Carnegie Mellon

Thesis Proposal

Institute for Software Research

Societal Computing

institute for SOFTWARE RESEARCH

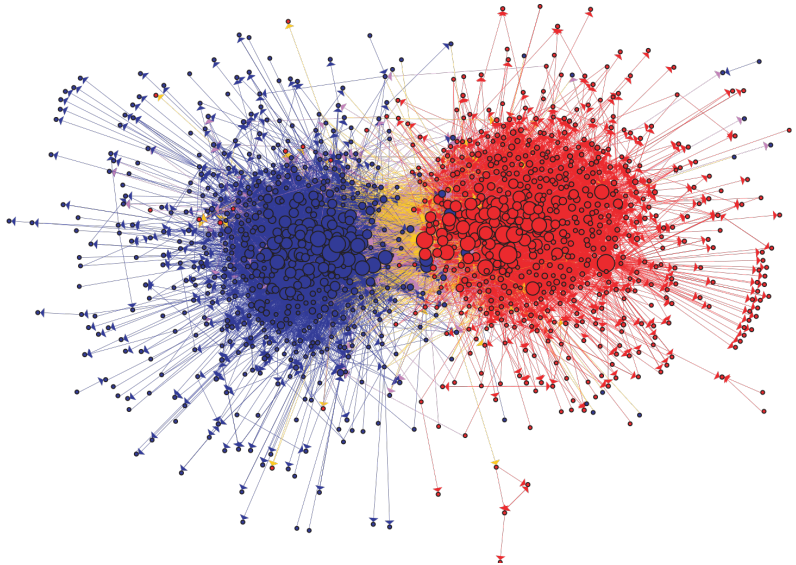
Bias and beyond in digital trace data

Momin M. Malik

Friday, 19 May 2017
9.15 am – 12 pm
Gates-Hillman Center 6501

Committee:
Dr. Jürgen Pfeffer *(co-chair)*, ISR, SCS, CMU
Dr. Anind K. Dey *(co-chair)*, HCII, SCS, CMU
Dr. Cosma Rohilla Shalizi, Statistics, CMU
Dr. David Lazer, Northeastern University

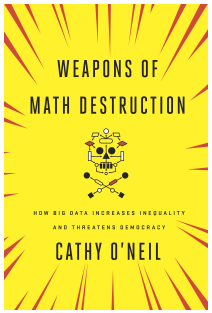
Digital trace data



"We check our ***e-mails*** *regularly, make* ***mobile phone calls****... We may post* ***blog entries*** *accessible to anyone, or maintain friendships through* ***online social networks****. Each of these transactions leaves* ***digital traces*** *that can be compiled into comprehensive pictures of both individual and group behavior, with the* ***potential to transform our understanding of our lives, organizations, and societies.****"*

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Alstyne, M. A. (2009). Computational social science. *Science*, *323*(5915), 721–723.

What could go wrong?



Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls

Zeynep Tufekci
University of North Carolina, Chapel Hill
zeynep@unc.edu

Commentary

Big Data and the danger of being precisely inaccurate

Big Data & Society
July–December 2015: 1–4

DOI: 10.1177/2053951715602495
bds.sagepub.com
SAGE

Big Data, Digital Media, and Computational Social Science: Possibilities and Perils

By
DHAVAN V. SHAH, JOSEPH N. CAPPELLA, and W. RUSSELL NEUMAN

Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites

By
ESZTER HARGITTAI

DOI:10.1145/2001269.2001297

The power to predict outcomes based on Twitter data is greatly exaggerated, especially for political elections.

BY DANIEL GAYO-AVELLO

Don't Turn Social Media Into Another 'Literary Digest' Poll

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

Classifying Political Orientation on Twitter: It's Not Easy!

Raviv Cohen and **Derek Ruths**
School of Computer Science
McGill University
raviv.cohen@mail.mcgill.ca, derek.ruths@mcgill.ca

danah boyd & Kate Crawford

CRITICAL QUESTIONS FOR BIG DATA
Provocations for a cultural, technological, and scholarly phenomenon

SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths[1]*** *and* **Jürgen Pfeffer[2]**

different social media platforms (*8*). For instance, Instagram is "especially appealing to

The rise of "embedded rese
searchers who have special

Why does it matter?

