Addressing Selection Bias in Event Studies with General Purpose Social Media Panels

HAN ZHANG, Princeton University
SHAWNDRA HILL and DAVID ROTHSCHILD, Microsoft Research

Data from Twitter have been employed in prior research to study the impacts of events. Conventionally, researchers use keyword-based samples of tweets to create a panel of Twitter users who mention event-related keywords during and after an event. However, the keyword-based sampling is limited in its *objectivity* dimension of data and information quality. First, the technique suffers from selection bias since users who discuss an event are already more likely to discuss event-related topics beforehand. Second, there are no viable control groups for comparison to a keyword-based sample of Twitter users. We propose an alternative sampling approach to construct panels of users defined by their geolocation. *Geolocated panels* are exogenous to the keywords in users' tweets, resulting in less selection bias than the keyword panel method. *Geolocated panels* allow us to follow within-person changes over time and enable the creation of comparison groups. We compare different panels in two real-world settings: response to mass shootings and TV advertising. We first show the strength of the selection biases of keyword-panels. Then, we empirically illustrate how geolocated panels reduce selection biases and allow meaningful comparison groups regarding the impact of the studied events. We are the first to provide a clear, empirical example of how a better panel-selection design, based on an exogenous variable such as geography, both reduces selection bias compared to the current state of the art and increases the value of Twitter research for studying events. While we advocate for the use of geolocated panels, we also discuss its weaknesses and application scenario seriously. This paper also calls attention to the importance of selection bias in impacting the objectivity of social media data.

CCS Concepts: •**Applied computing → Economics; Sociology;** *Marketing;* •**Information systems →** *Document filtering;*

Additional Key Words and Phrases: Twitter, social media, survey, panels, geolocation, selection bias, non-response bias, coverage bias

ACM Reference format:
Han Zhang, Shawndra Hill, and David Rothschild. 2016. Addressing Selection Bias in Event Studies with General Purpose Social Media Panels. *ACM J. Data Inform. Quality* 1, 1, Article 1 (January 2016), 24 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Recently, there has been a great deal of interest in how social media data can be used ex-post to answer questions about the influence of major events that cannot be captured ex-ante. Specifically, there have been many attempts to use Twitter to understand the impact of an event, concerning both interest and sentiment, after it occurs. Compared with traditional surveys, there are several advantages of using Twitter data (and social media data in general) to study events: low time latency, high time granularity, low financial cost, and large sample size. More importantly, Twitter

DOI: 0000001.0000001

also offers easy methods for gathering pre-event data. It is especially difficult to use survey methods to study rare events; it is hard to capture relevant respondents randomly, and retrospective polling is notoriously unreliable [13, 23]. Studies that have used Twitter to study events have attempted to answer questions that include how political protests impact political attitudes [4, 66], how pandemics raise public concern [55], how natural disasters give rise to sadness and anxiety [15], and how opinions about elections change[46]. Twitter can also be used to study advertisements, which can be considered a special kind of event.

Most scholars do not work with full Twitter datasets when studying events. Instead, they work with a sample of users or tweets selected using various sampling strategies. When considering Twitter samples, one question arises: whether the sampling procedures can provide objective data for event analytics. As survey research have known for a long time, even if the individual piece of information is accurate, different sampling strategies can result in different types of sampled respondents. Such sampling biases become even more acute regarding social media data because the accuracy of social media data itself is questionable.

Under the Total Data Quality Management framework (TDQM), the dominant framework of understanding data and information quality, data qualities related to sampling strategies of Twitter belongs to the dimension of *objectivity*. Objectivity refers to the degree that data are selected objectively and can be used without judgment in the process of creating the data [62]. Objectivity differs from the accuracy concerns of social media data, such as social misinformation, bots and censorship: the latter is related to whether each piece of individual information is correct comparing with the baseline values, while the former is related to whether samples of individual information can be used as an objective representation of the entire population. Objectivity impacts the believability of social data: a sample of tweets without elaboration on the process that created the data will have low believability. Objectivity stands as a dimension in its own right.

Despite the importance of objectivity, little work has been done to systematize the sampling strategies used to construct study events. In the following, we offer a categorization of sample selection methods used in Twitter event analytics. The majority of prior studies share a *keyword*-based sampling strategy. They use event-related keywords or hashtags to filter a cross-section of tweets that are used to describe the tweet-level conversation around a specific topic at a particular point in time. We call this the *keyword-based cross-section* method:

(1) Scholars use the Twitter search or streaming API to collect tweets mentioning keywords or hashtags that are directly relevant to the event(s) being examined. For instance, to study the Gezi protests, tweets mentioning related hashtags such as #occupygezi are chosen; to study the impact of the US presidential election, tweets mentioning "Obama" or "McCain" are chosen; to study the impact of influenza, tweets mentioning terms such as "H1N1" are chosen [4, 13, 22, 55].
(2) Scholars then analyze the change of counts, relative proportions, and sentiment of these tweets *caused* by events.

Only recently have researchers begun to extend tweet-centric cross-sections to user-centric panels, whose unit of analysis is users, instead of tweets. The major rationale for this shift is that tweet-level analysis does not consider the factor that some core discussants on Twitter are much more active tweeters than other non-active users, such that discussion trends are dominated by these active users. A user-level analysis, however, can incorporate users' socio-demographic characteristics as well as their engagement level on social media, thus providing a more accurate estimate of the impact of events.

A natural extension of tweet-level analysis is the expansion of a *keyword-based cross-section* to a *keyword-based panel* by collecting the historical tweets and characteristics of users that appear in the keyword-based cross-section sample. Shifting from *keyword cross-sections* to *keyword panels* also makes the sample selection process comparable to the classical survey framework. In an ideal classical social survey, scholars begin at the user level (instead of the tweet level) and determine a sample of users. They then ask the selected users attitudinal questions (cross-sectional survey), possibly tracking the same users over a period (panel survey). Under the perspective survey perspective, there are evident problems related to keyword selection using tweets rather than users as the main unit of analysis[13]. First, keyword cross-sections fluctuate in both who and how many people are responding, making it difficult to determine the meaning of fluctuation of interest and sentiment (i.e., does it indicate a shift in the interest of users or is it due to the selected topic?). Second, keyword cross-sections do not track each user's opinions over time to show within-person changes. *Keyword panels* are better at this than keyword cross-sections because they have a fixed population and attempt to track the population before and after events. The second advantage of *keyword panels* is that they can efficiently extract users of interest for a specific event.

Despite their advantages compared to the keyword cross-section method, we argue that there remain two critical problems with *keyword-based panels*: 1) selection bias regarding users and their tweets and 2) lack of control groups for comparison. These two biases are inherent in keyword-based cross-sections and panels, which rely on users' textual content to build the study population.

First, keyword-based sample construction (both *keyword-based panels* and *keyword-based cross-sections*) introduces selection bias. Event-related keywords are used to filter Twitter users who have discussed the event of study. The danger is that users who respond to events are usually already more interested in event-related topics beforehand than random users. Hence, it is unclear whether impacts identified using a *keyword-based panel* are truly driven by the event(s) or merely reflect self-selected engagement during events by an unrepresentative sample of Twitter users. Furthermore, users who are more likely to mention a certain event may also be systematically different regarding demographic characteristics than those who do not mention the event. As we show later, individuals who mention the terms “shooting” or “Xbox” are much more likely to mention exact terms and related keywords even before the triggering event (i.e., a mass shooting or Xbox advertisement) and more likely to be male than randomly selected users.

Second, to rigorously measure the impact of an event on users, scholars need a “control” group that is not systematically different from the exposed users, and could have been exposed to the event but not by chance. By comparing the study sample with a control group, we can ensure that changes are not driven by confounding trends. When using the *keyword-based panel,* method, all sampled users are, by definition, already affected by the event. Consequently, this method cannot offer the possibility of creating objective control groups. The control group here is denoted in a counterfactual way as it is defined in the causal inference literature [44]. Accordingly, random samples cannot typically serve as control groups for causal inference since they do not reveal how the impacted people would behave were they not impacted by an event [44].

There are scattered solutions in the literature that attempt to reduce the selection bias introduced by keyword panels [11, 66]. However, the literature lacks a systematic treatment of the sample methods scholars have used. Below, we summarize two additional types of panels that have been used.

Random panel: The *random panel* is based on a random sample of users. Random panels in Twitter are analogous to random samples in survey research. When possible, a probability-based random sample is an efficient alternative to a population census. In most cases, random panels serve as useful baselines in practice. However, *random panels* have two weaknesses. Theory on sampling for social media is yet to be developed while theory on probability population surveys is mature. Furthermore, random panels are often an inefficient way to study events since a sufficiently large sample must be collected to find users who have discussed a given topic.

Geolocated panel: To create a geolocated panel, scholars use geolocation information of tweets to collect a list of users who were *close* to an event regarding time and space and were thus likely to be "exposed" to the actual event instead of self-selecting themselves into the panel. Compared with the *keyword-based panel*, the *geolocated panel* can reduce selection bias since users do not actively select themselves into the panel; compared with the *random panel*, the *geolocated panel* provides more coverage of relevant users. However, geolocated panels may not be useful for some events whose effects unfold mainly in a non-spatial manner. A section later in the paper discusses when scholars should consider geolocated panels over other panels.

