Research Design and Computational Methods

Jae Yeon Kim

18 March, 2019

Going beyond data mining

Data mining: the process of divovering patterns in large data sets

But, mostly for associations. Big data could be a boon for p-hacking. (Remember, the statistical power is a function of the size of the data set.). Don't be fooled by statistical **significance**.

Goodhart's law: "When a measure [p-value] becomes a target, it ceases to be a good measure" (rephrased by Marilyn Strathern). FYI, what the heck is p-value?

For a credible research, we need to think more than associations. (e.g., Threats to internal and external validity by Campbell and Stanley)

For causal arguments, you need to think about interventions and counterfactuals.

Research Design and Computational Methods

▶ Research design: questions -> answers

▶ Research design is critical to identify causality in your project

▶ Computational methods:

  ▶ Along with qualitative, and quantitative methods, one of many tools that will help your journey.
  ▶ Research design is a compatative edge for social scientists (vis-à-vis data scientists) in doing computational social science.

Lesson objectives

1. How can computational methods lead to a new discovery?
2. How can leverage the strength of the data collected by computational methods?
3. How can we mix (including computational) methods?

Big descriptive inference

Yes, I know social media is a narrow sample. But what if my social media data covers 3 million posts…

How Censorship in China Allows Government Criticism but Silences Collective Expression

GARY KING   *Harvard University*
JENNIFER PAN   *Harvard University*
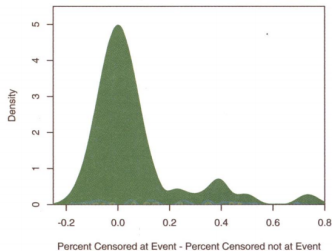MARGARET E. ROBERTS   *Harvard University*

*W*e offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to clearly expose government intent.*
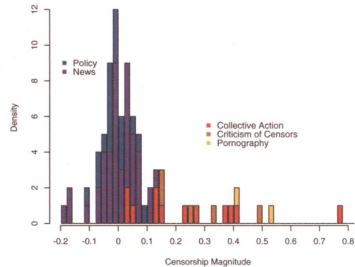
Research design

- ▶ Question: How Does Online Censorship in China Work?
- ▶ Platform: 1,400 Chinese social media services
- ▶ Unit of analysis: Online posts
- ▶ Data collection:
  - ▶ web crawling locating, accessing, and downloading posts from these from many Web sites before ISP or the government censored them.
  - ▶ N: 3,674,698
- ▶ Data analysis:
  - ▶ Stratied random sampling on hypothesized political sensitivity (created by reviewing prior literature, consulting with China specialists, and studying current events)
  - ▶ N: 127,283

Evidence

Figure 3. "Censorship Magnitude," The Percent of Posts Censored Inside a Volume Burst Minus Outside Volume Bursts.

(a) Distribution of Censorship Magnitude

(b) Censorship Magnitude by Event Type

Online Data Are Good and Bad at the Same Time
How to build a panel data using Web data

sociological science

Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement
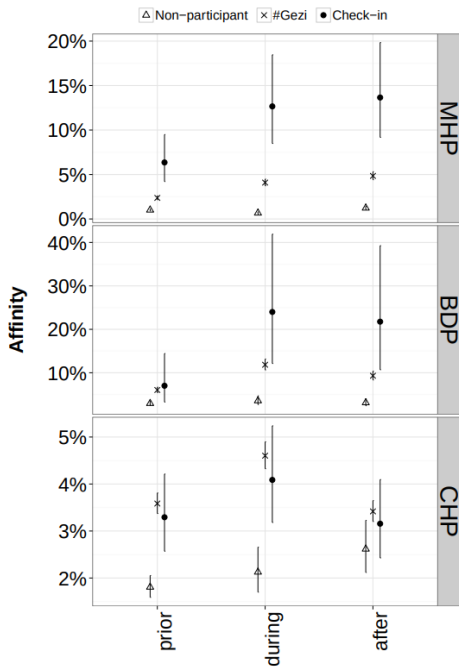
Ceren Budak, Duncan J. Watts

