Selection Bias in Event Studies with Twitter: How Geolocated Panel Can Help

Han Zhang, Princeton University
Shawndra Hill & David Rothschild, Microsoft Research (NYC)
Goal

• Understand population impacts of online/offline events through social media data
  • Electoral events on twitter mentions: Diaz et al. (2015)
  • Mass shooting on sadness: Dore et al. (2015)
  • Televised program on Twitter discourse: Walker and Muchnik (2013)
  • Natural disasters on public concerns: Signorini et al. (2011)
Why Social Media

• Why social media data?
  
  • Previous approach: social media data as a convenient sample
    
    • low time latency (Google Flu Trends) and low monetary cost (electoral studies)
    
    • Dilemma: offline data already exist
  
  • Our approach: social media data as a biased, but useful or even necessary data source over survey data:
    
    • Social survey are often not good at studying impacts of unexpected events:
    
    • Hard to construct study population in a post-hoc fashion.
    
    • Lack of ex-ante information: hard to compare changes
Twitter Data Collection

• Stream API: streams of new tweets; can filter by keywords

• Search API:
  • can search by keywords (only goes back to 1 week).
  • can search by lat/long (only goes back to 2 weeks).
  • cannot search by demographic characteristics, such as residences, genders etc.

• Firehose API: entire Twitter platform.
Keyword Cross-section

- Filter tweets that contain event relevant keywords/hashtags from streaming API.
- Analyze shifts of volumes of tweets mentioning events.

<table>
<thead>
<tr>
<th>Time 0</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>M-1</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a tweet with event-related keywords</td>
<td>This is a tweet with event-related keywords</td>
<td>...</td>
<td>This is a tweet with event-related keywords</td>
<td>This is a tweet with event-related keywords</td>
<td>This is a tweet with event-related keywords</td>
</tr>
</tbody>
</table>
Example

Daily volumes of mentions of tweets that contain hashtags related to the Gezi protests in Turkey (Budak and Watts, 2015)
Problems

• Tweet-based:
  • cross-section responses of varying quantities of users, rather than panels of the same users
  • Lack historical tweets: whether users change?
  • From tweet-based to user-based panels.
Twitter as an “panel”

• Each tweets are generated by responses to latent questions

• Users can response in any time at multiple times

• Users engagement varies considerably

• **Problem: how to construct ex-post panels to study events**
Keyword Panel

- From tweets to users:
  - identify users who discuss the event from keyword cross-section; then collect historical tweets
  - each user may have vary number of tweets discussing the event.
- Advantage: can compare ex-post with ex-ante levels.
Still Problematic

- **Selection Bias**: Users in the keyword panel are different from a random sample of Twitter users in three ways:
  
  - Demographics: users in the keyword panel maybe systematically different than other Twitter users before events.
  
  - Selecting on outcome: use outcome keywords to select user themselves.
  
  - Content: users in the keyword panel are more likely to discuss relevant topics before event.
  
  - Are observed shifts driven by people who already interested in the topic, hence lacks external validity?
Random Panel

- Randomly sample users
- Hunt for tweets that are relevant to events.
- High time/monetary cost vs. high representativeness.
Use geolocation to construct a panel of users that are “close” to event in terms of time and space, and are likely to be “exposed” to event.
Features

• Pros
  • Reduce selection bias:
    • if users’ residences are exogenous to events (mass shooting).
    • In some situations, users’ residences are endogenous to events (e.g., local homicides)
  • Potential to construct comparison group
  • Trade-off between time cost and representativeness.
  • Can be matched with offline population survey such as census

• Cons:
  • Restrict to users with geolocated tweets
Data and Events

• Data: Twitter Firehose

• Events:

  • 14 large mass shootings in USA in 2014.

• Advertisement:

  • One large national TV advertisement on Xbox on January 19, 2014.

  • A batch of local ads that were shown at same time in 14 market areas (DMA) on January 12, 2014.
Empirical Analysis

1. Compare selection bias in keyword, geolocated, and random panels

2. Analyze different outcomes for three panels
Constructing Ex-post Panels

• Keyword panel: Use event related keywords to filter users who mentioned the keyword within 7 days.

• Random Panel: randomly select users and collect their tweets

• Geolocated Panel:
  • Population: twitter users with more than 5 geolocated tweet
  • Map geo-tweets of users into census tracts, and identify their frequent census tracts as home/work location.
  • Sample users who live within 100 miles of the shooting.
### Table 2: Gender ratio of three panels

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

- Users in keyword panels are significantly more likely to mention relevant words, and being males, even before event.

- Users in geolocate panel are similar to random panels: minimal selection bias.
Results: mention of Xbox
Effect by distance for geolocated panel: shootings
Effect by distance for geolocated panel: Ads

- treated (in DMA w/ advertisement)
- untreated (not in DMA w/ advertisement)

Graph showing proportion of users who mentioned advertisement vs. minutes from advertisement.
Fear after shooting
Findings

• Geolocated panel can replicate results of random panel: a population based measure

• Keyword panel gives much higher estimate than the real impacts on the entire population
Details of Constructing a Geolocation Panel

1. Determine users’ location and sample user based on their distance to event:
   1. Geolocated tweets vs. profile locations
   2. Instance location vs. frequent location
2. Treatment/control group:

- Binary treatment: Users inside DMA can see Xbox advertisements.

- Continuous treatment: No cut-off for mass shooting. The closer to shootings, the higher the likelihood to be influenced

3. Endogenous event: events whose occurrences are not geographically random (e.g., local homicides)

   1. Construct comparison groups within a geolocated panel (no hope in keyword panels).

      1. first find users living in the same neighborhood (no selection now)

      2. distinguish between users by their instant location
Take away

• Taxonomy of ex-post panel construction with social media data

• Selection bias in ex-post panels:
  • Keyword cross-section/panel create selection bias
  • Geolocated panel can reduce selection bias and efficiently construct samples to study events simultaneously

• Geolocated panel can give population-based estimate of impacts of events.
Questions

• Han Zhang: hz2@princeton.edu

• http://www.princeton.edu/~hz2