In the remainder of the paper, we first define the concept of selection bias, build its connection with data and information quality, categorize three types of Twitter panels and describe their selection biases (Section 2). We then empirically show that *keyword panels* indeed exhibit strong selection biases, while *random* and *geolocated panels* can reduce selection biases (Section 3). Following this, we show how *geolocated* and *random* panels can give different estimations of the impacts of two very different sets of events—mass shootings and TV advertisements—compared with *keyword* panels (Section 4). We then discuss the application scenario of geolocated panels, detail the steps to construct them, and summarize the strengths and weaknesses of each type of panels. The final section provides a conclusion and a discussion of future directions for research.

This paper makes four major contributions. First, we summarize previous research and categorize three Twitter panels that have been used to study events, borrowing knowledge from survey research. Second, raise the importance of focusing on selection bias in user content, which is often more difficult to reduce. Third, we empirically compare the strength of selection bias in each type of panel, using multiple types of events. Last, while scholars, practitioners, and industry have widely used social media in event analytics, very few have utilized the existing research in data and information quality theories. We link the selection biases caused by sampling in social media with the objectivity concerns under the TDQM framework in data and information quality research. This paper thus explores how research frameworks of data and information quality can be used to address the data challenges of social media.

2 CATEGORIZING DATA COLLECTION METHODS OF EVENT ANALYTICS WITH SOCIAL MEDIA DATA

The use of social media, especially Twitter, has been a recent trend in event analytics. Twitter offers critical advantages to survey methods when studying unexpected events. Besides well-known advantages concerning cost and speed, Twitter can track the discussion trends of an unexpected event before, during, and after the event, while conventional surveys cannot begin before an event. Twitter thus has advantages over surveys for dynamically tracking the impacts of events. Twitter-like social media platforms have been widely used in various event analytics, including 1) event detection [10, 42, 50]; 2) prediction such as for worldwide election outcomes [14, 46, 56]; and 3) measurement of the impacts of events, including public concern about the influence of [55] and responses to political protests[4]. In this paper, we focus on the third line of research: measuring the impacts of events using Twitter data. Specifically, we restrict our research to scholarship that uses *textual* information of tweets as outcome measures.

Despite the wide use in event analytics, Twitter is also known for its data quality issues. While previous studies more or less acknowledge that social media data are not perfect in measurements, research frameworks in data and quality have just been applied to understand biases in social media data. Agarwal and Sureka [1] is an early study that documents how social media data will affect the 16 dimensions of the TDQM framework. They find that three unique features of microblogs such as Twitter—spams, colloquial usage and misspelling, and constant flow of information—increase timeliness and conciseness, but decrease qualities of all other dimensions of data quality. Shankaranarayanan and Blake [54] argue that accuracy, consistency, timeliness, and completeness, which are traditionally important, will have declining importance in relation to social media data. On the other hand, believability was barely studied previously but have increasing dimensions in the age of social media. Emamjome et al. also suggest that information quality in the context of social media is different compared to traditional IQ in information systems. They propose an IQnSM model that is built upon the TDQM framework but contains revised dimensions of information quality for analyzing social media data [18].

In this paper, we extend previous discussions by making the connection between selection biases in Twitter samples and data and information quality. We first define selection bias, argue that sampling Tweets or users create selection biases, and suggest that such biases should be put into the category of objectivity dimension of data and information quality. We then offer a categorization of the data collection procedures of previous Twitter-based event studies. We believe that our categorization illustrates the strengths and weaknesses of each type of Twitter data collection method.

2.1 Selection bias and its relation to data and information quality

Selection bias is well-known as a major threat to causal inference in empirical social science research. Selection bias is the bias introduced by the selection of individuals, groups, or data for analysis in such a way that scholars do not have control over the randomized selection of individuals into their study, as individuals self-select themselves. Addressing and reducing selection bias have also become major requirements for social science research using quantitative approaches (to name a few, economics, political science, and sociology, see [3, 34]). In event studies using social media data, however, researchers have just begun to realize the importance of recognizing and evaluating selection bias. For instance, Tufekci [61] showed the danger of deleting hashtags during major events. Lin et al. [40] compared users selected from keyword sampling with a focus group and noted considerable differences between the two. Culotta [11] clearly mentioned selection bias in keyword-centric design and argued that geolocation can be used to address this problem. Other approaches have used experimental, quasi-experimental, or matching methods to correct selection bias [36, 47, 53], highlighting the ubiquity of concerns about data and information quality across fields.

While the aforementioned studies mention selection biases and how they impact data and information quality at a conceptual level, they have yet to fully incorporate the rich studies of data and information quality. We use the TDQM framework, developed by Wang and Strong [62], to understand selection biases. TDQM divides data and information quality into four large categories: intrinsic, contextual, representational, and accessibility. Wang and Strong argue that the intrinsic category contains arguably the essential aspects of data and information quality. They summarize four dimensions of the intrinsic category: accuracy, believability, objectivity, and reputation.

We propose that selection biases belong to the *intrinsic* category and it addresses the dimension of *objectivity* of data and information quality. Objectivity refers to the process in which data are generated. As Wang and Strong nicely put it, objective data are generated through unbiased and impartial ways, while unobjective data have judgment in the process of creating the data [62]. Selection biases exactly address this problem: what matters for selection biases is the criterion scholar select tweets or users in the study. Objectivity is not the same as accuracy. Even if tweets are perfectly accurate (which is itself impossible), different sample strategies will create different samples of tweets and give different estimates of the impact of events, as we will show later in this paper. On the other hand, objectivity impacts the believability, while the latter focus more on the subjective evaluation of data quality. In general, as long as users generate data, and scholars do not have control over 1) whether users can opt-out or opt-in to contribute to the data and 2) what types of data opted-in users report, selection bias will persist.

Selection bias is not limited to social media data; it prevails in any social data, in which unit values are reported or generated by people, instead of objectively measured. For instance, even in well-designed surveys, non-responses and misreporting can still introduce selection biases. Yet, selection biases in social media worth more attention, because current research focuses dominantly on the accuracy problem of social media. In next several sections, we summarize existing sampling approaches in Twitter. For each approach, we also discuss its's selection biases and practical steps for data collection.

2.2 Keyword-based cross-sections

As defined in the introduction, *keyword cross-sections* are collections of tweets that contain event-related keywords within the desired time frame. Scholars use *keyword cross-sections* to analyze discussion trends surrounding certain keywords or topics. For example, Thelwall, Buckley, and Paltoglou [59] analyzed sentiments of tweets containing event-related hashtags for top events mined from Twitter. Lehmann, Goncalves, Ramasco, and Cattuto [38] identified events whose related hashtags had a sudden spike in use and analyzed the context of tweets containing these hashtags. Tsytsarau, Palpanas, and Castellanos [60] examined how news events, as revealed by online memes, trigger social media attention by selecting tweets containing keywords related to specific news events. Importantly, all these studies, among others, used keywords that were directly relevant to particular events. Hence, the selecting-on-keyword bias could be significant.

From an engineering perspective, *keyword cross-sections* are the easiest type to collect. Scholars must follow the stream of public tweets on Twitter using the standard Twitter Stream API[1] or search for certain keywords that are relevant to the event using the Twitter Search API[2]. Most of the time, scholars do not collect further data about users who posted the tweets, such as whether they posted something similar before the event (which would require collecting a user's historical tweets).

2.3 Keyword panels: from tweet-centric to user-centric.

Despite the intuitive interpretation and simple implementation of Twitter *keyword cross-sections*, they have a major disadvantage, namely that they are tweet-centric instead of user-centric. Each user makes drastically different contributions to the discussion of an event. A shift in the trend of discussion about a certain topic may be caused by several enthusiastic users, and the use of a *keyword cross-section* wrongly regards this as reflecting a true shift in the interest of a more general population. For instance, when using *keyword-based cross-sections* to study outbreaks of disease, Kanhabua and Nejdl [30] noted that they are very sensitive to random turbulence in users' discussion levels.

Recently, Diaz, Gamon, Hofman, Kcman, and Rothschild [13] suggested viewing the entire Twitter platform as if it were produced by a hypothetical pseudo-survey, in which users opt in to answer pseudo-questions on topics in which they are interested and can answer multiple times. We call this approach *keyword panels*. We follow the conventional use of the word "panel" in social science survey research: to keep track of a set of stable users over certain periods, ideally

[1] https://dev.twitter.com/streaming/public

[2] https://dev.twitter.com/rest/public/search

before, during, and after an event of interest. *Keyword panels* offer two advantages over *keyword cross-sections*: 1) they are user-centric and 2) they contain data on the ex-ante behaviors of users. Because they are user-centric, scholars can follow a stable set of users over time, which is prone to random drifting in terms of user participation. For instance, Weber, Garimella, and Batayneh[64] and Budak and Watts [4] analyzed a fixed set of users' political behaviors before and after protests in Egypt and Turkey using a panel of users who mentioned protest-related hashtags. When predicting election outcomes, scholars found that counting one vote for each user who mentioned a party or candidate often outperformed the approach of counting one vote per tweet [17, 51]. Outside the election prediction realm, An and Weber [2] confirmed that user-centric analysis outperformed counting tweet occurrences for predicting both flu activity and unemployment rates. In addition, Lin, Margolin, Keegan, and Lazer [40] explicitly addressed the shortcomings of using large volumes of *keyword cross-section* data without tracing the history or context of the individuals generating the tweets.