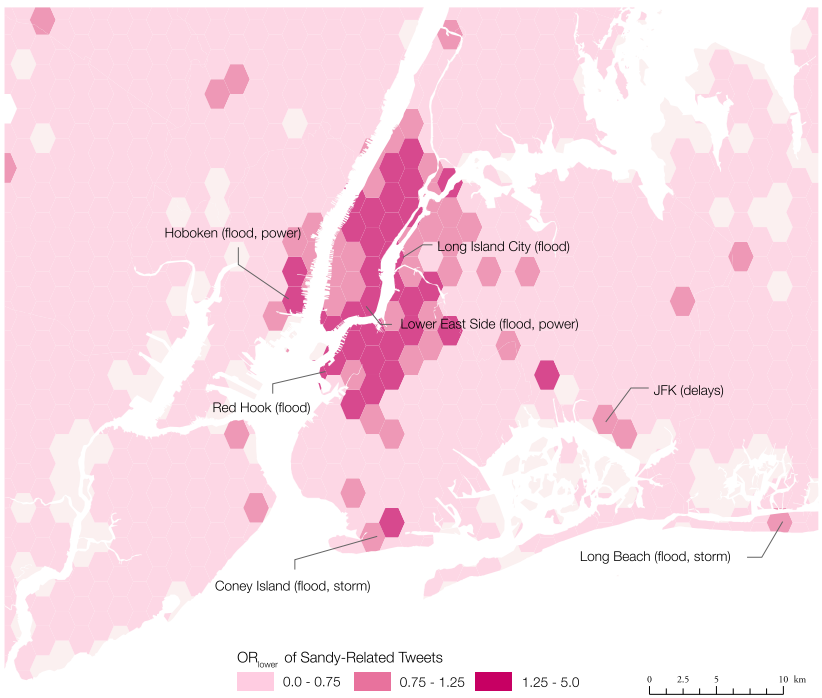
"In the early days of the program, Street Bump found something fascinating: **there were more potholes reported in wealthy areas of the city than in poor ones.**"

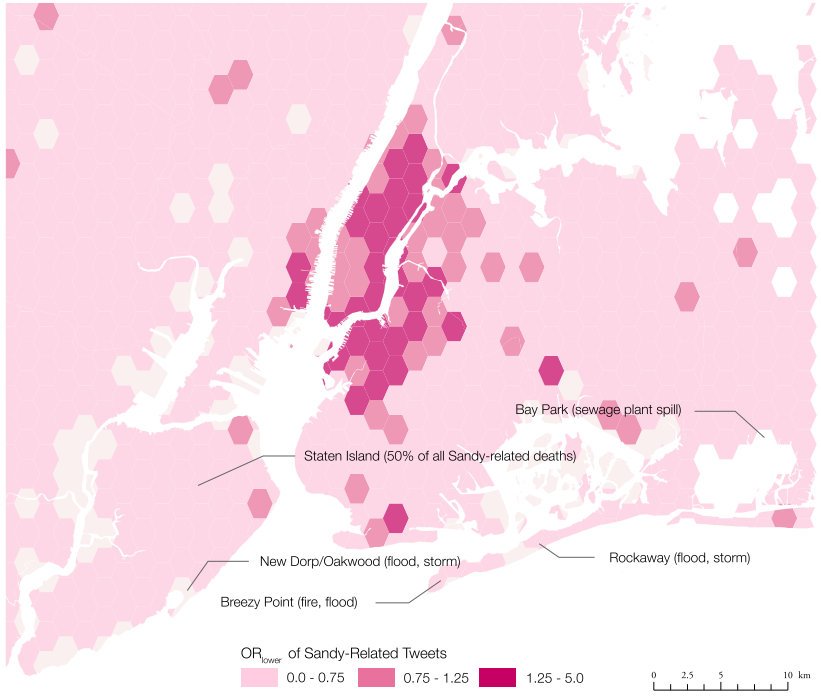
Why?

"**Wealthy people were far more likely to own smart phones and to use the Street Bump app.** Where they drove, potholes were found; where they didn't travel, potholes went unnoted."

Christopherson, E. G. (2013, July 25). Confronting the data dilemma. Rita Allen Foundation.

Why does it matter?





Shelton, T., Poorthuis, T., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. Geoforum, 52, 167–179.

Why does it matter?

Identifying Networks of Criminals

"Facebook has helped me by identifying suspects that were friends or associates of other suspects in a crime and all brought in and interviewed and later convicted of theft and drug offenses."

"My biggest use for social media has been to locate and identify criminals. I have started to utilize it to piece together local drug networks."

LexisNexis® Risk Solutions (2014). Survey of law enforcement personnel and their use of social media.

Goal:	Predict, Monitor, and Prevent Risk In/Around Protests
Anticipated Activity:	Protests, Riots, Looting
Overt Threats:	Unions, Activist Groups, Etc.
Locations:	Schools, Public Spaces, Malls, High-Rent Districts
Actions Taken:	During Event(s), Post-Event

Ozer, N.(2016, September 23). Police use of social media surveillance software is escalating, and activists are in the digital crosshairs. American Civil Liberties Union of Northern California.

Authorization and authentication based on an individual's social network

US 9432351 B2

ABSTRACT

In particular embodiments, a method includes receiving a request for a first user to access a loan from a lender, the request identifying a user identifier (ID) of the first user; determining whether the first user is authorized to access the loan based at least in part on a gray list comprising user IDs of the users who are not authorized to access loans, wherein the gray list is based on a black list; and permitting the loan to be accessed by the first user if the first user is authorized to access the loan based on the gray list.

