Geolocated Twitter Panels to Study the Impact of Events

Abstract

Data from Twitter have been employed in prior research to study the impact of events. Historically, researchers have relied on keyword-based samples of tweets to create a panel of Twitter users that mention event-related keywords during and/or after an event. There are limitations to the *keyword-based panel* approach. First, the technique suffers from selection bias since users who discuss an event are already more likely to discuss event-related topics beforehand; it is unclear whether observed impacts are merely driven by a set of users who are intrinsically more interested in events. Second, there are no viable groups for comparison to a keyword-based sample of tweeters. We propose an alternative sampling approach to studying response to events on Twitter that addresses the aforementioned two issues. We construct panels of users defined by their geolocation. These panels are exogenous to the keywords in users' tweets, resulting in less selection bias than the *keyword-based panel* method. *Geolocated Twitter panels* allow us to follow within-person changes over time and also enable the creation of comparison groups. We evaluate our panel selection approach in two real-world settings: response to mass shootings and TV advertising. We illustrate how our approach limits selection bias introduced by the keyword-based approach, how discussion among the panel of users shifts before and after an event, and how geography can provide meaningful comparison groups regarding the impact of these events. We believe that we are the first to provide a clear empirical example of how a better panel selection design, based on an exogenous variable like geography, both reduces selection bias compared to the current state-of-the-art and increases the value of Twitter research for studying events.

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, in terms of both the 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. Beyond these advantages, Twitter also offers easy methods for gathering data around events (Diaz et al., 2014), making it especially popular for scholars. Studies that have used Twitter to study events have attempted to answer questions that include how political protests impact political attitudes (Budak and Watts, 2015; Zhang, 2015) and how pandemics raise public concern (Signorini, Segre, and Polgreen, 2011), how natural disasters give rise to sadness and anxiety (Doré et al., 2015), and how opinions about elections change(O'Connor et al., 2010).

Despite the wide range of topics and questions scholars have pursued with Twitter, many prior studies share a common method of keyword-based data collection. Historically, researchers have used keywords to select a cross-section of tweets to be used to describe the tweet-level conversation around a specific topic at a particular point in time. Only recently, have researchers extended cross-sections to user-centric panels, by collecting the historical tweets of the users that appear in the keyword-based cross-section sample. We call this sampling strategy, the *keyword-based panel* approach throughout this paper:

1. Scholars use the Twitter search or streaming API to collect tweets that mentioned keywords or hashtags that are directly relevant to the event(s) being examined.
2. Scholars then analyze the change of counts, relative proportions, and sentiment of those tweets *caused* by events.

We argue that there are two critical problems of the *keyword-based panel* method: 1) selection-bias regarding users and their tweets, and 2) lack of comparison samples.

First, the *keyword-based panel* approach introduces selection bias. It uses event-related keywords to choose Twitter users that are used to study events themselves, which is known as sampling on outcome bias. The danger is that users who respond to events are usually more interested in event-related topics beforehand than random users. Hence, it is unclear whether impacts observed from the *keyword-based panel* method 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 in terms of demographic characteristics than those who do not mention the event. As we show later, individuals who mention the terms "shoot-

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ing" or "Xbox" are much more likely to mention exact terms and related keywords even before a triggering event (i.e. a mass shooting or Xbox advertisement), and more likely to be male than randomly selected users.

Second, to measure the impact of an event on users, scholars need a "control" group not influenced by an event to ensure that changes are not driven by confounding trends. When using the *keyword-based panel* method, all sampled users are already affected by the event. Consequently, this method cannot offer the possibility of drawing objective comparisons.

Further, When considering new sample selections for Twitter, we are responding to three traditional survey biases that are particularly acute in Twitter. First, there are coverage issues because only 19% of the adult population uses Twitter [1]. In comparison, most of the US population is still reachable by either a landline or cell phone, the traditional approach for collecting survey data. Second, Twitter suffers from extreme non-response issues, as users can opt-in and opt-out at will. A 5% response rate is expected for election surveys collected by phone in 2016, but that is still many magnitudes more than the percentage of Twitter users that discuss a given topic in any week or day (let alone an hour or minute). Finally, Twitter data are noisy, suffering from extreme measurement error, because unlike surveys, the responses on Twitter are unstructured.

Despite the aforementioned limitations, Twitter offers several advantages over traditional surveys. For instance, Twitter is a panel by definition. Whereas most survey research is conducted using cross-sections, Twitter has repeat users. Therefore, Twitter allows for within-person examination using panels. We compare three panel selection strategies.

Keyword-based panel: The *keyword-based panel* is based on users that mention event-relevant keywords within a limited time period after an event. In other words, the users "opted in" to the panel during the time frame in which scholars collected data by discussing the event on Twitter. This sampling strategy adds further selection bias to an already unrepresentative sample of Twitter users as most users "opted out" of the panel and were thus omitted from the sample. By collecting tweets of keyword-based sampled users before and after an event, we can empirically measure the selection problem by comparing the *keyword-based panel* around events to the two panels described below.