Furthermore, with the ex-ante behaviors of users, scholars can construct pre-event features of users, which can help to correct biases in keyword-based panels. User features that can be inferred from profiles and historical tweets include demographics [20, 37], geography [12], and interest in discussing the event or related topics [2, 9, 63]. For instance, Wang, Rothschild, Goel, and Gelman [63] used post-stratification techniques to adjust the unbalanced demographic composition in the 2012 US presidential election; their results are comparable to those of representative polls. De Choudhury, Diakopoulos, and Naaman [12] adjusted the types of participants in events on Twitter, building a classifier to differentiate between organizations, journalists and media bloggers, and ordinary individuals.

From an engineering perspective, *keyword panels* require an extra step beyond *keyword cross-sections*. One must first create a *keyword cross-section* and then follow tweets in the *keyword cross-section* to obtain profiles and historical tweets of the users who posted the tweets. The extra step, however, significantly helps scholars to estimate the impact of events more accurately.

Keyword cross-sections and panels are the most popular method in event analytics using social media data so far, as they both rely on keyword filtering. There have been recent attempts to collect user-centric data directly from user characteristics, such as demographics and geolocations, which we discuss below.

2.4 Panels based on demographics

Panels based on demographics are similar to *keyword panels* in the sense that both are user-centric and follow users over a certain period. The difference lies in how data are collected: Panels based on demographics search for results based on certain demographics traits, while *keyword panels* filter users based on keyword searches. Unfortunately, Twitter does not offer an API that allows direct selection of individuals with certain traits, such as being female. However, third-party tools such as Followerwonk allow scholars to search for users of interest and then collect the tweets of these users[3]. Scholars can search for users with interests in specific topics such as the use of certain music services or religious affiliations [7, 49]. Panels based on demographics are thus user-centric panels that are collected based on user characteristics.

2.5 Geolocated panels

Geolocated panels are a special type of panel based on demographics, with geolocation being the demographic trait. Prior research has used Twitter samples based on geolocation [27, 33, 48]. For example, Zhang [66] used the check-in history of Weibo (Chinese Twitter) to find users near

[3] https://moz.com/followerwonk/bio

protests, collect their tweet history, and compare changes in political discussion at the user level before and after the protests. Most prior work has aimed simply at reconstructing a representative panel of users based on their geolocations to study users in a particular location for a particular subject. Culotta [11], on the other hand, attempted to reduce selection bias by searching for users who had geolocated tweets in 100 US counties and evaluated how the proportion of users who mentioned health-related keywords within each county correlated with the offline health behaviors of users by geography. However, the authors did not address the potential for geolocated panels to enable the construction of comparison groups. Our paper extends the current literature by addressing this gap, which is one of our contributions to information and data quality research.

From an engineering perspective, *geolocated panels* are the easiest to create among all panels based on demographics, since scholars can collect geolocated users through a geolocation search using Twitter API or Weibo. In other words, they can search for users who had a check-in proximate to a certain place.

3 EVENTS AND DATA

Most previous research using Twitter to understand the impact of events has focused on a specific type of event (or sometimes even a single event). In our empirical analysis, for each type of panel, we evaluate Twitter's effectiveness in studying two very different types of events: the 15 largest mass shootings in the United States in 2014 and a mix of local and national Xbox advertisements aired in 2014. Both these event types are exogenous to users in terms of time and location: users know these types of events exist but cannot predict the exact time and location before the event occurs. This is exactly the type of event for which social media is better suited than traditional survey data. However, the events differ in the ability of researchers to assign treatments. For mass shootings, researchers typically cannot predict when and where an event will occur. Therefore, it is nearly impossible to perform surveys on the ex-ante behavior of users before the occurrence of an event. Other unpredicted events, such as terrorist attacks and natural crises, also fall into this category. In this case, scholars can treat an unexpected event as a natural experiment to evaluate the effect of the event [16]. On the other hand, when and where an advertisement will occur are known to scholars. Scholars thus have more power to manipulate treatment assignment to understand the causal impact of events.

Next, we describe how we constructed the three types of panels for all the events in our study. We used all mass shootings in 2014 documented in the Stanford Mass Shootings of America (MSA) data project[4]. The Xbox advertisements included a batch of local ads that were all aired at 2:22 PM ET in 14 market areas on January 12, 2014, as well as a national ad that aired later the same day.

For each event, we accessed the full Twitter stream via the Twitter Firehose provided by Microsoft. Twitter Firehose is a Twitter API that delivers full access to the public stream of tweets[5]. The use of Twitter Firehose is necessary for our purposes since our aim is to *methodologically evaluate* the selection bias associated with Twitter. The strength of this bias can be used in future research when considering sample selection procedures. Only by using full Twitter data can we ensure that the selection bias detected in this paper is not due to sampling bias from Twitter API [45]. In other words, our results provide a lower bound; if the selection bias in our paper is significant, it may be even larger when scholars do not have full access to Twitter data through the Firehose API.

For practical users of Twitter panels, however, we emphasize that scholars *do not* need Twitter Firehose access; they can construct their own panels through the streaming or geolocation search API of Twitter. The streaming API provides a random sample of the full Twitter data. Therefore,

[4] Data can be downloaded from the following site: https://library.stanford.edu/projects/mass-shootings-america
[5] http://support.gnip.com/apis/firehose

estimations of quantities such as group proportions and sizes can be obtained, following the theories of random sampling [52].

For each mass shooting, we used eight words related to "shooting"–attack, cop, jail, kill, murder, shot, Trayvon (for the Trayvon Martin case), and shooting –to build eight stylized keyword panels. For instance, we constructed the *keyword-based panel* for the word "shooting" by finding all the tweets that mentioned the word "shooting" within seven days after each shooting and collecting historical tweets of the users posting these tweets within seven days before the event. Similarly, for each Xbox advertisement, we constructed three panels, each using one of the keywords Xbox, PlayStation, or PS3. For instance, for the Xbox *keyword-based panel*, we selected users who mentioned "Xbox" within seven days of the advertisement and collected their historical tweets for seven days before the advertisement. Based on prior research, we know that online response to TV advertisements on Twitter takes place within minutes as opposed to days [35]. With access to the Twitter Firehose, we were able to find all tweets that mentioned the keywords of interest. Each *keyword-based panel* is a full census of users who mentioned the keywords within the designated time frame. We used one word to construct each panel, instead of a set of keywords, which is more common in the literature, to ensure that the effect is not confounded by other keywords in the keyword set.

For comparison, we created *random panels* for each event. Again, we used Twitter Firehose and selected 30,000 users at random who tweeted at least once during the specified time frame after each event. We then collected all these users' tweets for the same designated time frame as the keyword panels (seven days)[6].

Finally, we constructed *geolocated panels* for the mass shooting and Xbox advertisement events. We identified geolocated users who had at least five geolocated tweets within the United States in 2014 and used this as the population of geolocated users. We then projected their geolocated posts over the course of the year onto census tracts and used the two most frequent census tracts as the frequent locations. For the mass shooting events, we randomly sampled users whose two most frequent locations were within 100 miles of a particular shooting. For Xbox ads, we randomly sampled users whose most frequent location was within 100 miles of the center of a Designated Market Area (DMA). Each DMA is an "exclusive geographic area of counties in which the home market television stations hold a dominance of total hours viewed"[7]. In other words, broadcasting companies choose to air the same set of programs within the same DMA. Some DMAs span the boundaries of counties. By drawing a circle around a DMA, we created an area in which we could find two groups of users who have a similar distance to the center of the DMA. However, one group could see some advertisements since they were within the DMA region while the group living outside the DMA could not see the advertisements [31]. This method was used to create a comparison group.

The choice of 100 miles in both of our examples reflects the appropriate distance for each problem. In the mass shooting example, we found that using a range of more than 100 miles did not impact the pattern of our outcomes (see later discussion for details). In the Xbox example, all DMAs were covered within a 100-mile radius from the center of each DMA. For other types of problems, researchers should choose a distance range to reflect their knowledge about how far an event can exert influence.

[6] We initially choose 30,000 since it is a convenience number. We tested the statistical power of using this number. We find that a sample size of 30,000 can distinguish the observed proportion of 0.01 of the random panels and the observed proportion of 0.02 of the geolocated panels at the 0.01 level, with 99.9% probability. The power analysis is done with 2-side t-tests and using R package "pwr".

[7] http://www.nielsenmedia.com/glossary

In next section, using the events and panels constructed as described above, we empirically show the selection bias that is embedded in *keyword panels*. Following this, we empirically compare three types of panels for evaluating the impacts of mass shootings and advertisements and show how geolocated panels can improve the determination of the estimated impact of events. Finally, we discuss steps to construct *geolocated panels*, which can help reduce selection bias in certain scenarios.

4 SELECTION BIAS IN TWITTER PANELS

4.1 Selection Bias in Keyword Panels

As briefly discussed in the introduction, the *keyword-based panel* method introduces selection bias into a sample. In this section, we explicate types of selection bias in three *keyword panels* and empirically reveal the strength of the selection bias in the panels. We differentiate between three aspects of selection bias introduced in the *keyword-based panel* design as follows:

- Selection-on-outcome bias: Users have different probabilities to be selected into the study population based on outcomes. One of the first lessons in social science research is that the study population should not be selected based on the dependent variable [21, 57]. When the outcome of an event is measured by users' text, *keyword panels* are subject to selection bias since they mention the event used to construct the study population. For instance, when studying the impact of mass shootings on Twitter, one should not select users already influenced by a shooting or use the fact that the identified users are tweeting about the event to justify its impact.
- Content bias: Users mentioning certain keywords on Twitter after an event may be systematically biased towards mentioning the keywords in general, compared to users who did not mention the words after the event. Regarding mass shootings, keyword searches are likely to find users who have an interest in the topic beforehand and think and talk about mass shootings differently than the general population. Hence, it is unclear whether impacts of events are driven by the events themselves or merely encourage discussion among users who are already interested in a particular event.
- Demographic bias: Sampled users may differ from the general population in terms of their demographics. For example, males are more likely to be contained in keyword-based panels for both mass shootings and Xbox events.