Publication number	US9432351 B2
Publication type	Grant
Application number	US 14/299,391
Publication date	Aug 30, 2016
Filing date	Jun 9, 2014
Priority date ⓘ	Jul 22, 2004
Also published as	CN101036366A, 26 More »
Inventors	Christopher Lunt
Original Assignee	Facebook, Inc.
Export Citation	BiBTeX, EndNote, RefMan

Patent Citations (81), Non-Patent Citations (13), Classifications (23)

External Links: USPTO, USPTO Assignment, Espacenet

Why does it matter?

What can we do?

Reducing biases and flaws in social media data

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
  a. Corrects for platform-specific and proxy population biases
  OR
  b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
  a. Shows results for more than one platform
  OR
  b. Shows results for time-separated data sets from the same platform
- 7. For new methods: compares results to existing methods on the same data
- 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct data sets (one of which was not used during classifier development or design)

- Study the nature of social platforms
- Extend investigations to the next big frontier of social (network) data: sensors
- *Define scopes and design studies around what data can tell us*

Ruths, D. ,& Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Thesis statement

Social media and sensor data do not give unbiased, generalizable findings about human behavior: inferences about constructs are complicated by selection bias, medium-specific norms and culture, and algorithmic user manipulation, and raw measurements are of physical quantities rather than of causal underlying social constructs. But by studying these forms of bias and the data-generating processes of such data and understanding their limitations, we can establish proper scopes and study designs within which findings will be accurate, reliable, and fair for use in business decision-making, scientific research, and public policy.

Thesis statement (breakdown)

Social media and sensor data do not give unbiased, generalizable findings about human behavior.

- Inferences about constructs are complicated by:
  - Selection bias (Chapter 1)
  - Medium-specific norms and culture
  - Algorithmic user manipulation (Chapter 2)
- Raw measurements are of physical quantities rather than of causal underlying social constructs. (Chapter 3)

But by studying these forms of bias and the data-generating processes of such data and understanding their limitations, we can

- Establish proper scopes (Chapter 4-5)
- Create appropriate study designs (Chapter 6)

And get findings that are accurate, reliable, and fair for use in business decision-making, scientific research, and public policy.

Outline

Part I: Critiques

1. Demographic biases
2. Platform effects
3. Sensors for social network data collection

↓

(Central argument)

