, risk and trust (PASSAT), and 2011 IEEE third international conference on social computing (SocialCom)*. IEEE pp. 192–199.

De Leeuw, Edith and Wim De Heer. 2002. Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In *Survey Nonresponse*, ed. Robert M. Groves, Don A.

Dillman, John L. Eltinge and Roderick J. A. Little. New York: John Wiley& Sons pp. 41–54.

Diaz, Fernando, Michael Gamon, Jake Hofman, Emre Kiciman and David Rothschild. 2014. "Online and Social Media Data as a Flawed Continuous Panel Survey." Working paper.

Duggan, Maeve and Joanna Brenner. 2015. *The Demographics of Social Media Users, 2014*. Vol. 14 Pew Research Center's Internet & American Life Project Washington, DC.

Eichstaedt, Johannes C, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap et al. 2015. "Psychological language on twitter predicts county-level heart disease mortality." *Psychological science* 26(2):159–169.

Fang, Anjie, Iadh Ounis, Philip Habel and Craig Macdonald. 2015. "Topic-centric Classification of Twitter User's Political Orientation.".

Farrell, Henry and Daniel W Drezner. 2008. "The Power and Politics of Blogs." *Public Choice* 134(1-2):15–30.

Flicker, Sarah, Dave Haans and Harvey Skinner. 2004. "Ethical dilemmas in research on Internet communities." *Qualitative Health Research* 14(1):124–134.

Franco, Annie, Justin Grimmer and Monica Lee. 2016. "Changing the Subject to Build an Audience: How Elected Officials Affect Constituent Communication." Manuscript.

Gayo-Avello, Daniel. 2011. "Don't Turn Social Media into Another 'Literary Digest' Poll." *Communications of the ACM* 54(10):121–128.

Golbeck, Jennifer and Derek Hansen. 2011. Computing Political Preference among Twitter Followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM pp. 1105–1108.

Golder, Scott A. and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333(6051):1878–1881.

González-Bailón, Sandra and Georgios Paltoglou. 2015. "Signals of Public Opinion in Online Communication A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659(1):95–107.

González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer and Yamir Moreno. 2014. "Assessing the bias in samples of large online networks." *Social Networks* 38:16–27.

Groves, Robert. 2011. ""Designed Data" and "Organic Data".".
URL: *http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/*

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70(5):646–675.

Groves, Robert M. and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: a Meta-analysis." *Public Opinion Quarterly* 72(2):167–189.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer and

Roger Tourangeau. 2011. *Survey methodology.* Vol. 561 John Wiley &amp; Sons.

Gruzd, Anatoliy and Caroline Haythornthwaite. 2013. "Enabling Community Through Social Media." *Journal of Medical Internet Research* 15(10).

Hampton, Keith, Lauren Sessions Goulet, Lee Rainie and Kristen Purcell. 2011. "Social networking sites and our lives." Pew Internet & American Life Project Report.

He, Ran and David Rothschild. 2014. "Who are People Talking about on Twitter?" Working paper.

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.* ACM pp. 237–246.

Hillygus, D. Sunshine. 2011. The Practice of Survey Research: Changes and Challenges. In *New Directions in Public Opinion*, ed. Adam Berinsky. Routledge Press.

Joinson, Adam. 1999. "Social Desirability, Anonymity, and Internet-based Questionnaires." *Behavior Research Methods, Instruments, & Computers* 31(3):433–438.

Jungherr, Andreas. 2014. "Twitter in politics: a comprehensive literature review." *Available at SSRN 2402443* .

Jungherr, Andreas, Harald Schoen, Oliver Posegga and Pascal Jürgens. 2016. "Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support." *Social Science Computer Review* .

Jungherr, Andreas, Pascal Jürgens and Harald Schoen. 2012. "Why the Pirate Party Won the German Election of 2009 or the Trouble with Predictions: A Tesponse to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment"." *Social Science Computer Review* 30(2):229–234.