Random Panel: While it would be ideal to compare the *keyword-based panel* to the entire population of Twitter, this is not feasible, largely due to expense. Instead, we select a random sample of users, obtaining their entire tweet histories, and identifying the selected users' characteristics (i.e., available demographics, geography, etc.). This mirrors survey theory, in that if we collect a random sample of users, we can avoid additional bias, beyond the sample bias. We show that this is often not an efficient way to study events since a sufficiently large sample must be collected to find users that actually discuss any given topic. However, this type of panel serves as a useful baseline.

Geolocated Panel: We compare the *keyword-based panel*, along with a *random panel* of Twitter users, to a new concept, the *geolocated Twitter panel*. To create this geolocated panel, we used geolocation information of tweets to construct a panel of users (and their full tweet history) that are *close* to an event in terms of time and space and are thus likely to be "exposed" to the actual event, instead of self-selecting themselves into the panel. This can reduce problems of selection bias since users do not actively self-select themselves into the panel and the panel has more coverage of relevant users than a truly *random panel*.

This paper advocates that social media research on events should move from *keyword panels* to *geolocated Twitter panels*. We believe our paper is the first to provide a clear empirical example of how a better sample design both reduces bias in the current sample design and increases the value of this type of social media research.

Literature Review

This section provides an overview of selected previous research that has examined the impact of temporal events on Twitter users by analyzing their tweets. The 500 million tweets generated daily are timestamped and a proportion of tweets are geotagged. Therefore, the convenience of Twitter enables the study of a wide range of events that occur at specific times and locations, for various purposes: 1) for prediction, for example, Skoric et al. (2012); DiGrazia et al. (2013); O'Connor et al. (2010) have examined how tweets which mention candidates can be used to predict election outcomes; 2) for event detection (Cui et al., 2012; McMinn, Moshfeghi, and Jose, 2013), for example, researchers can detect that an earthquake has occurred in near real time (Sakaki, Okazaki, and Matsuo, 2010) ; and 3) to measure the impact of events (Budak and Watts, 2015; Signorini, Segre, and Polgreen, 2011). In the subsections below, we focus attention on strategies from prior research, used to sample Twitter users to measure the impact of events.

Keyword-based cross-sections

Researchers have used Twitter to describe the discussion about events. For example, Thelwall, Buckley, and Paltoglou (2011) analyzed sentiments of tweets containing event-related hashtags for top events mined from Twitter. Lehmann et al. (2012) identified events whose hashtags had a sudden spike and analyzed the context of tweets containing the hashtags. Tsytsarau, Palpanas, and Castellanos (2014) examined how news events, as revealed by online memes, trigger social media attention, by selecting tweets containing keywords related to the news events. While using keyword-based cross sections to study outbreaks of disease, Kanhabua and Nejdl (2013) noted that *keyword cross-sections* often include irrelevant tweets in the data. They argued that scholars should pay attention to the temporal and spatial dynamics of tweets.The studies mentioned above, all used keywords that are directly relevant to the events. However, the studies did not further associate users' historical tweets and attributes in their analysis, and hence adopted a keyword-based cross-section approach rather than a panel design.

[1] http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

Keyword panels: From tweet-centric to user-centric.

Keyword-based panels offer two advantages over *keyword cross-section* samples: 1) they are user-centric and 2) they contain data on the ex-ante behaviors of users. Regarding the first feature, scholars have recently cataloged the bias caused by tweet-centric analysis and urged a transition to user-centric analysis. Weber, Garimella, and Batayneh (2013) and Budak and Watts (2015) analyzed a fixed set of users' political behaviors before and after protests in Egypt and Turkey using a panel of users that mentioned protest 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 (Sang and Bos, 2012; Dwi Prasetyo and Hauff, 2015). Outside the election prediction realm, An and Weber (2015) confirmed that user-centric analysis outperformed counting tweet occurrences for predicting both flu-activity and unemployment rates.

Scholars have attempted to correct biases in keyword-based cross-sections and panels by including user features such as demographics, geography and tweeting behaviors in models (Lampos and Cohn, 2013; Gayo-Avello, 2011; De Choudhury, Diakopoulos, and Naaman, 2012), or by reweighting samples toward the underlying population truth with respect demographics and geography (Wang et al., 2014; Choy et al., 2011; An and Weber, 2015).

However, attempts for reweighting data does not fully solve the selection bias issue. In principle, there is potential to reweight biases on all sorts of demographic properties towards the distribution of the offline population, and calibrating user behaviors on Twitter such as number of tweets. Nevertheless, it is not clear how to reweight bias introduced by self-selection into both the platform and the particular topic; the variables may not be recoverable. Hence, it is possible that reweighting procedures result in a panel whose demographics distributions looks very similar to that of random users, but the panel may still discuss different topics of interest beforehand.

Panels based on demographics

Even though keyword-based sampling is the most prevalent in prior research, there have been recent attempts to directly collect user-centric data from user profiles. Tools such as Followerwonk allow scholars to first search for users of interest and then collect tweets of these users[2]. Scholars search for users with interests in the specific topics, such as using certain music services or religious affiliations(Chen, Weber, and Okulicz-Kozaryn, 2014; Park et al., 2015). This strategy, however, does not reduce selection bias since it still relies on a group of users who show interest in the topic being examined. Furthermore, by only focusing on searching users of interest, there lacks a meaningful comparison group.