↓

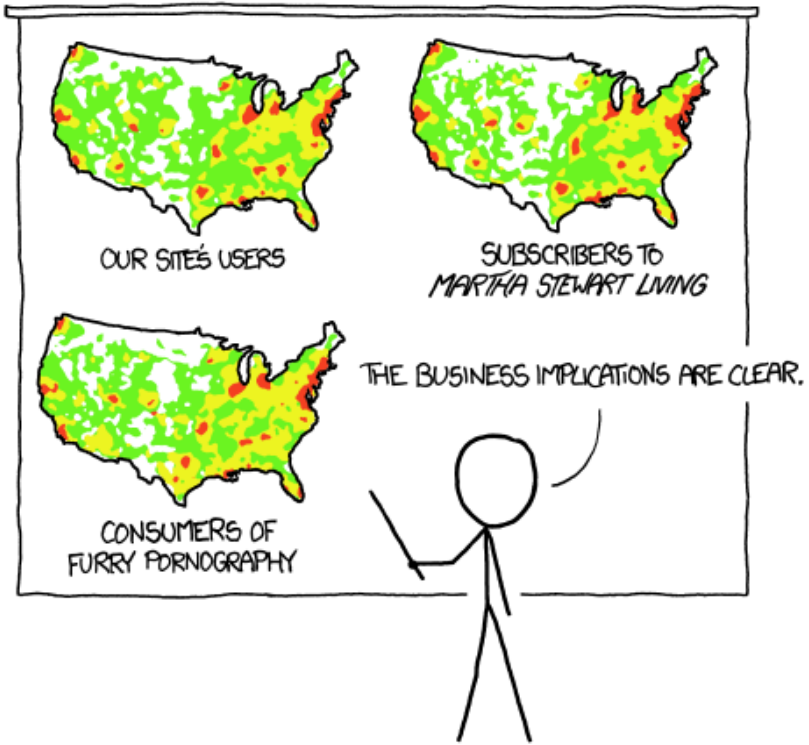
Part II: Responses

4. Studies of small communities
5. Social media for applied research and interventions
6. Cohort studies

Part I: Critiques


YOU KEEP ON USING THESE DATA
I DO NOT THINK THEY MEAN WHAT YOU THINK THEY MEAN

Chapter 1: Demographic biases



Introduction

PART I: CRITIQUES

Demographic biases

Platform effects

Sensors and social networks

PART II: RESPONSES

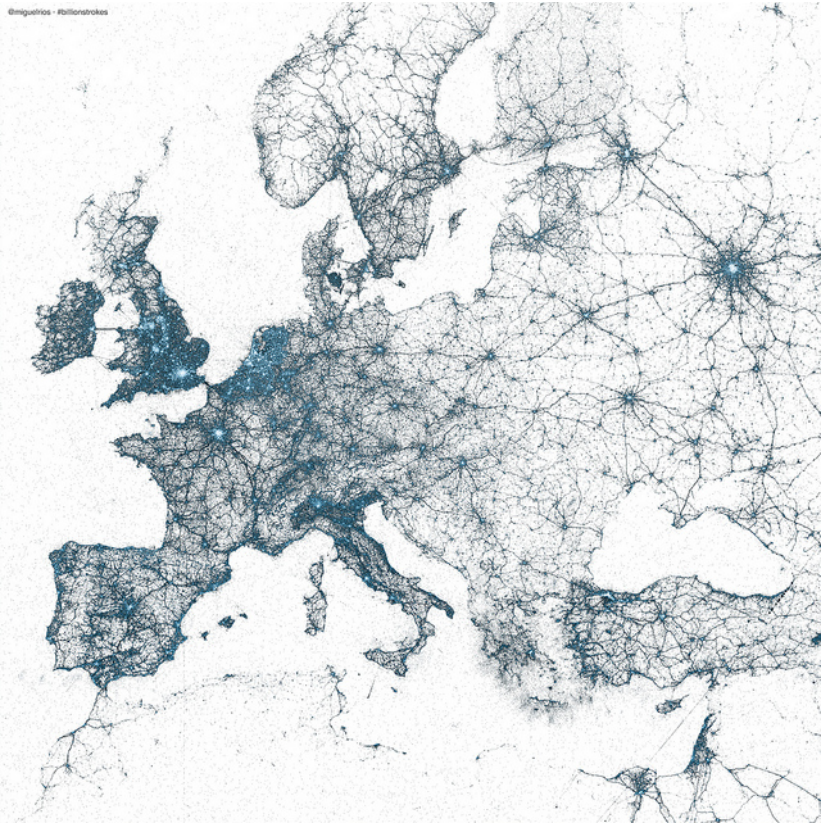
Small communities

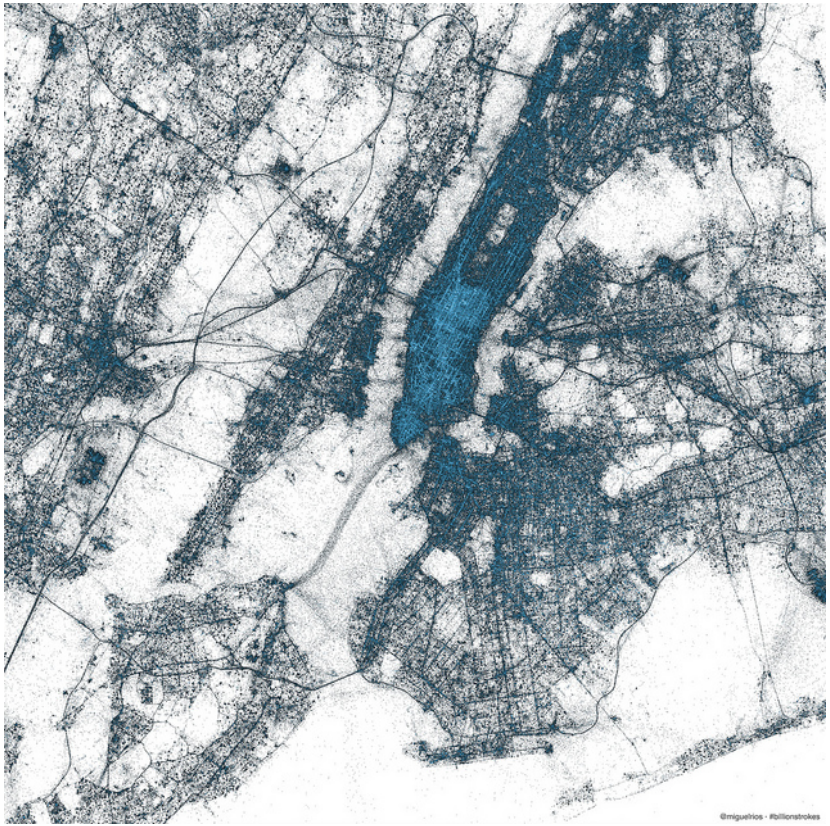
Intervention design

Sensor-based study design

Conclusion

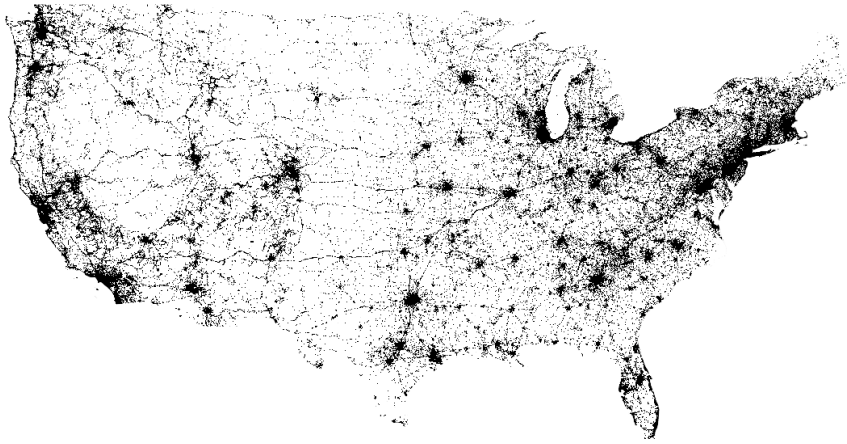
Chapter 1: Demographic biases





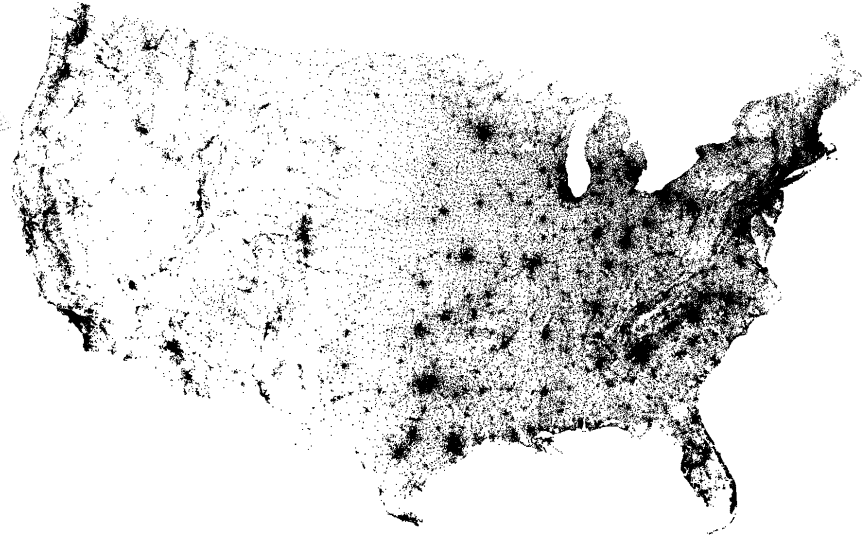
Chapter 1: Demographic biases

Geotagged tweets



Adapted from 'Contiguous United States geotag map (2009)' by Eric Fischer (https://www.flickr.com/photos/walkingsf/5985800498)

Population