Keeter, Scott. 2012. "Presidential Address: Survey Research, its New Frontiers, and Democracy." *Public Opinion Quarterly* 76(3):600–608.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2016. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." Manuscript.

King, Gary, Patrick Lam and Margaret Roberts. 2014. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *Copy at http://j. mp/1qdVqhx Export BibTex Tagged XML Download Paper* 456.

Kreiss, Daniel. 2014. "Seizing the Moment: The Presidential Campaigns' Use of Twitter During the 2012 Electoral Cycle." *New Media & Society* .

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.* ACM pp. 591–600.

Larson, Jennifer, Jonathan Nagler, Jonathan Ronen and Joshua A Tucker. 2016. "Social networks and Protest Participation: Evidence from 93 Million Twitter Users." *SSRN* .

Lax, Jeffrey R and Justin H Phillips. 2009. "How Should we Estimate Public Opinion in the States?"

American Journal of Political Science 53(1):107–121.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343(14 March):1203–1205.

Leetaru, Kalev, Shaowen Wang, Guofeng Cao, Anand Padmanabhan and Eric Shook. 2013. “Mapping the global Twitter heartbeat: The geography of Twitter.” *First Monday* 18(5).

Lin, Yu-Ru, Drew Margolin, Brian Keegan and David Lazer. 2013. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 737–748.

Little, Roderick JA. 1993. “Post-Stratification: a Modeler’s Perspective.” *Journal of the American Statistical Association* 88(423):1001–1012.

Malik, Momin M, Hemank Lamba, Constantine Nakos and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. In *Ninth International AAAI Conference on Web and Social Media*.

Marwick, Alice E. and Danah Boyd. 2011. “I Tweet Honestly, I Tweet Passionately: Twitter Users, Context collapse, and the Imagined Audience.” *New Media & Society* 13(1):114–133.

Metaxas, Panagiotis Takis, Eni Mustafaraj and Daniel Gayo-Avello. 2011. How (Not) to Predict Elections. In *Privacy, security, risk and trust (PASSAT), and 2011 IEEE third international conference on social computing (SocialCom)*. IEEE pp. 165–171.

Metzger, Megan, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker. 2016. “Tweeting Identity? Ukranian, Russian, and #Euromaidan.” *Journal of Comparative Economics* 44(1):16–50.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J Niels Rosenquist. 2011. “Understanding the Demographics of Twitter Users.” *ICWSM* 11:5th.

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang and Alessandro Vespignani. 2013. “The twitter of babel: Mapping world languages through microblogging platforms.” *PloS one* 8(4):e61981.

Mokrzycki, Michael, Scott Keeter and Courtney Kennedy. 2009. “Cell-phone-only Voters in the 2008 Exit Poll and Implications for Future Noncoverage Bias.” *Public Opinion Quarterly* 73(5):845–865.

Morris, Meredith Ringel, Scott Counts, Asta Roseway, Aaron Hoff and Julia Schwarz. 2012. Tweeting is Believing? Understanding Microblog Credibility Perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM pp. 441–450.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. In *ICWSM*.

Munger, Kevin. 2015. “Elites Tweet to get Feet off the Streets: Measuring Elite Reaction to Protest Using Social Media.”.

Mustafaraj, Eni and Panagiotis Metaxas. 2010. “From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search.” Paper presented at WebSci10: Extending the Frontiers of

Society On-Line, April 26-27th, 2010, Raleigh, NC.

Mustafaraj, Eni, Samantha Finn, Carolyn Whitlock and Panagiotis Takis Metaxas. 2011. Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail. In *SocialCom/PASSAT*. IEEE pp. 103–110.

Neuman, W. Russell, Lauren Guggenheim, S. Mo Jang and Soo Young Bae. 2014. "The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data." *Journal of Communication* 64(2):193–214.

Newman, Mark W., Debra Lauterbach, Sean A. Munson, Paul Resnick and Margaret E. Morris. 2011. It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health. In *Proceedings of the ACM 2011 Conference on Computer-Supported Cooperative Work*. ACM pp. 341–350.