Geolocation panels

Prior research has also used Twitter samples based on geolocation (Kim et al., 2014; Olteanu et al., 2014; Imran et al., 2015). For example, Zhang (2015) used check-in history of Weibo (Chinese Twitter) to find users nearby protests, collect their tweet history, and compared changes of political discussion on the user level before and after protests. Most prior work simply aims to reconstruct a representative panel of users based on their geolocations in order to study users in a particular location for a particular subject. Culotta (2014), on the other hand, offers an attempt to reduce selection bias by searching for users that had geolocated tweets in 100 counties of the USA, 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 do not address the potential for geolocated panels to enable the construction of comparison groups. Our paper extends the current literature by addressing this gap.

Selection bias

Selection bias is well known to be a major threat to causal analysis in empirical research (King, Keohane, and Verba, 1994). In event studies with social media data, researchers have acknowledged the importance of recognizing and evaluating selection bias. For instance, Tufekci (2014) showed the danger of deletion of hashtags during major events. Lin et al. (2013) compared users form keyword sampling with their focus group and noted considerable differences between the two. Culotta (2014) clearly mentioned the selection bias in keyword-centric design and argue that geolocation can reduce the problem. Other approaches use experimental, quasi-experimental or matching methods to correct the bias (Kohavi et al., 2009; Schonlau et al., 2009; Oktay, Taylor, and Jensen, 2010). Based on this previous thought, this paper explicitly listed forms of selection bias, and discusses ways to correct each type of bias. Furthermore, we also propose the advantages of a *geolocated Twitter panel*– to be able to draw objective comparisons by distinguishing treatment and control groups when possible, which is not feasible in keyword-based panels.

Selection bias has been rarely examined in scholarly works that use Twitter as a data source. Critics have addressed the problem of selection bias within social media, but this paper addresses the issue in the specific setting of keyword-centric design and is one of the first to empirically demonstrate the significance of biased samples leading to very different estimates of outcomes.

Events and Data

Most previous research using Twitter to understand the impact of events has focussed on a specific type of event (or sometimes even a single event). Prior research has had a substantive rather than methodological focus and hence fails to place the data collection procedure within a general framework. This paper compares three data collection frameworks–keyword, random and geolocated panels for understanding the impact of events.

For each framework, we study 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

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

in 2014. We use mass shootings in 2014 documented in the Stanford Mass Shootings of America (MSA) data project. [3] The Xbox advertisements include 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.

Both of 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. However, there are extreme differences in how people are experience these events and therefore how we expect them to react.

Next we describe how we construct the three types of panels for all the events in our study. We accessed the full Twitter stream, via the Twitter Firehose provided by Microsoft. 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 itself–to build eight keyword panels. For example, we constructed the *keyword-based panel* on the word "shooting" by finding all the tweets that mentioned the word "shooting" within seven days after each specific shooting, and collecting historical tweets of the users posting those tweets within seven days before the event. Similarly, for each Xbox advertisement, we constructed three panels using the keywords Xbox, playstation and ps3 respectively. 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 within 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 (Kitts et al., 2014). With the 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 that mentioned the keywords within the designated timeframe.

For comparison, we created random panels for each event. Again, we used the Twitter Firehose and selected 30,000 users at random who tweeted at least once during the specified timeframe after each event. We then collected all their tweets for the noted timeframe before the event.

Finally, we constructed *geolocated Twitter panels* for each event. 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. Then we mapped their geolocated posts over the year onto census tracts, and use the most two frequent tracts as the frequent locations. For mass shooting events, we randomly sampled users whose two most frequent location is within 100 miles of the particular shooting. For Xbox ads, we randomly sampled users whose most frequent location is within 100 miles of the center of a designated market area (DMA). In Section we discuss in more detail how we define a user's most frequent location. In Sections and we compare the three panels in different scenarios.

[3]Data can be downloaded from the following site https://library.stanford.edu/projects/mass-shootings-america

Selection Bias of Keyword-based Panels

As we briefly discussed in the introduction, the *keyword-based panel* method introduces selection bias into the sample, beyond the coverage and non-response bias inherent to Twitter. We differentiate between three aspects of the selection bias introduced in the *keyword-based panel* design as follows:

- Selecting based on outcome bias: Users have different probabilities to be selected into the study population, based on their outcomes. In this case, the outcomes are tweets that contain specified keywords. One of the first lessons in social science research is that the study population should not be selected based on the dependent variable(Suchman, 1962). For instance, when studying the impact of mass shootings on Twitter, one should not select users already influenced by a shooting (i.e, those who posted tweets about the shooting), and use the fact that the identified users are tweeting about the event to justify the impact.
- Content bias: Users mentioning certain keywords on Twitter after an event may be systematically biased towards mentioning the keywords in general compared to those users who did not mention the words after the event 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 with respect to their demographics. For example, males are more likely to be contained in keyword-based panels for both mass shooting and Xbox events.

In this section, we demonstrate that the three types of bias discussed above exist. For illustration, we analyze the Fort Hood mass shooting, which occurred on April 2, 2014 and was the largest mass shooting in terms of injuries in 2014, and a national Xbox advertisement that was aired on January 19, 2014.