Population density in 2010 US Census. Adapted from 'Nighttime Population Distribution Wall Map' by Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau. Each square represents 1,000 people.

Chapter 1: Demographic biases

- **Completed.** Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. (2015). Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)* (pp. 18–27).
- Do geotagged tweets represent the population?
- Geotagged tweets used to study mobility, urban life, transportation, natural disaster crisis response, public health, and more
- Null hypothesis: users of geotagged tweets are distributed randomly over the US population

Chapter 1: Demographic biases

- Noise proportional to population, $U = \alpha P + \varepsilon P$, and take a log transformation,

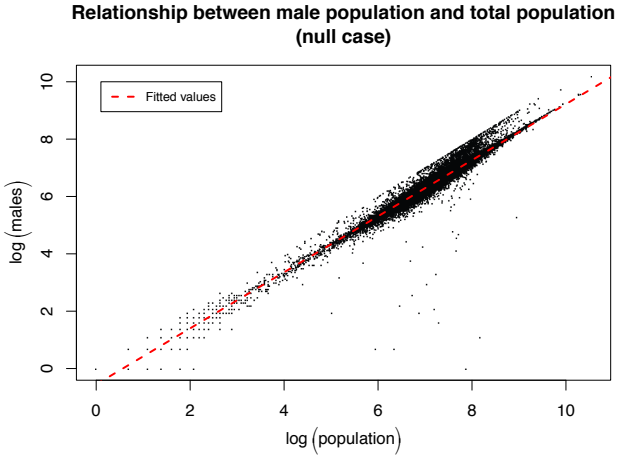
$$\log U = \log \alpha + \log P + \varepsilon'.$$

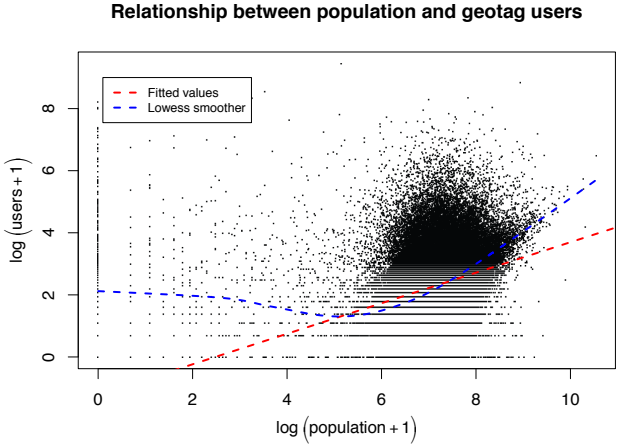
- For linear model

$$\log U = \beta_0 + \beta_1 \log P + \varepsilon',$$

- test H_0: $\beta_1 = 1$, $\exp(\beta_0)$ should be the proportion.