Among studies seeking to correct the bias of Twitter samples, most scholars have focused on demographic bias (e.g., [11, 43]). We acknowledge the importance of this, but we believe that selection-on-outcome bias is a more serious issue since it is more difficult to reduce compared to demographic bias. There is no doubt that to reduce demographic bias, scholars can re-weight samples based on various demographic properties towards the distribution of the offline population and calibrate user behaviors on Twitter, such as the number of tweets. There have been some good recent attempts to correct selection bias in big data using post-stratification techniques that re-weight the Twitter sample to the underlying population distribution [11, 41, 63]. Nevertheless, re-weighting bias caused by self-selection is difficult because it requires either a strong theory about factors that contribute to sample selection (e.g., the famous Heckman estimators in econometrics) or exogenous variables (e.g., instrumental variables) to help model the selection bias [21, 26, 65]. The former is difficult in social media analytics since theories about why and how people post are underdeveloped; the latter is difficult because the causal inference literature has only recently begun to be applied in social media analytics [28]. Hence, even if re-weighting procedures result in a panel with a demographics distribution that look very similar to that of random users, the panel may still exhibit selection-on-outcome bias.

We empirically reveal the three types of bias discussed above. For illustration, we analyze the Fort Hood mass shooting on April 2, 2014, which was the largest mass shooting in terms of injuries in 2014, and a national Xbox advertisement that was aired on January 19, 2014. The results are similar to other mass shootings and Xbox advertisements.

For both events, users did not anticipate the occurrence of the event, especially the time and location of the event. Hence, if users were selected randomly, we would expect that the level of discussion about the event, as measured by mentions of the exact keyword of interest, those used to construct the *keyword-based panels*, and other similar words, would remain at a minimal level and then spike when the event occurred. Figure 1 shows the proportion of users who discussed event-related keywords before the event for each of the three types of panels–random, keyword, and geolocated. The horizontal axis lists the different panels. For instance, the first column corresponds to a *keyword-based panel* whose users mentioned "shooting" in their tweets after the event. The vertical axis shows the proportion of users in certain panels who mentioned the respective keywords within seven days before the event. For instance, the cell with the horizontal value "shooting" and the vertical value "kill" indicates the proportion of users in the keyword panel constructed using the keyword "shooting" and the keyword "kill" in the seven days before the event.

Users from the *keyword-based panel* showed selection-on-outcome bias: they were more likely to mention the exact keyword that was used to evaluate the impact of events. For the mass shooting *keyword-based panel*, around 30% of users mentioned the words "shooting" and "kill" in the seven days prior to the event. Therefore, the users selected from keyword panels were already more likely to tweet about the exact keyword used to construct the panels, even before the event. Therefore, it is difficult to determine whether the observed changes are due to the impact of the event or the ex-ante interests of users.

Users from the *keyword panel* also exhibit outcome bias: they were more likely to mention other keywords relevant to the event *before* the event. A total of 20% mentioned related keywords such as murder, jail, attack, and shooting itself. The existence of outcome bias indicates that even though scholars use some keywords to construct panels and other related keywords to evaluate outcomes, the *keyword panel* is still subject to content bias.

Theoretically, we expect the *random panel* to exhibit minimal selection bias and the *geolocated panel* to reduce selection bias, compared with the *keyword panels*. Confirming this expectation, we find that the random and geolocated panels mentioned "shooting" and "kill" less often than any of the keyword panels (Figure 1). Furthermore, the *geolocated panel* is comparable to a random panel in terms of reducing selection bias. Figure 1 highlights the discussion in the seven days prior to the events. In the next section, we discuss bias after the event.

The results reveal the first two aspects of selection bias in the *keyword panel*: a nontrivial proportion of users discussing the keyword used to construct the panel and related keywords before the event. However, these selection bias issues are not prevalent in random and geolocated panels.

Further, we compare the differences in an important demographic variable—gender distribution—among the three panels in Table 1[8]. The *keyword-based panels* were 65% male, and the geolocated panels were 53% male. A Pew survey found that 53% of American Twitter users are male. Thus, the *geolocated panel* is similar to our best estimate of the ground truth based on the Pew survey, while *keyword panels* are severely biased in their gender distribution [9]. While gender is just one of many ways to consider user demographics, the purpose of verifying the selection bias with keyword

[8] We used the Discussion Graph Tool to identify users' gender [32]. Across the panels, the algorithm identified the same proportion of users based on usernames, which suggests that the tool does not favor a certain panel over others.

[9] http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/. Overall, 21% of women and 24% of men use Twitter, and the population is 50% male. Thus, 24/(21+24) of Twitter users are male.

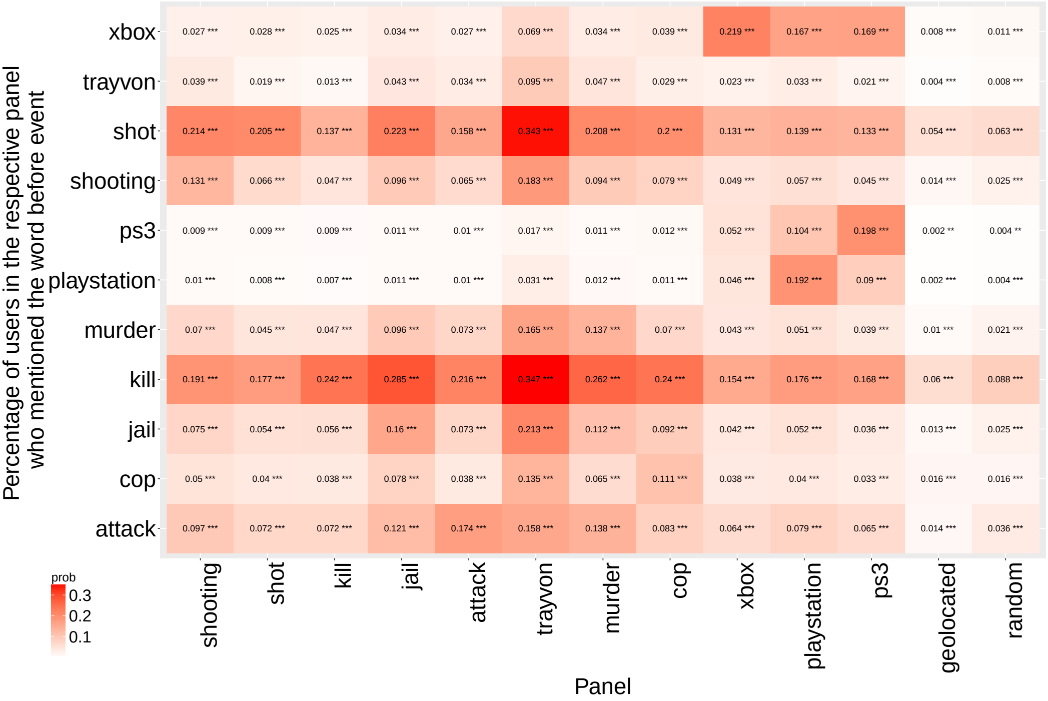

Fig. 1. Proportions of users mentioning the respective keywords in each panel before the occurrence of an event. Each cell comprises the proportions and its p-value. We perform statistical test through a one-side t-test. The p-values are: $< 0.1, *; < 0.01, **; < 0.001, ***$.

panels is to show the strength of the bias. Scholars are encouraged to consider more thoroughly what kinds of demographic traits are possibly correlated with an event being studied and thus evaluate appropriate selection biases.

In sum, we empirically demonstrate the bias introduced by keyword panels: users in a *keyword panel* are more likely to mention the keyword used to construct the panels and related words even before an event. The sample is also biased in terms of demographic characteristics compared with random panels. *Random panels* can reduce selection bias at first sight. However, this approach has two major shortcomings: efficiency and lack of a comparison group. First, for many events, few users are identified since the vast majority never discuss a particular event. Second, for many events, we care about the identification of a subgroup that defines a comparison group (in some cases, a treatment and a control group).

4.2 Estimating the impact of events

Next, we display the empirical estimation of the impact of mass shootings and advertisements on Twitter users' discussion of these events. The purpose is to compare how the three types of panels provide different estimates of the impacts of events at the population level.

We begin by showing the proportion of users who mentioned events before/after the events of all three types of panels: geolocated, random, and keyword-based. Figure 2 shows the proportion of users who mentioned the word "shooting" as a measure of the impact of the first set of mass shootings. As expected, there is a spike in mentions of "shooting" on the first day of all three panels.