For both event types, 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 of 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 stay 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 the three panel types–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 respective keywords within seven days before the event. For instance, the cell with horizontal value "shooting" and vertical value "kill" indicates

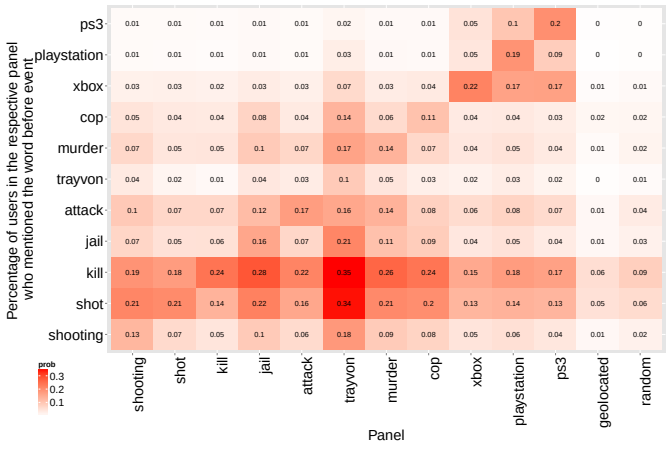

Figure 1: Proportions of users who mention the respective keywords in each panel before the occurrence of events.

the proportion of users in the keyword panel constructed using the keyword "shooting" and the keyword "kill" within seven days before event.

The *keyword-based panels* tweeted about the outcome variable more than the panels based on unrelated keywords, the random or geolocated panels. Users from the *keyword-based panel* were more likely to mention the exact keyword, and other keywords relevant to the event *before* the event. For the mass shooting *keyword-based panel*, around 30% of users mentioned the word shot and kill in the seven days prior to the event, 20% mentioned related keywords such as murder, jail, attack, and shooting itself.

Theoretically, we expect the *random panel* to exhibit minimal selection bias. Figure 1 shows the random and geolocated panels mentioned shot and kill less than any of the keyword panels. Indeed, they are similar in their discussion of the topics topics. This suggests that our *geolocated Twitter 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 selection bias in the *keyword-based panel*. A non-trivial proportion of users are discussing the keyword used to construct the panel and related keywords before the event. However, these selection issues are not prevalent the random and geolocated panels.

Further, we compare the differences in gender distribution among the three panels in Table 1. We used the Discussion Graph Tool to identify user's gender (Kıcıman et al., 2014). Across the panels, the algorithm identified the same proportion of users based on usernames, which suggests that the tool does not favor certain panel over others. The *keyword-based panels* were 65% male, while the geolocated panels were 53% male. A Pew survey of 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 Pew survey, while *keyword panel* are biased in their demographic characteristics[4].

Table 1: Gender ratio of three 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

Geolocated Panels

The previous section empirically demonstrated the biases introduced by keyword panels: users in a *keyword-based 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: scholars construct a panel of users either randomly or following some probability sampling procedure, and estimate impacts of events from those users. However, this approach has two major shortcomings. First, for many events there is little identification since the vast majority of users never discuss a particular event. Second, for many events we care about the identification of a subgroup that defines a comparison group (and in some cases a treatment and control group).

Geolocation helps with both these concerns. First, as the previous section shows, geolocated panels reduce selection bias compared to keyword-based panels. Incorporating geolocation into data collection allows scholars to create panels that are more likely to discuss a topic than random users, without using the outcome variable to construct the sample.[5] Second, when defining a treatment that varies by geography, geolocation allows comparison across treated areas.

There are five steps in the framework for *geolocated Twitter panel* construction:

1. *Control for time*: Create a full list of geolocated users within the necessary timeframe.
2. *Control for location*: Construct a panel of users who are "close" enough to the event in terms of location, and are

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

[5] Of course, users could move to a location endogenously after an event, but that is something we can define and control for in our panel selection.

thus exposed to events.

3. *Create control*: Further distinguish between those who were exposed to the event and those that were not. For instance, advertisers often choose where to air an advertisement and hence users who are nearby each other may not have the equal probability to be exposed to the event. By constructing comparison groups we can perform rigorous causal analysis. In contrast, mass shooting do not have a clear boundary, but the treatment fades with distance.
4. *Gather full tweet history*: After constructing a panel of users who are possibly "exposed" to events, scholars should further collect all their tweet history before and after event. The ability to collect ex-ante information allow scholars to compare changes of individual outcomes after an event, which is one of the major advantage of social media data over traditional social survey.
5. *Consider outcomes*: Choose the specific outcome measures for the problem of interests. This is only done after the panel is fully created.

Despite these seemingly straightforward procedures, executing these steps is not as simple as one might think. After the first step, the choices are tricky and scholars may need to make adaptations for different events. The rest of the section outlines the practical problems one may encounter when applying the framework to a specific event, and different choices a scholar can make regarding this.

Choice of geolocation

For a geolocated panel, we need to determine what location we assign to users. Then we are able to find users who are spatially proximate to events, and hence are possibly exposed to events.