Chapter 1: Demographic biases





- Spatial multivariate modeling reveals the nature of biases: towards Asian, Hispanic, and black populations; towards the coasts, and especially the east coast; higher income; urbanization; and more young people.
- (This does not take into account systematic differences among geotagged tweet users vs. non-users)

Chapter 1: Demographic biases

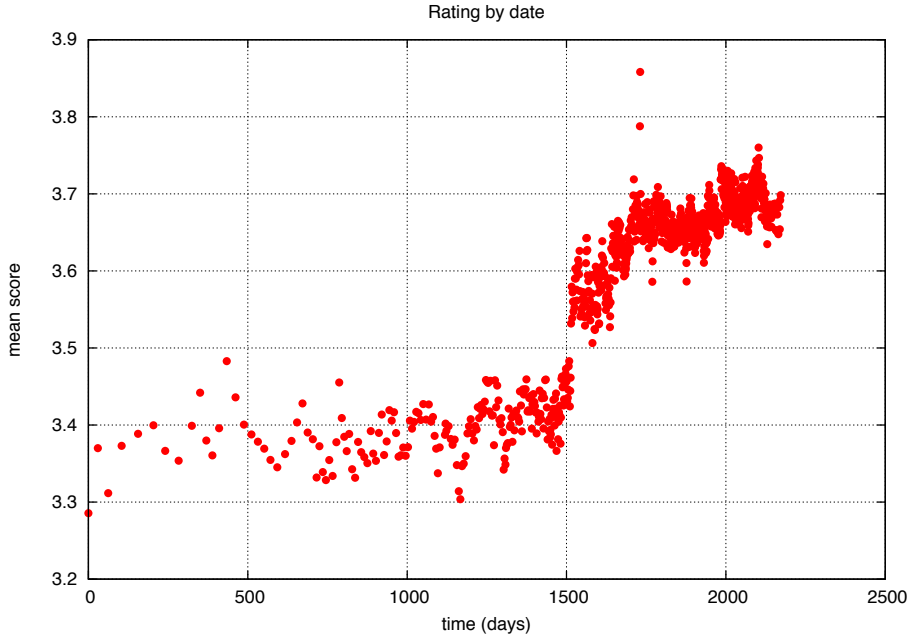
Lessons:

- Geotagged tweets are not representative!
- So, they will not give unbiased, generalizable findings about human behavior

Contributions:

- First country-wide, multivariate, spatial model linking Census and social media data

Chapter 2: Platform effects



Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

Koren, Y. (2009). The BellKor solution to the Netflix Grand Prize.

Chapter 2: Platform effects

- **Completed.** Momin M. Malik and Jürgen Pfeffer. (2016). Identifying platform effects in social media data. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM-16)* (pp. 241–249).
- Social media platforms not neutral utilities, but commercial entities
- Platforms are engineered to manipulate user behavior
- When we measure behavior, what are we really measuring?
- Data artifacts can give a valuable peek inside what's happening
- Case: Facebook's "People You May Know" recommendation system
- Facebook New Orleans crawl from Viswanath et al. (2009)

Chapter 2: Platform effects

- Regression Discontinuity (RD) Design is the use of a treatment that is effective strictly above some cutoff value c of a covariate X_i, so $T = \mathbf{1}(X_i > c)$.

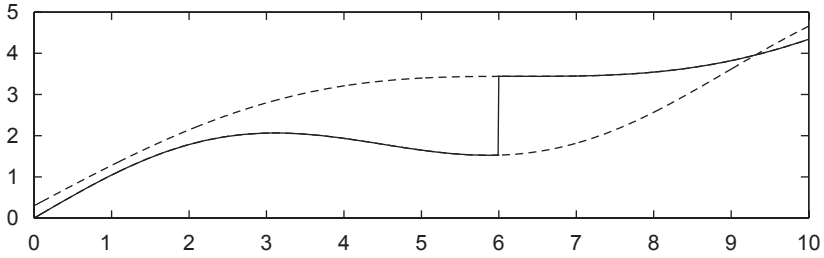


Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

- Linear univariate case:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 T + \beta_3 x_i T + \varepsilon_i$$