Table 1. Gender ratio of the three types panels

Panel	Total Number of users	%Gender Identified	%Male
attack	265, 326	45	60
cop	171, 370	45	65
jail	182, 200	40	65
kill	764, 917	41	55
murder	197, 792	43	62
playstation	33, 921	45	80
ps3	42, 207	43	82
shooting	239, 106	47	65
shot	595, 104	46	65
trayvon	6, 842	40	72
xbox	131, 520	48	80
geolocated	116, 737	50	53
Pew	1,597	1.00	53

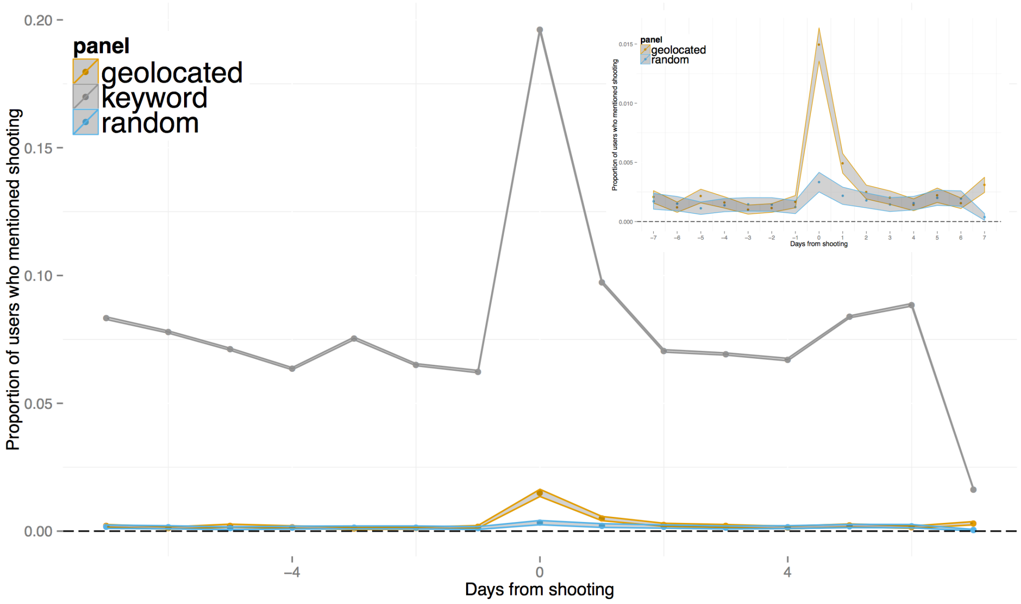

Fig. 2. Proportion of users within each panel who mentioned "shooting" by days from shooting at Fort Hood on April 2, 2014. A 95% confidence interval was obtained from bootstrapping methods and is plotted in shadow. The points for the geolocated and random panels overlap each other so a zoomed-in subplot is provided in the top-right corner.

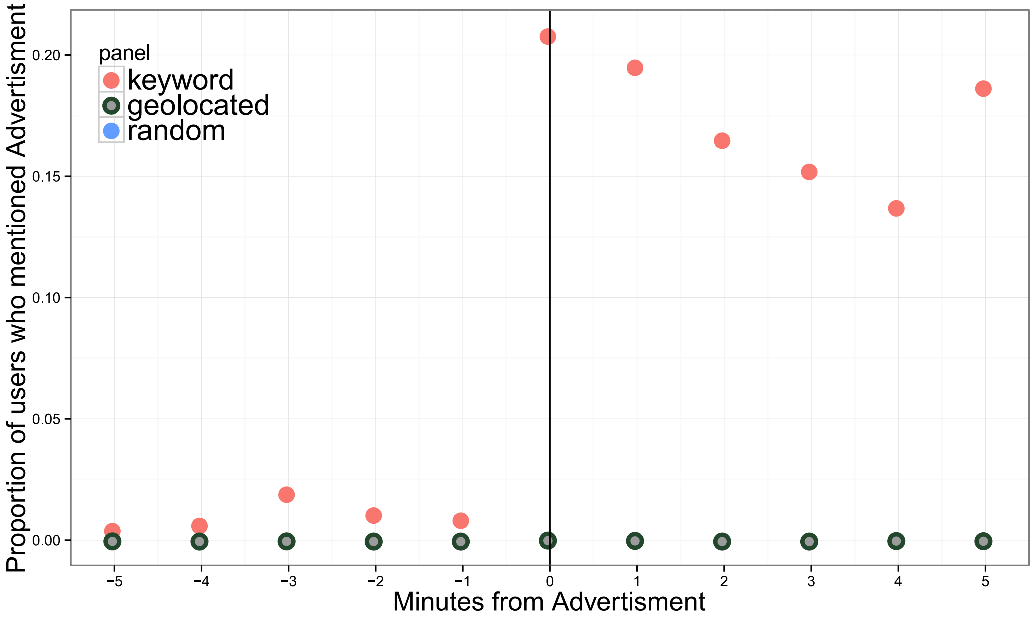

Fig. 3. Proportion of users within each panel who mentioned “Xbox” by minutes from batch of local advertisements at 2:22 PM ET on January 12, 2014.

However, the proportion of users who mentioned “shooting” on the first day of the Fort Hood mass shooting in 2014 is much higher in keyword-based panels than in random and geolocated panels. The *keyword panel* evaluates the impact on users who are already interested in the topic and hence provides a much larger estimation. The real impact of events for the entire Twitter population, however, is greatly exaggerated in this *keyword panel*. It is unsurprising that the effect measured in the *geolocated panel* is slightly larger than that of the *random panel*, because we limited the geolocated users to within 100 miles of the shooting and thus they are more likely to be treated by the event. Hence, we show that the *geolocated panel* can replicate the estimations of the *random panel* in a more efficient way.

In Figure 3, the impacts of the Xbox advertisements are displayed at the minute level since previous research has found that advertisements trigger responses on Twitter often within seconds of an event [35]. All of the panels were reconstructed to fit the shorter time frame. The Xbox advertisement has a much smaller overall effect than the shootings. This is partially due to the general level of chatter and the smaller impact of a single advertisement. Here, we can see how a *random panel* is unable to pick up any response because it is simply too small to capture this rare event. The *geolocated panel* has a small bump, as we confined the panel to the 14 DMAs that received the treatment, but it is impossible to see this in the provided figure. Only the *keyword-based panel*, with its selection bias, shows a clear impact.

Next, we show results that can only be measured meaningfully with geolocated panels: in other words, how impacts of events change spatially. There is not enough geolocation information in the *random panel* to determine this. Further, the *keyword-based panel* suffers from all of the selection bias issues noted in the previous sections.

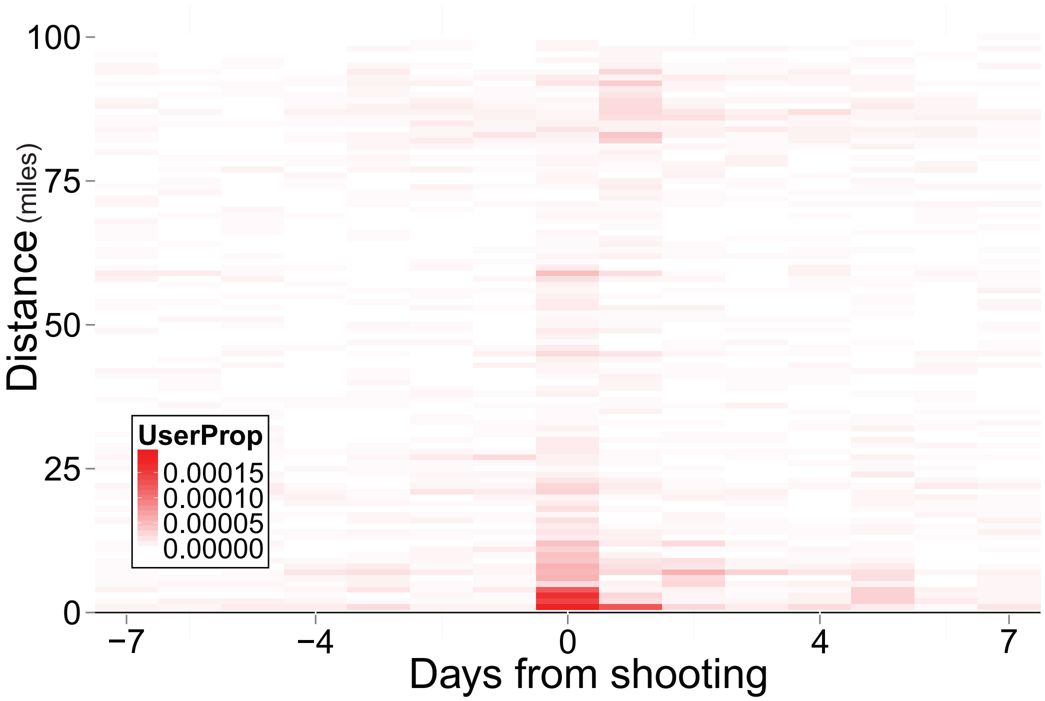

Fig. 4. Proportion of users within the geolocated panel who mentioned "shooting" by days from shooting and by distance to any of the 15 mass shootings in the US in 2014.

Figure 4 shows how the proportion of users who mentioned "shooting" changes by days from the shooting and by spatial distance. The proportion is an average effect over all 15 mass shootings in 2014. As expected, the proportion of users who mentioned "shooting" is randomly distributed before events by time and spatial distance. It indicates that events are exogenous to users in the *geolocated panel*. On the first day of shooting, there is a huge spike of users discussing the event at locations nearest to the event (within 5 miles). The impact quickly declines by distance on the day of the shooting but remains high compared with other dates. For all users, the impact decays quickly after the first day, remaining high for the second day in only the closest regions, before approaching pre-event numbers. Impacts remain for another three days after the second day only for users living within 10 miles of an event. The figure confirms that impacts of events decay by spatial and temporal distance.