To identify user locations, we can use either geolocated posts or locations that were provided in profiles or user tweets. We first discuss different choices scholars can make when they decide to use geolocated posts to define users' locations. When tweeting, users can opt to turn on location service and their tweets will hence become a geolocated tweet, which can reveal people's instant location at the level of meters and and the time of appearance at the level of seconds. Assume we have n geolocated tweets from a user which span certain time ranges (can be ex-ante or ex-post of events). Then, there are three ways to calculate users' spatial distance to events, based on different ways to define user's location.

- Instant Distance: The distance of users to the event directly during the event. Such distance reveals the distance bewteen the event and the user when the event occurred. If scholars who care about how physical exposures to events exert influences on individuals, then the precise time and location of individuals are essential for knowing whom are possibly physically exposed to events, and whom are not(Culotta, 2014; Zhang, 2015). However, using the instant location of the user when an event occurred to construct a panel of users has two disadvantages. First, many users may be at the location, but not tweet as the event unfolds. Thus, by constricting to instant location we are underestimating the group of people at the event. Second, for many events physically being there is not the key constraint but having a tie to the area. Third, instant location are endogenous to events sometimes: users may change their location, either approaching or leaving the location of event due to the event (Sundsoy et al., 2012).
- Shortest Distance: The minimal distance between an event and any geolocated tweet of the user would be meaningful in a unique event. The scholar would need to care about if the user had ever been to the location; if so, this would be the right metric.
- Distance between users' frequent locations to event: This is our final and generally preferred definition. For many events, such as mass shootings or national election etc, their influences can be spread out through nonphysical fashion, such as news-media or diffusion through friends and families. For those events, influences of events goes much 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 checked-in matters more in this situation. It may be users home or work locations.(González, Hidalgo, and Barabási, 2008; Cho, Myers, and Leskovec, 2011). One's frequent location directs correlates with the probability that one is exposed to the event, though local neighborhoods, local newspapers, news medias and other channels.

Based on the observations, we prefer to identify Twitter users' frequent locations from her history of geolocated tweets over a long time period of time (a year in our empirical analysis), and then sample users based on distances of their frequent locations to the event. In the empirical part we mapped users' geolocated posts onto census tracts, and choose the most two frequent ones as their frequent locations. This approach has the advantage that it does not require users to post a geolocated tweets nearby event when the event occurred, and hence can increase the size of study population, as well as better approximate who is actually affected by the unfolding events. Furthermore, keyword panel are often sampling over from the activity stream over a limited time window but misses users who were inactive during that time period. By identify users who stayed around the event for a much longer time frame, we are able to include users who were inactive when the event occurred.

Using geolocated tweets has the advantage of precision, but requires user to have geolocated tweets; such users are however a small proportion of Twitter users. Hence using profile location has the advantage to increase the size of study population, but it suffers from several disadvantages, such as, coding cost, reliability, and coarseness of measures (Burton et al., 2012; Hecht et al., 2011). Recent attempts that try to identify users' locations based on text and friendship patterns also have similar problems of reliability and coarsenessChandra, Khan, and Muhaya (2011); Cho, Myers, and Leskovec (2011).

Scholars can use profile location 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 influences individuals and 2) the tradeoff between granularity of geolocation measures and size of population.

Constructing Comparison Group

Another advantage of a *geolocated panel*, compared with a *keyword panel*, is that it leaves spaces to construct comparison groups such that scholars can draw objective comparisons. If the event has a geographical center there are going to be users that are more treated or, if the treatment is binary, either treated or not treated by the event.

We classify events by whether the time and place of its occurrence are endogenous to events. 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 distance from the exact location as a proxy for the quantity of treatment. For an advertisement, anyone within a Designated Market Area (DMA) is treated in the same way. Therefore there is a binary control of people within the DMA that has an advertisement and those outside of that DMA but in nearby counties.

It is possible that the places of an event are endogenous to Twitter users, such as 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 hard to prove whether crimes have negative impacts on local residents, or such impacts are due to their disadvantages situations in other aspects such as economic conditions. That means our *geolocated panel* reduces selecting-on-outcome bias of the *keyword panel*, but may introduce other demographic biases. Furthermore, users' decisions to check-in could themselves not be random(Jurgens et al., 2015).

Under this situation, *geolocated panel* still have the possibility to correct such bias during sampling procedures while it is hard to do so in 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 which satisfy the following criteria

- The two groups share the same frequent locations and hence are not biased on where they self-select to live/work.
- Two groups differ on their instant locations: one group remained at their frequent location when the event occurred, and the other groups were far away from it. The former have a higher likelihood to be exposed to events.

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

Outcomes

Scholars can come up with the outcomes that interest them. We propose some possible choices:

- Mention of event related words. This serves as a most natural way to validate that events indeed have an impact on individuals. Naturally, one would assume that users are more likely to discuss the event the more interested they are in the event.
- Sentiment scores of users tweets (Kıcıman et al., 2014). This goes past interest and relates to the nature of the impact of the event on the users.
- Movement over time of users. Scholars can can also address the movement of users over time and geolocation as events unfold.