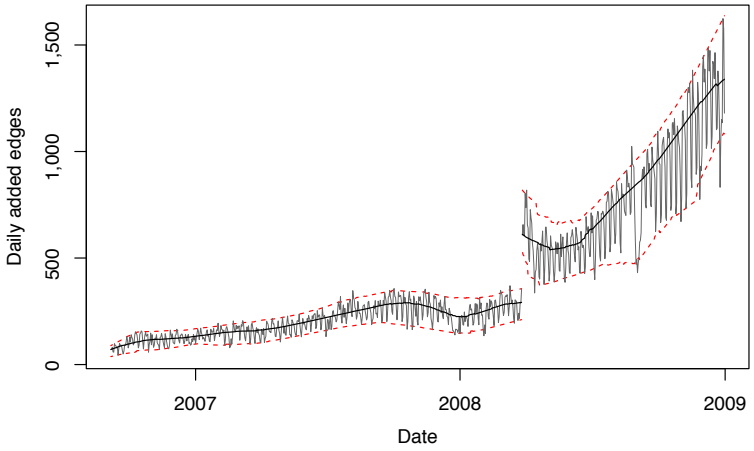
- Change at the point is $\beta_2 - \beta_3 c$, so we can estimate local causal impact

Chapter 2: Platform effects

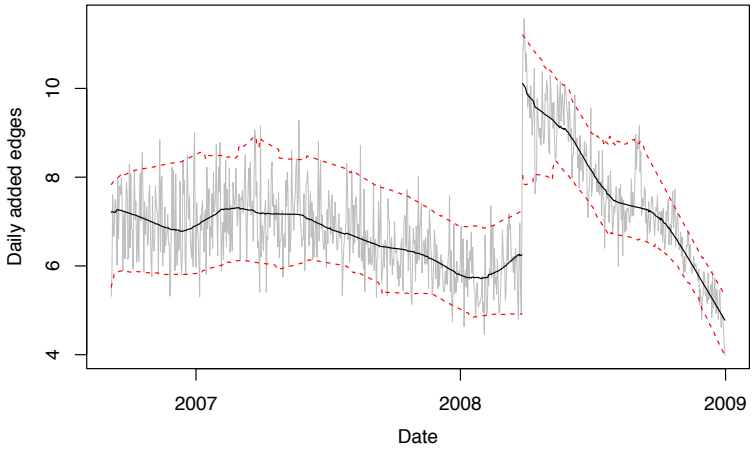
- For time series, appropriate analog is Interrupted Time Series
- ITS doesn't have the same framework as regression discontinuity
- No reason that the discontinuity can't be in time, but
  - Temporal processes (moving average, temporal autocorrelation)
  - Cause deflated standard errors
- The variance of a time series is strongly centered on an autocorrelated mean; I use quantile regression (5% and 95%) to get tolerance bands as a way to incorporate the variance

Chapter 2: Platform effects

- Facebook links: +300 new edges per day (~200%)



- Triangles: +3.8 triangles per edge (~64%)



Chapter 2: Platform effects

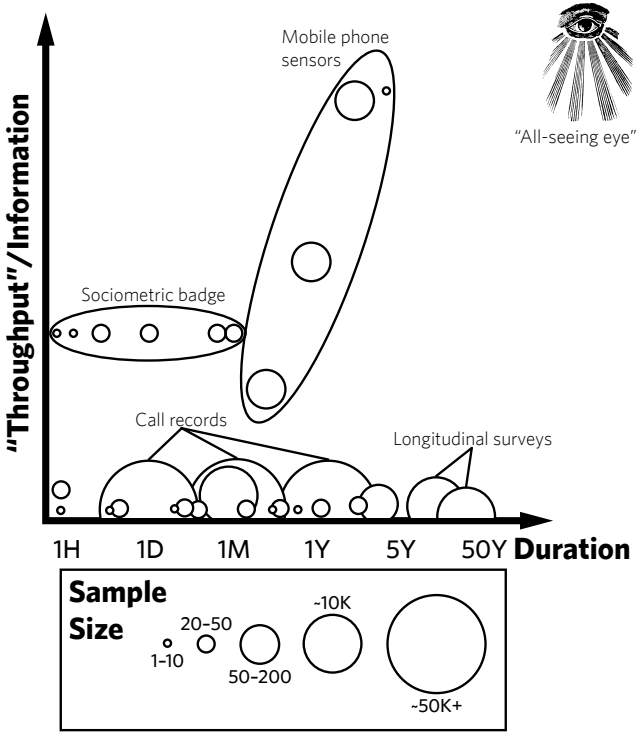
Lesson:

- Facebook networks reflect platform engineering
- Facebook networks do not give an unbiased, generalizable look at human social networks

Contributions:

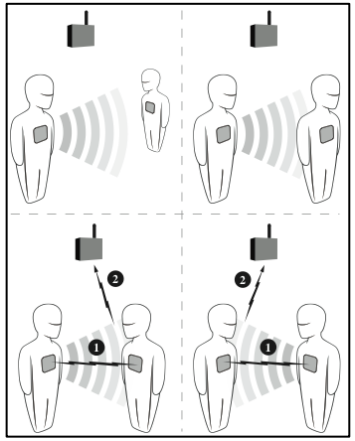
- Perhaps the first empirical demonstration of algorithmic user manipulation as a platform effect
- Provides an external measurement of the effect of a triadic closure-based recommendation systems on user behavior
- Model of applying causal inference designs in social media
- Theorizing the use of data artifacts as valuable opportunities for study

Chapter 3: Sensor data



Aharony, N., Pan, W., Ip, C., Khayal, I., & Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing, 7*(6), 643–659.