Nexgate. 2013. "2013 State of Social Media Spam." Nexgate Report, http://go.nexgate.com/nexgate-social-media-spam-research-report.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R Routledge and Noah A Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *ICWSM* 11:122–129.

Park, David K, Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Pavalanathan, Umashanthi and Munmun De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee pp. 315–321.

Pennacchiotti, Marco and Ana-Maria Popescu. 2011. "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11:281–288.

Pew Research Center. 2012. "Assessing the Representativeness of Public Opinion Surveys." Report, For the People & the Press, .

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*. ACM pp. 1025–1030.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Rao, Delip, David Yarowsky, Abhishek Shreevats and Manaswi Gupta. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. ACM pp. 37–44.

Ratkiewicz, Jacob, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini and Filippo Menczer. 2011. Detecting and Tracking Political Abuse in Social Media. In *ICWSM*. pp. 297–304.

Richman, Wendy L., Sara Kiesler, Suzanne Weisband and Fritz Drasgow. 1999. "A Meta-analytic Study of Social Desirability Distortion in Computer-administered Questionnaires, Traditional Questionnaires, and Interviews." *Journal of Applied Psychology* 84(5):754.

Sanovich, S. 2015. "Government Response Online: New Classification with Application to Russia." Unpublished Manuscript, New York University.

Skoric, Marko, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim and Jing Jiang. 2012. Tweets and Votes: A Study of the 2011 Singapore General Election. In *System Science (HICSS), 2012 45th Hawaii International Conference on Systems Science (HICSS-45 2012)*. IEEE pp. 2583–2591.

Solberg, Lauren B. 2010. "Data mining on Facebook: A free space for researchers or an IRB nightmare?" *Journal of Law, Technology and Policy* 2010(2).

Thomas, Kurt, Chris Grier and Vern Paxson. 2012. Adapting social spam infrastructure for political censorship. In *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*. USENIX Association pp. 13–13.

Tourangeau, Roger, Robert M. Groves and Cleo D. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link Between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74(3):413–432.

Tucker, Joshua A., Jonathan Nagler, Megan MacDuffee Metzger, Pablo Barberá, Duncan Penfold-Brown and Richard Bonneau. 2016. Big Data, Social Media, and Protest: Foundations for a Research Agenda. In *Computational Social Science: Discovery and Prediction*, ed. R. Michael Alvarez. Cambridge University Press chapter Big Data, Social Media, and Protest: Foundations for a Research Agenda, pp. 199–224.

Tufekci, Zeynep and Christopher Wilson. 2012. "Social media and the decision to participate in political protest: Observations from Tahrir Square." *Journal of Communication* 62(2):363–379.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G Sandner and Isabell M Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10:178–185.

Tuunainen, Virpi Kristiina, Olli Pitkänen and Marjaana Hovi. 2009. "Users' awareness of privacy on online social networking sites-case Facebook." *Bled 2009 Proceedings* p. 42.

Vaccari, Cristian, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T Jost, Jonathan Nagler and Joshua Tucker. 2013. "Social Media and Political Communication. A survey of Twitter Users during the 2013 Italian General Election." *Rivista Italiana di Scienza Politica* 43(3):381–410.

Veenstra, Aaron, NNarayanan Iyer, Namrata Bansal, Mohammad Hossain and Jiachun Park, Jiwoo & Hong. 2014. "#Forward!: Twitter as Citizen Journalism in the Wisconsin Labor Protests." Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication. St. Louis, MO.

Wang, Wei, David Rothschild, Sharad Goel and Andrew Gelman. 2014. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting. Forthcoming* .

Wong, Felix Ming Fai, Chee Wei Tan, Soumya Sen and Mung Chiang. 2013. Quantifying Political Leaning from Tweets and Retweets. In *ICWSM*. pp. 640–649.

Wu, Shaomei, Jake M Hofman, Winter A Mason and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM pp. 705–714.

Zimmer, Michael. 2010. ““But the data is already public”: on the ethics of research in Facebook.” *Ethics and information technology* 12(4):313–325.