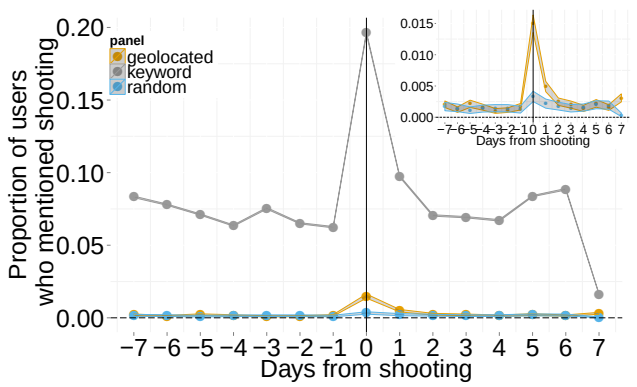

Figure 2: Proportion of users within each panel who mentioned "shooting" by days from shooting at Fort Hood on April 2, 2014. The 95% confidence interval was obtained from bootstrapping methods and plotted in shadow. The points for geolocated and random panel are overlapped together so we put a zoomed-in subplot on the top right corner.

Empirical Results

In this section we display the empirical results about how mass shootings and advertisements influence Twitter users. We start by showing the proportion of users in our panel who mentioned events before/after events, for all three types of panels: geolocated, random, and keyword-based panels. Figure 2 displays the results for the Fort Hood Mass Shooting and Figure 3 shows result for a set of same Xbox advertisements (specifically on January 12, 2014 where there was one set of local advertisements that ran concurrently in 14 DMAs). Procedures to construct respective panels are described in Section .

Figure 2 shows that the proportion of users that mentioned "shooting" on the first day of mass shootings is much higher in keyword-based panels than it is in random and geolocated panels. As expected, there is a spike in mentions of shooting at the first day for all three panels. The *keyword-based panel* evaluates the impact on users who are interested in the topic already, and hence give an estimation which is much larger in size. The real impact of events for the entire Twitter population however is greatly exaggerated in this *keyword-based panel*. It is unsurprising that the effect measured in *geolocated panel* is slightly larger than that of the *random panel*, this is by design: we limited the geolocated users to within 100 miles of the shooting, thus they are more likely to be treated by the event. Hence, we show that the *geolocated*

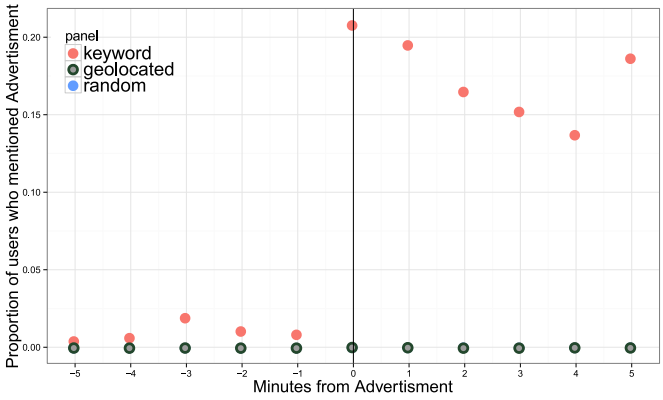

Figure 3: Proportion of users within each panel who mentioned "xbox" by hours from batch of local advertisements at 2:22 PM ET on January 12, 2014.

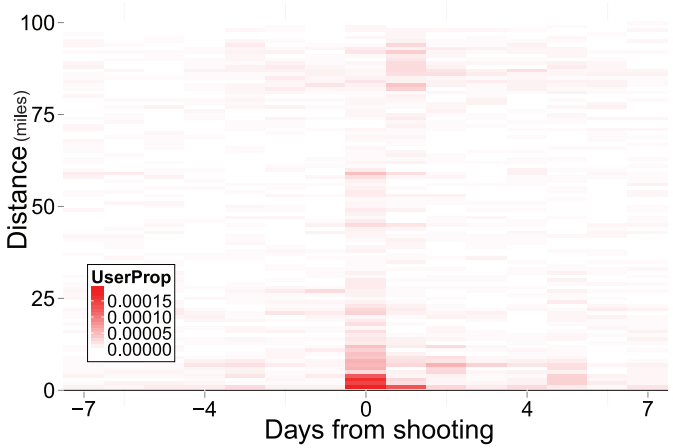

Figure 4: Proportion of users within the geolocated panel who mentioned "shooting" by days from shooting, and by distance to any of the 15 mass shootings in the US in 2014.

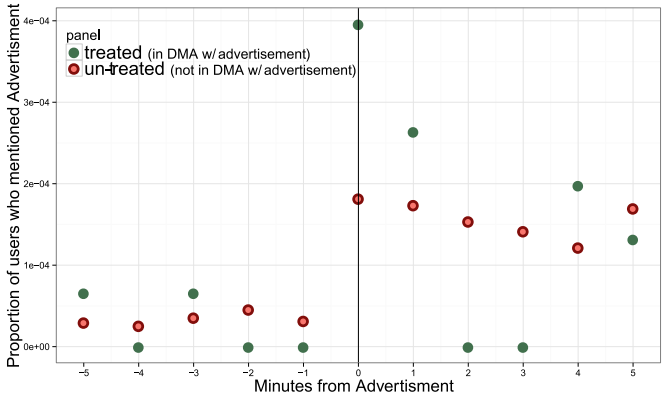

Figure 5: Proportion of users within the *geolocated panel* who mentioned "Xbox" by minutes from the batch of local advertisements at 2:22 PM ET on January 12, 2014, and within and outside of DMA that saw it.

panel are able to replicate the estimations of *random panel* but need less data.