Microsoft Research

Abstract: Do social movements actively shape the opinions and attitudes of participants by bringing together diverse groups that subsequently influence one another? Ethnographic studies of the 2013 Gezi uprising seem to answer "yes," pointing to solidarity among groups that were traditionally indifferent, or even hostile, to one another. We argue that two mechanisms with differing implications may generate this observed outcome: "influence" (change in attitude caused by interacting with other participants); and "selection" (individuals who participated in the movement were generally more supportive of other groups beforehand). We tease out the relative importance of these mechanisms by constructing a panel of over 30,000 Twitter users and analyzing their support for the main Turkish opposition parties before, during, and after the movement. We find that although individuals changed in significant ways, becoming in general more supportive of the other opposition parties, those who participated in the movement were also significantly more supportive of the other parties all along. These findings suggest that both mechanisms were important, but that selection dominated. In addition to our substantive findings, our paper also makes a methodological contribution that we believe could be useful to studies of social movements and mass opinion change more generally. In contrast with traditional panel studies, which must be designed and implemented prior to the event of interest, our method relies on ex post panel construction, and hence can be used to study unanticipated or otherwise inaccessible events. We conclude that despite the well known limitations of social media, their "always on" nature and their widespread availability offer an important source of public opinion data.

Keywords: social networks; social movements; ex post panels; Twitter

Research design

- ▶ Question: Do social movements actively shape the opinions and attitudes of participants by bringing together diverse groups that subsequently influence one another? -> How can we tease "influence" from "selection" mechanism?
- ▶ Platform: Twitter
- ▶ Unit of analysis: individual users
- ▶ N: 30,000
- ▶ Ex-post panel construction:
    - ▶ Following both participants (also there's a subcategory divided by verified physical variable) of the movement (Occupy Gezi movement) and non-participants
    - ▶ Based on their retweet behavior in the 17 months leading up to the Gezi movement (before), during the movement (during), and the five months after the movement (after)
    - ▶ Trade-off between external and internal validity

Evidence

Going into the causal direction

▶ "Onilne" survey and lab experiments: web technology is used for recruitment and sampling and that's it.

▶ Online field experiments: "experiments that leverage platforms or systems that already exist on the Internet to study motivations and behaviors of individuals, organizations, and even governments".

▶ "The treatments and outcome of online field experiments are often native to the online setting." (Muise and Pan 2018, 2-5)

Sources of Authoritarian Responsiveness: A Field Experiment in China

Jidong Chen   Beijing Normal University
Jennifer Pan   Stanford University
Yiqing Xu   Massachusetts Institute of Technology

A growing body of research suggests that authoritarian regimes are responsive to societal actors, but our understanding of the sources of authoritarian responsiveness remains limited because of the challenges of measurement and causal identification. By conducting an online field experiment among 2,103 Chinese counties, we examine factors that affect officials' incentives to respond to citizens in an authoritarian context. At baseline, we find that approximately one-third of county governments respond to citizen demands expressed online. Threats of collective action and threats of tattling to upper levels of government cause county governments to be considerably more responsive, whereas identifying as loyal, long-standing members of the Chinese Communist Party does not increase responsiveness. Moreover, we find that threats of collective action make local officials more publicly responsive. Together, these results demonstrate that top-down mechanisms of oversight as well as bottom-up societal pressures are possible sources of authoritarian responsiveness.

Research design

- ▶ Question: measuring the authoritarian government responsiveness
- ▶ Platform: county-level government websites (identified by the web-crawling scripts)
- ▶ Unit of analysis: county
- ▶ N: 2,103
- ▶ Randomization and treatment: the authors randomly assigned message types to county websites.
  - ▶ Treatment: 1) proxying collective action potential and 2) the threat of tattling to higher-level authorities , and party loyalty
  - ▶ Control
- ▶ Outcomes: 1) whether the message received a response, 2) how many days took to get it, 3) content of the response

FIGURE 4 The Causal Effects of Treatments on Responsiveness