The results for the *geolocated panel* for Xbox are shown in Figure 5. Figure 5 is similar to Figure 3 but distinguishes users in the geolocated panels by whether their frequent location is in one of the 14 DMAs in which the advertisement was aired. Users in one of the 14 DMAs could see the advertisement (i.e., serving as the treatment group), while users who lived outside the boundary but close to the 14 DMAs could not see the advertisement (i.e., serving as the control group). Therefore, by comparing the treatment and control groups, we can rule out other types of bias such as that related to demographics. We see that treated users respond more than untreated users to a given advertisement. It is possible that the difference is downwardly biased in that some untreated users are actually in the DMA but have not been identified as such. The difference between the two figures is startling, in that the impact of the advertisement is actually much larger, as a percentage bump,

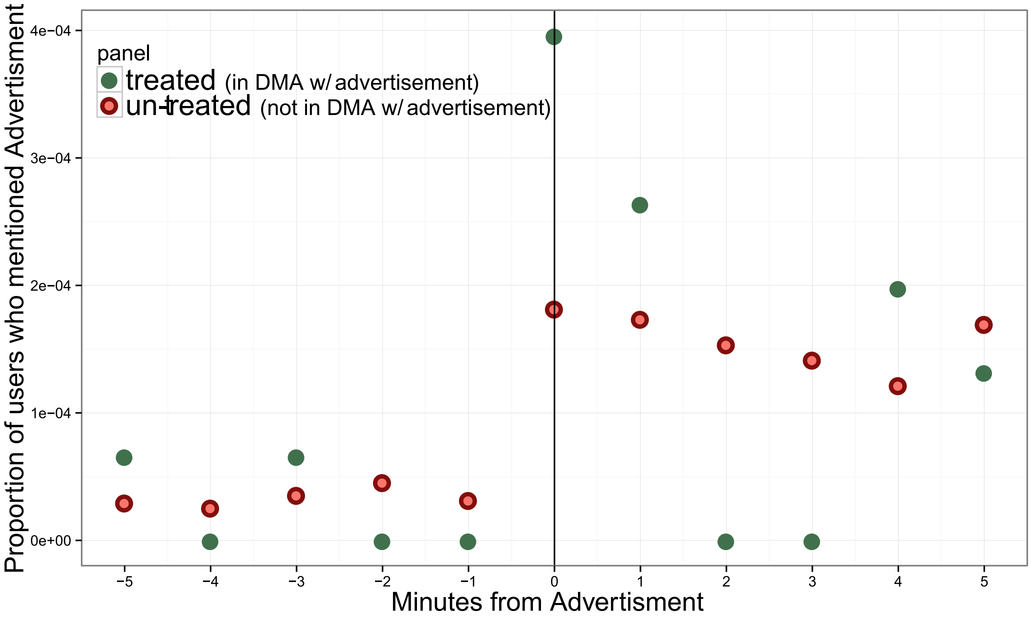

Fig. 5. Proportion of users within the *geolocated panel* who mentioned "Xbox" by minutes from the batch of local advertisements at 2:22 PM ET on January 12, 2014, within and outside of DMA in which it was aired.

for the panel of users who may not regularly think about Xbox, compared with the *keyword-based panel*.

Furthermore, we know that prior to the airing of the Xbox advertisements, there was no other confounding effect, such as the promotion effect, from other types of advertisement since the treatment and the control groups behaved similarly before the Xbox advertisement. This argument can be made since we collected the ex-ante posts of users. For a typical *keyword cross-section*, which lacks historical tweets, we cannot make such an argument.

Lastly, Figure 6 shows how sentiment changes by days and spatial distance from mass shootings. Again, the effect is averaged over all mass shootings. Specifically, we measured the score of fear using the Discussion Graph Tool (DGT)[32]. The DGT produced a joint distribution of seven types of mood—joviality, fatigue, hostility, sadness, serenity, fear, and guilt—for each tweet. Here, we use only the fear scores. Fear does not disappear, even after seven days, when people mention a shooting. There is a clearer pattern of decay by spatial distance after an event. The ex-ante fear scores are randomly distributed by time and distance, which suggests that our panel does not exhibit strong selection bias. Further, it shows that people mention "shooting" without fear; the phenomenon of fear is attached to the presence of a specific event.

5 GEOLOCATED PANELS

The previous section showed that *geolocated panels* not only reduce selection bias but can also be used to obtain population estimates for the impact of events. The idea of geolocated panels is relatively easy to understand, but there are many practical considerations when constructing a geolocated panel for a specific problem. In this section, we discuss aspects for scholars to consider when constructing their own *geolocated panels*.

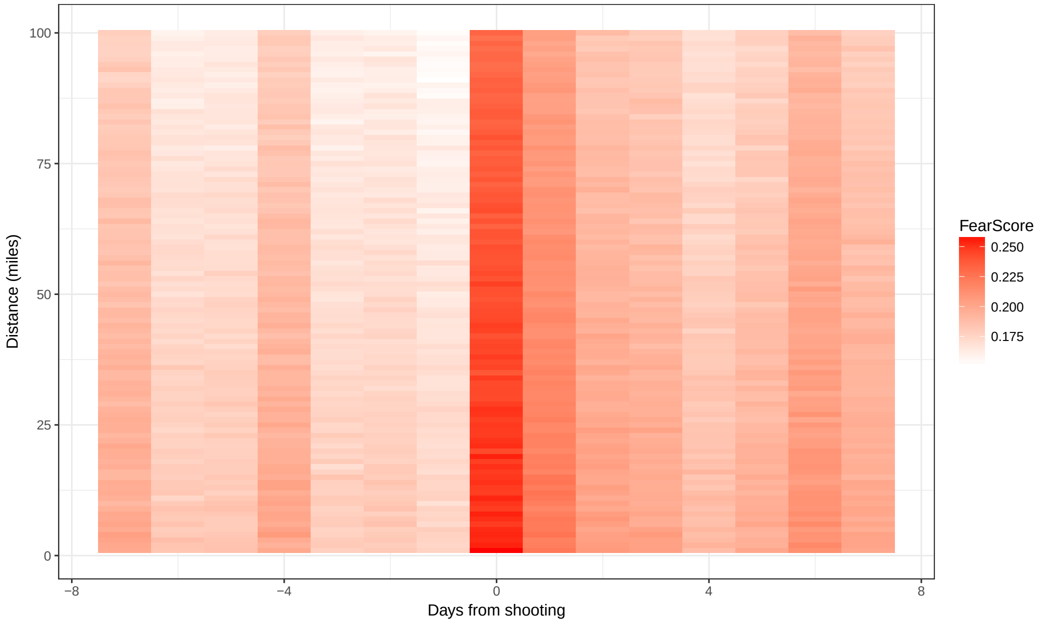

Fig. 6. Predicted fear scores of users' tweets mentioning "shooting" by days from shooting and by distance from any of the 15 mass shootings in the US in 2014.

5.1 When should we use geolocated panels

In Section 3, we provide evidence that *geolocated panels* reduce selection bias for the events we are studying. However, because it is optional to turn on geolocation services when tweeting, only a small proportion of tweets are geolocated. There have been recent advances in the data mining literature on geolocation inferences of social media users using profile and historical tweets, thus making it possible to significantly enhance the pool of users who have some version of geolocation information from which we can construct geolocated panels [6, 19, 29, 39]. Still, geolocated panels require that we know the geolocation information (either from check-ins, profiles, or historical tweets) of users, which shrinks the size of panels.

When should we use geolocated panels, and when should we keep conventional keyword panels? We argue that the decision should be based on the following considerations:

(1) Whether the outcome is evaluated through texts of tweets or not.
- When outcomes are evaluated through the textual information of tweets: Using text to evaluate outcomes is still the far more popular choice in the literature (including the examples in Section 3), as summarized in Section 2. *Keyword panels* exhibit selection-on-outcome bias, since the outcomes are mixed with the data collection itself; *random/geolocated panels* reduce the selection bias.
- When outcomes are evaluated through the non-textual information of tweets: For instance, when the outcome of interest concerns the spatial pattern of users after an event [8], *geolocated panels* may be more biased than keyword panels in this case.

(2) Whether the geolocation of users is exogenous (orthogonal) to the outcome of interest.
- If geolocation is exogenous to outcomes, geolocated users and random users are equally likely to discuss event-related topics, as we show in Table 1, using gender as an

example. In other words, the proportion of event-related discussion among geolocated and random users will be similar, even if geolocated users check-in more often than random users. In this case, even though geolocated users check-in more frequently than random users, and may differ in some other aspects, such biases are orthogonal to the outcome. Thus, we do not expect coverage issues in geolocated panels to impact the outcome of interest. We recommend using *geolocated panels* in this case.

- If geolocation is endogenous to outcomes, geolocated panels are subject to bias introduced by geolocation. In this case, geolocated panels reduce the selection bias on content but may exaggerate the selection bias on demographics. Therefore, we recommend continuing to use keyword panels.