In Figure 3, the impacts of Xbox advertisement are displayed at minutes level since previous research found that advertisement trigger responses on Twitter often after seconds of events (Kitts et al., 2014). All of the panels were reconstructed here to fit the shorter timeframe. 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 just too small to capture this rare event. The *geolocated panel* has a small bump, as we have confined the panel to the 14 DMA that received the treatment, but it is impossible to see that on this figure. Only the *keyword-based panel*, with its selection bias, shows any impact visually.

Next we show results which can only be measured meaningfully with geolocated panels: how impacts of events changed spatially. There is not enough geolocation information in a *random panel* to answer this. And the *keyword-based panel* suffers from all of the selection issues noted in the previous sections.

Figure 4 shows how the proportion of users who mentioned "shooting" changes by days to 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 are randomly distributed before events by time and spatial distance. This means that events are exogenous to our users in the *geolocated panel*. On the first day after shooting, there is a huge spike of users who discussed the event at nearest place of events (within 5 miles). The impact declines by distance quickly on the day of shooting, but still remains high compared with other dates. For all users, the impact decays quickly after first day, remaining high for the second day in only the closest regions, before approaching pre-event numbers. Impacts remains for another three days after second day only for the users who lived within 10 miles of events. The figure confirms that impacts of events decayed by spatial and temporal distance.

The results for the *geolocated panel* of Xbox are shown in Figure 5. This is the same as Figure 3, but in this figure we distinguish users in our geolocated panels by whether their frequent locations resides in one of the 14 DMA which received the advertisement. Users inside the DMA can possibly see the advertisement while users outside the DMA cannot. 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 are startling, in that the impact of the advertisement is actually much larger, as a percentage bump, towards the panel of users who may not regularly think about Xbox, compared with the *keyword-based panel*.

Lastly, Figure 6 shows how sentiment changes by days and spatial distance to mass shootings. Again, the effect is

averaging over all mass shootings. Specifically, we measure the score of fear using the Discussion Graph Tool (DGT)(Kıcıman et al., 2014). The DGT produced a joint distribution of seven types of moods–joviality, fatigue, hostility, sadness, serenity, fear, guilt–for each tweet. Here we only use only the fear scores it produced. We can see that generally the fear does not disappear even after seven days, when people are bringing up the shooting. There is a more clear pattern of decay by spatial distance after events. The ex-ante fear scores are randomly distributed by time and distance which suggest that our panel does not exhibit strong selection bias. Also, it shows that people do generally mention shooting without fear; this phenomena of fear is attached to the presence of this specific event.

Discussion

This paper argues that social media research on events should move from *keyword-based panels* to *geolocated panels*. First, cross-sectional samples without stable quantities of users conflate changes in the likeliness to opt-in to a discussion, with changes in interest and sentiment. Either of them can be important, but a stable panel can help identify each of these independently. Second, keyword samples suffer from a host of selection biases, while geolocated samples not only eliminate that problem, but also provide a useful demographic for creating comparison groups.

Geolocated panels are preferable to random panels both theoretically and empirically. First, geolocated panels make it possible for scholars to create comparison groups along geolocated treatments. Second, geolocated panels efficiently on where events have impacts. A random sample of Twitter, is sometimes too narrow to capture outcomes of interest. Geolocated panels allow scholars with restricted data access to answer questions that would be impossible to reach with a random panel. We are fortunate to work with the Twitter Firehose and examine the full Twitter stream. Geolocated panels can allow scholars without access to fight selection bias as demonstrated in this paper.

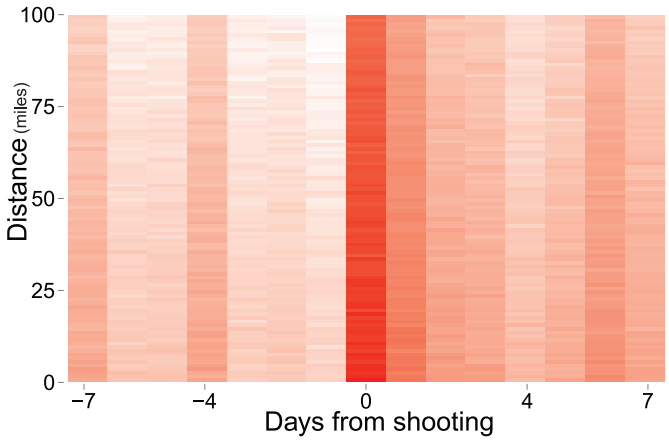

Figure 6: Predicted fear scores of users tweets which mentioned "shooting" by days from shooting, and by distance to any of the 15 mass shootings in the US in 2014.

References

An, J., and Weber, I. 2015. Whom should we sense in "social sensing" - analyzing which users work best for social media nowcasting. *EPJ Data Sci.* 4(1):1–22.