5.2 Construct geolocated panels

Five steps are required for constructing a *geolocated panel*:

(1) *Control for time*: Create a full list of geolocated users within the necessary time frame.
(2) *Control for location*: Construct a panel of users who were sufficiently "close" to the event in terms of location and were thus exposed to it.
(3) *Create control*: Distinguish between those who were exposed to the event (the treatment group) and those who could have been but were not (the control group). For instance, advertisers selectively air an advertisement over a DMA and thus users who live near the boundary of a DMA have different probabilities of receiving the same advertisement. By constructing comparison groups across the boundary of a DMA, we can perform a rigorous causal analysis. In contrast, a mass shooting does not have a clear boundary, but the treatment fades with distance.
(4) *Gather full tweet history*: After constructing a panel of users who have possibly been "exposed" to an event, scholars should further collect their entire tweet history before and after the event. The ability to collect ex-ante information allows scholars to compare changes in individual outcomes after an event, which is a major advantage of social media data over traditional social survey data.
(5) *Consider outcomes*: Choose the specific outcome measures for the problem of interest.

Despite these seemingly straightforward procedures, executing these steps is not as simple as one might think. Often, the choices are tricky, and scholars may need to make adaptations for different events. The remainder of this section outlines the practical problems one may encounter when applying the framework to a specific event, and the choices a scholar can make regarding this. We specifically focus on steps 2 and 3.

5.3 Control for location

For a geolocated panel, we need to determine what location to assign to users. Then, we can find users who were spatially proximate and hence possibly exposed to an event. To identify user locations, we can use either geolocated posts or locations provided in profiles or user tweets. Scholars can make various choices when they decide to use geolocated posts to define users' locations. When tweeting, users can opt to turn on the location service, and their tweets will hence become geolocated, which instantly reveals people's location at the level of meters and the time of appearance at the level of seconds. We propose three ways to calculate users' spatial distance to events based on different ways of theorizing the impact of events.

- Instant distance: The distance of users to the event during the event. Instance distance reveals the distance between the event and the user at the time the event occurred. Instant distance should be used if scholars care about how physical exposure to events exerts influence on individuals because the precise time and location of individuals are essential for knowing who was possibly physically exposed to an event and who was not [11, 66].

The shortcomings of using users' instant location at the time an event occurred to construct a panel of users are threefold. First, many users may be at the location but not tweet as the event unfolds. Thus, by restricting the panel to users' providing their instant location, we underestimate the size of the group of users at the event. Second, for many events, being physically present is not a key constraint but may indicate a tie to the area. Third, instant location is sometimes endogenous to an event: users may change their location, either approaching or leaving the location of an event, sometimes due to the event [58].

- Shortest distance: The minimal distance between an event and any geolocated tweet of a user is meaningful for a unique event. Shortest distance should be used if the event of interest will have an impact on users if they have ever been to the location.
- Frequent distance: The distance between users' frequent locations to the event of study. Using frequent distance is preferable for events where the frequent distance to an event determines the probability of being exposed to the event, as in the case of mass shootings or national elections. The influence of these events can be spread out in a nonphysical fashion, such as through the news media or diffusion through friends and family. For these events, their influence goes far beyond a reasonable range of physical exposure, say, hundreds of meters, and hence using instant location may not be the optimal choice. Places where users frequently check-in matters more in this situation as there is a high likelihood these are users' home or work locations[8, 24].

In our empirical analysis, we use users' frequent locations to construct *geolocated panels*. We identify Twitter users' frequent locations from their history of geolocated tweets over a long period (a year in our empirical analysis) and then sample users based on distances of their frequent locations from the event. An advantage of this approach is that it does not require users to post geolocated tweets near an event when the event occurs. Hence, it can increase the size of the study population and better approximate who is actually affected by the unfolding event. Furthermore, keyword panels are often sampled from the activity stream over a limited time window, missing users who were inactive during this period. By identifying users who remained close to the event for a much longer time frame, we can include users who were inactive when the event occurred.

Using geolocated tweets provides the advantage of precision but requires users to have geolocated tweets; such users are, however, a small proportion of Twitter users. Hence, while using profile location provides the advantage of increasing the size of the study population, it also brings several disadvantages such as coding cost, reliability, and coarseness of measures [5, 25]. Recent research in data mining has advanced methods of identifying users' locations based on text and networks, greatly improving the size of the population identified with geolocations [6, 8, 19, 39].

Scholars can use profile locations provided by users if they find that the need to collect more data efficiently outweighs the requirement for location precision. Ultimately, the choice of how to use geolocation information depends on trade-offs between: 1) the mechanisms through which events influence individuals and 2) the trade-off between granularity of geolocation measures and population size.

5.4 Creating control groups

Another advantage of the *geolocated panel*, compared with the *keyword panel*, is that it leaves space to construct control groups so that scholars can draw comparisons to infer causal effects. Creating controls for *keyword panels* is difficult since we need to find a group of users that could

Table 2. Comparison between different panels

	Keyword	Geolocated	Random
Data Collection Method	Keyword-filtering through search API	Geolocation-filtering through search API	Streaming API/Firehose.
Selection bias	Major selection bias on content; selection bias on demographics	Selection bias on demographics (especially geolocations)	Minimal
Speed	Fast	Intermediate	Slow
Coverage	Subset of Twitter that has discussed the relevant event	Subset of Twitter that is exposed to the event	1% of Twitter users/all Twitter users

discuss the event but did not. For rare events, it has not been determined how to find such a group. On the other hand, constructing control groups in *geolocated panels* is easy due to natural geographical boundaries. If an event has a geographical center, some users will be more treated or, if the treatment is binary, either treated or not treated by the event. These strategies have been utilized often in natural experimental designs [31].

We classify an event by whether the time and place of its occurrence are endogenous to the event. For instance, general users cannot predict when and where a mass shooting will occur. Hence, both time and place are exogenous to Twitter users. In this case, we use the distance from the exact location as a proxy for the quantity of treatment. For an advertisement, anyone within a DMA is treated in the same way. Therefore, there is a binary control of people within the DMA in which an advertisement is aired and those outside of the DMA but in nearby counties.

It is possible that the place of an event is endogenous to Twitter users, such as with local crimes. Users living in a neighborhood with higher crime rates are more likely to be exposed to crimes than those who live far away, and they are likely to be poorer. Hence, it is difficult to prove whether crimes have negative impacts on local residents, or if such impacts are due to their disadvantageous situation regarding other aspects such as economic conditions. This means our *geolocated panel* reduces selection-on-outcome bias of the *keyword panel* but may introduce other demographic biases. Furthermore, users' decisions to check-in could be not random[29].

In this situation, there is still the possibility to correct such bias during sampling procedures for *geolocated panels*, while it is difficult to do so for *keyword panels*. The second way to correct the bias introduced by endogenous events is to use both instant and frequent locations of users to construct comparison groups. The key is to find two groups of users who satisfy the following criteria:

- The two groups share the same frequent locations and hence there is no bias regarding where they self-select to live/work.
- The two groups differ in terms of their instant locations: users of one group were at their frequent location when the event occurred, and the other group was far away from it. The former group has a higher likelihood of being exposed to the event.

By comparing two groups, we can gauge how different levels of exposure to events influence similar users differently. However, in keyword panels, it is difficult to determine how to sample users who did not mention events into the panel.

We summarize the data collection methods, biases, speed, and coverage issues of each type of Twitter panel in Table 2.

6 DISCUSSION

This paper categorizes three commonly used panel types in the literature on event analytics using social media data—keyword, geolocated, and random panels. We compare the advantages and disadvantages of using each type of panel to evaluate the impact of real-world events. Specifically, we find that when the outcome is evaluated through text, the most widely used type of panel, the *keyword panel*, is inherently biased due to selecting on outcomes. The bias from selecting on outcomes is harder to resolve compared with the widely recognized bias caused by the unrepresentative and shifting distribution of Twitter demographics in previous research.

In the literature using social media data to study events, this paper is the first to empirically demonstrate the significance of biased samples leading to very different estimates of outcomes. This paper explicitly lists forms of selection bias and discusses ways to correct each type. Furthermore, we also propose the advantages of *geolocated panels*, namely the ability to draw objective comparisons by distinguishing treatment and control groups when possible, which is not feasible in keyword-based panels. This paper calls for scholars using social media data for event analytics to recognize and begin to address the commonly exhibited selection bias in keyword-based data collection methods.

Geolocated and *random* panels can reduce the bias caused by selecting outcomes. Geolocated panels are preferable to random panels both theoretically and empirically when the outcome is measured through tweets. First, geolocated panels make it possible for scholars to create comparison groups along geolocated treatments. Second, geolocated panels are efficient for studying where events have impacts since they geographically narrow down the sample population to users who could be impacted by an event. A random sample of Twitter usually requires a large population to capture outcomes of interest. Geolocated panels allow scholars with restricted data access to answer questions that would be impossible to consider with a random panel. In summary, as demonstrated in this paper, geolocated panels allow scholars to reduce selection bias.

This paper reveals two possible directions for the future social media analytics: 1) turning away from tweet-centric to user-centric analysis and 2) building panels based on users' demographic characteristics. Specifically, we strongly advocate for scholars to use users as the unit of analysis, instead of posts, due to the drifting nature of posts and the lack of historical tweet data. Furthermore, while this paper uses geolocation information to construct panels, Twitter panels can be built with other demographic characteristics such as gender, region, and user labels. Twitter panels based on other demographics can be used to extract study users for events that do not exert influence over physical space but rather through social networks, which can be imagined as a kind of social space. For these events, panels based on factors other than physical distance to an event, such as social distance to the originator of an event (in the case of a protest) can efficiently select a study population while also reducing selection bias. While a few works have considered user attributes, there is room for improvement. Scholars should choose suitable methods to build appropriate Twitter panels, based on our suggestions in Section 5.1 and their specific research contexts.

Finally, regarding the study of data and information quality, we introduce the selection bias issue. We argue that selection biases should be understood as an objectivity issue of data and information quality. Selection bias not only leads to unobjective measures of the impact of events but can also decrease the believability of social media data. Our research thus calls for greater consideration of how data quality concepts should be applied to social media.

REFERENCES

[1] Nitin Agarwal and Yusuf Yiliyasi. 2010. Information quality challenges in social media. In *International Conference on Information Quality (ICIQ)*.

[2] Jisun An and Ingmar Weber. 2015. Whom should we sense in “social sensing” - analyzing which users work best for social media now-casting. *EPJ Data Sci.* 4, 1 (Nov. 2015), 1–22.

[3] Joshua D. Angrist and Jrn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist’s companion.* Princeton university press. http://books.google.com/books?hl=en&lr=&id=ztXL21Xd8v8C&oi=fnd&pg=PR8&dq=most+harmless+econometrics&ots=Ui21TD0LyN&sig=u7i15nMhi6WZSZoStva4ESjhCPY

[4] Ceren Budak and Duncan J Watts. 2015. Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement. *Sociological Science* (2015), 370–397.

[5] Scott H Burton, Kesler W Tanner, Christophe G Giraud-Carrier, Joshua H West, and Michael D Barnes. 2012. ”Right time, right place” health communication on Twitter: value and accuracy of location information. *Journal of medical Internet research* 14, 6 (2012).

[6] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. 2011. Estimating twitter user location using social interactions–a content based approach. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on.* IEEE, 838–843.

[7] Lu Chen, Ingmar Weber, and Adam Okulicz-Kozaryn. 2014. US Religious Landscape on Twitter. In *Social Informatics.* Springer, 544–560.

[8] Eunjoon Cho, 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.* ACM, 1082–1090.

[9] Murphy Choy, Michelle L F Cheong, Ma Nang Laik, and Koo Ping Shung. 2011. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. (Aug. 2011). arXiv:1108.5520

[10] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. 2012. Discover Breaking Events with Popular Hashtags in Twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM ’12).* ACM, New York, NY, USA, 1794–1798. DOI:http://dx.doi.org/10.1145/2396761.2398519

[11] A Culotta. 2014. Reducing Sampling Bias in Social Media Data for County Health Inference. *Joint Statistical Meetings Proceedings* (2014).

[12] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.* ACM, 241–244.

[13] Fernando Diaz, Michael Gamon, Jake M Hofman, Emre Kıcıman, and David Rothschild. 2016. Online and social media data as an imperfect continuous panel survey. *PloS ONE* 11, 1 (2016), e0145406.

[14] J DiGrazia, K McKelvey, J Bollen, and F Rojas. 2013. More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLoS ONE* 8, 11 (2013).

[15] Bruce Doré, Leonard Ort, Ofir Braverman, and Kevin N Ochsner. 2015. Sadness shifts to anxiety over Time and distance from the national tragedy in Newtown, Connecticut. *Psychological science* 26, 4 (2015), 363–373.

[16] Thad Dunning. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach.* Cambridge University Press.

[17] Nugroho Dwi Prasetyo and Claudia Hauff. 2015. Twitter-based Election Prediction in the Developing World. In *the 26th ACM Conference.* ACM Press, New York, New York, USA, 149–158.

[18] Fahame F. Emamjome, Ahmad A. Rabaa’i, Guy G. Gable, and Wasana Bandara. 2013. Information quality in social media : a conceptual model. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2013)*, Jae-Nam Lee, Ji-Ye Mao, and James Thong (Eds.). AIS Electronic Library (AISel), Jeju Island, Korea. http://aisel.aisnet.org/pacis2013/72

[19] David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. 2015. On the accuracy of hyper-local geotagging of social media content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* ACM, 127–136.

[20] Daniel Gayo-Avello. 2011. Don’t turn social media into another ’Literary Digest’ poll. *Commun. ACM* 54, 10 (2011), 121–128.

[21] Barbara Geddes. 1990. How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political analysis* 2, 1 (1990), 131–150.

[22] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[23] Amir Goldberg. 2015. In defense of forensic social science. *Big Data & Society* 2, 2 (Dec. 2015), 2053951715601145. DOI:http://dx.doi.org/10.1177/2053951715601145

[24] Marta C González, César A Hidalgo, and Albert-László Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (June 2008), 779–782.

[25] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. *Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles.* ACM, New York, New York, USA.

[26] James J Heckman. 1977. Sample selection bias as a specification error (with an application to the estimation of labor supply functions). (1977).

[27] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 67.

[28] Andreas Jungherr. 2015. *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research.* Springer Publishing Company, Incorporated.

[29] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media (ICWSM).*

[30] Nattiya Kanhabua and Wolfgang Nejdl. 2013. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd international conference on World Wide Web companion.* International World Wide Web Conferences Steering Committee, 1335–1342.

[31] Luke J. Keele and Roco Titiunik. 2014. Geographic Boundaries as Regression Discontinuities. *Political Analysis* (Oct. 2014), mpu014. DOI: http://dx.doi.org/10.1093/pan/mpu014

[32] Emre Kıcıman, Scott Counts, Michael Gamon, Munmun De Choudhury, and Bo Thiesson. 2014. Discussion Graphs: Putting Social Media Analysis in Context. In *Eighth International AAAI Conference on Weblogs and Social Media.*

[33] Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media.* ACM, 243–248.

[34] Gary King, Robert O Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research.* Princeton University Press.

[35] Brendan Kitts, Michael Bardaro, Dyng Au, Al Lee, Sawin Lee, Jon Borchardt, Craig Schwartz, John Sobieski, and John Wadsworth-Drake. 2014. Can Television Advertising Impact Be Measured on the Web? Web Spike Response as a Possible Conversion Tracking System for Television. In *Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* ACM, 1–9.

[36] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18, 1 (2009), 140–181.

[37] Vasileios Lampos and Trevor Cohn. 2013. A user-centric model of voting intention from Social Media. In *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL.*

[38] Janette Lehmann, Bruno Gonçalves, José J Ramasco, and Ciro Cattuto. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web.* ACM, 251–260.

[39] Chenliang Li and Aixin Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* ACM, 43–52.

[40] Yu-Ru Lin, 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, 737–748.

[41] Roderick JA Little. 1993. Post-stratification: a modeler's perspective. *J. Amer. Statist. Assoc.* 88, 423 (1993), 1001–1012.

[42] Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a Large-scale Corpus for Evaluating Event Detection on Twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information &#38; Knowledge Management (CIKM '13).* ACM, New York, NY, USA, 409–418. DOI: http://dx.doi.org/10.1145/2505515.2505695

[43] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *ICWSM* 11 (2011), 5th.

[44] Stephen L Morgan and Christopher Winship. 2014. *Counterfactuals and causal inference.* Cambridge University Press.

[45] Fred Morstatter, Jrgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv:1306.5204 [physics]* (June 2013). http://arxiv.org/abs/1306.5204 arXiv: 1306.5204.

[46] Brendan O'Connor, 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 (2010), 1–2.

[47] Hüseyin Oktay, 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.* ACM, 1–9.

[48] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM" 14).*

[49] Minsu Park, Ingmar Weber, Mor Naaman, and Sarah Vieweg. 2015. Understanding Musical Diversity via Online Social Media. In *Ninth International AAAI Conference on Web and Social Media.*

[50] Takeshi Sakaki, 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.* ACM, 851–860.

[51] Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics, 53–60.

[52] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. 2003. *Model assisted survey sampling*. Springer Science & Business Media.

[53] Matthias Schonlau, Arthur Van Soest, Arie Kapteyn, and Mick Couper. 2009. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research* 37, 3 (2009), 291–318.

[54] G. Shankaranarayanan and Roger Blake. 2017. From Content to Context: The Evolution and Growth of Data Quality Research. *J. Data and Information Quality* 8, 2 (Jan. 2017), 9:1–9:28. DOI: http://dx.doi.org/10.1145/2996198

[55] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS one* 6, 5 (2011), e19467.

[56] Marko Skoric, 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*. IEEE, 2583–2591.

[57] Edward A Suchman. 1962. An analysis of fibiasfi in survey research. *Public Opinion Quarterly* 26, 1 (1962), 102–111.

[58] Pal Roe Sundsoy, Johannes Bjelland, Geoffrey Canright, Kenth Engo-Monsen, and Rich Ling. 2012. The activation of core social networks in the wake of the 22 July Oslo bombing. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*. IEEE, 586–590.

[59] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 62, 2 (2011), 406–418.

[60] Mikalai Tsytsarau, Themis Palpanas, and Malu Castellanos. 2014. Dynamics of news events and social media reaction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 901–910.

[61] Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[62] Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12, 4 (1996), 5–33.

[63] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2014. Forecasting elections with non-representative polls. *International Journal of Forecasting* (2014).

[64] I Weber, VRK Garimella, and A Batayneh. 2013. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the 2013 IEEE/ ...*. ACM Press, New York, New York, USA, 290–297.

[65] Christopher Winship and Robert D. Mare. 1992. Models for Sample Selection Bias. *Annual Review of Sociology* 18 (1992), 327–350. http://www.jstor.org/stable/2083457

[66] Han Zhang. 2015. Witnessing Political Protest on Civic Engagement and Political Attitudes: A Natural Experiment. (2015).