Budak, C., and Watts, D. J. 2015. Dissecting the spirit of gezi: Influence vs. selection in the occupy gezi movement. *Sociological Science* 370–397.

Burton, S. H.; Tanner, K. W.; Giraud-Carrier, C. G.; West, J. H.; and Barnes, M. D. 2012. " right time, right place" health communication on twitter: value and accuracy of location information. *Journal of medical Internet research* 14(6).

Chandra, S.; Khan, L.; and Muhaya, F. B. 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*, 838–843. IEEE.

Chen, L.; Weber, I.; and Okulicz-Kozaryn, A. 2014. Us religious landscape on twitter. In *Social Informatics*. Springer. 544–560.

Cho, E.; Myers, S. A.; and Leskovec, J. 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*, 1082–1090. ACM.

Choy, M.; Cheong, M. L. F.; Laik, M. N.; and Shung, K. P. 2011. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction.

Cui, A.; Zhang, M.; Liu, Y.; Ma, S.; and Zhang, K. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 1794–1798. New York, NY, USA: ACM.

Culotta, A. 2014. Reducing Sampling Bias in Social Media Data for County Health Inference. *Joint Statistical Meetings Proceedings*.

De Choudhury, M.; Diakopoulos, N.; and Naaman, M. 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*, 241–244. ACM.

Diaz, F.; Gamon, M.; Hofman, J.; and Rothschild, D. 2014. Online and social media data as a flawed continuous panel survey.

DiGrazia, J.; McKelvey, K.; Bollen, J.; and Rojas, F. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE* 8(11).

Doré, B.; Ort, L.; Braverman, O.; and Ochsner, K. N. 2015. Sadness shifts to anxiety over time and distance from the national tragedy in newtown, connecticut. *Psychological science* 26(4):363–373.

Dwi Prasetyo, N., and Hauff, C. 2015. Twitter-based Election Prediction in the Developing World. In *the 26th ACM Conference*, 149–158. New York, New York, USA: ACM Press.

Gayo-Avello, D. 2011. Don't turn social media into another 'literary digest' poll. *Communications of the ACM* 54(10):121–128.

González, M. C.; Hidalgo, C. A.; and Barabási, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. *Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles*. the dynamics of the location field in user profiles. New York, New York, USA: ACM.

Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47(4):67.

Jurgens, D.; Finethy, T.; McCorriston, J.; and Xu, Y. T. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 9th ...*.

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

Kıcıman, E.; Counts, S.; Gamon, M.; De Choudhury, M.; and Thiesson, B. 2014. Discussion graphs: Putting social media analysis in context. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 243–248. ACM.

King, G.; Keohane, R. O.; and Verba, S. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

Kitts, B.; Bardaro, M.; Au, D.; Lee, A.; Lee, S.; Borchardt, J.; Schwartz, C.; Sobieski, J.; and Wadsworth-Drake, J. 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*, 1–9. ACM.

Kohavi, R.; Longbotham, R.; Sommerfield, D.; and Henne, R. M. 2009. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery* 18(1):140–181.

Lampos, V., and Cohn, T. 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*.

Lehmann, J.; Gonçalves, B.; Ramasco, J. J.; and Cattuto, C. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, 251–260. ACM.

Lin, Y.-R.; Margolin, D.; Keegan, B.; and Lazer, D. 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*, 737–748. International World Wide Web Conferences Steering Committee.

McMinn, A. J.; Moshfeghi, Y.; and Jose, J. M. 2013. Building a large-scale corpus for evaluating event detection on twitter. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, 409–418. New York, NY, USA: ACM.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11(122-129):1–2.

Oktay, H.; Taylor, B. J.; and Jensen, D. D. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, 1–9. ACM.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 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)*, number EPFL-CONF-203561.

Park, M.; Weber, I.; Naaman, M.; and Vieweg, S. 2015. Understanding musical diversity via online social media. In *Ninth International AAAI Conference on Web and Social Media*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860. ACM.

Sang, E. T. K., and Bos, J. 2012. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, 53–60. Association for Computational Linguistics.

Schonlau, M.; Van Soest, A.; Kapteyn, A.; and Couper, M. 2009. Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research* 37(3):291–318.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 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):e19467.

Skoric, M.; Poor, N.; Achananuparp, P.; Lim, E.-P.; and Jiang, J. 2012. Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2583–2591. IEEE.

Suchman, E. A. 1962. An analysis of bias in survey research. *Public Opinion Quarterly* 26(1):102–111.

Sundsoy, P. R.; Bjelland, J.; Canright, G.; Engo-Monsen, K.; and Ling, R. 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*, 586–590. IEEE.

Thelwall, M.; Buckley, K.; and Paltoglou, G. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology* 62(2):406–418.

Tsytsarau, M.; Palpanas, T.; and Castellanos, M. 2014. Dynamics of news events and social media reaction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 901–910. ACM.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Wang, W.; Rothschild, D.; Goel, S.; and Gelman, A. 2014. Forecasting elections with non-representative polls. *International Journal of Forecasting*.

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

Zhang, H. 2015. Witnessing political protest on civic engagement and political attitudes: A natural experiment.