- What social networking behavior can sensors measure?
- Mobile phone sensors and sensor badges measure proximity, not interaction, and not actual friendship (an underlying psychological state)



Panisson, A., Gauvin, L., Barrat, A., & Cattuto, C. (2013). Fingerprinting temporal networks of close-range human proximity. *Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications (PERCOM) Workshops* (pp. 261-266).

Chapter 3: Sensor data

- Can we use proximity/interaction to get friendship? Proximity plays a role in friendship (called "propinquity"), but is not the only thing.

Festinger, L., Back, K. W., & Schachter, S. (1950). *Social pressure in informal groups: A study of human factors in housing.* Stanford, CA: Stanford University Press.

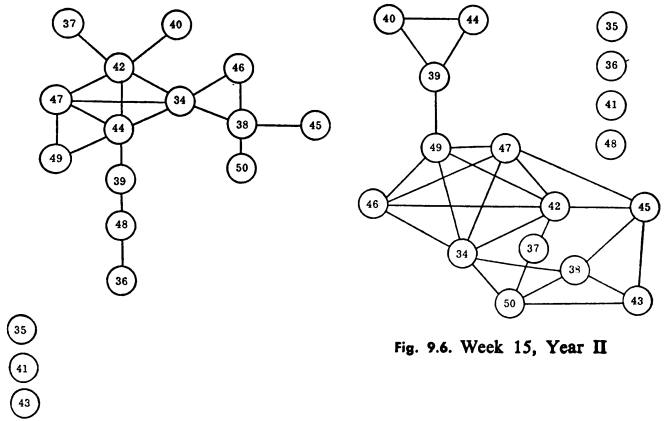


Fig. 9.6. Week 15, Year II

Fig. 9.5. Week 1, Year II

Sociograms showing all pairs and larger sets of individuals whose attraction relationships reach the 95-percent criterion. Unconnected individuals have no such relationships.

Newcomb, T. M. (1961). *The acquaintance process.* New York, NY: Holt, Reinhard & Winston.

Chapter 3: Sensor data

Work in progress. (With Jürgen Pfeffer, Afsaneh Doryab, & Anind Dey)

- Critique of existing sensor studies
- Link to work from the 1940s and 1950s examining basic friendship formation mechanisms
- Link also to work on social network data quality ("informant accuracy") from the 70s and 80s
- Compare surveys and sensor data as preliminary findings from a cohort study (Chapter 6)

Chapter 3: Sensor data

Anticipated lessons:

- Sensor data are raw measurements are of physical quantities, not causal underlying social constructs
- Physical quantities are theoretically interesting, and deserve examination on that basis

Anticipated contributions:

- First theorization of mobile phone sensor data in terms of social network analysis
- Tying together sensor and SNA literature

Introduction

PART I: CRITIQUES

Demographic biases

Platform effects

Sensors and social networks

PART II: RESPONSES

Small communities

Intervention design

Sensor-based study design

Conclusion

Part II: Responses

(Coffee break, and any questions so far!)

Part II: Responses


ONE DOES NOT SIMPLY
GENERALIZE

Central argument

Central argument

- Survey data:
  - Unrepresentative samples addressed with sampling strategies and weighting
  - Respondent biases addressed with careful survey design and delivery
- For digital trace data, *such technical approaches will not necessarily work*
  - Uncertainty about sampling frames, unknown biases, platform effects
- Study such problems, but also, work to *establish proper scopes*
- Figure out what we can study with available digital trace data, rather than how to use available digital trace data to study what we want
  - And to study what we want, design data collection procedures

Chapter 4: Studies of small communities

- Ethnography, case studies, "small-*n* analysis": for meaning-making, patterns, similarities and contrasts in (small, defined) social groups or settings
- No attempt to (formally) generalize to a (hypothetical) larger population
- Exploring the *range of behavior and meaning-makings*
- Why can't we do something similar with computational approaches? Does large-scale data have to be generalizable?

Chapter 4: Studies of small communities

- Erowid.org, an online drug-information portal active for 20 years
- Over 15 years of reader-submitted "experience reports," an important (but so far mostly unmined) source of psychopharmacological data

Chapter 4: Studies of small communities

Work in progress. (With Hemank Lamba, Jürgen Pfeffer, and TBD Erowid contributors)

- Components:
  - Topic modeling, temporal keyword extraction
  - Internal influence of experience reports?
  - Co-usage networks over time
- Especially with small communities, important to be *ethical*
  - Consult with relevant community, do relevant analyses, respect norms
  - Give credit to content producers, curators, and maintainers
  - Give opportunities for feedback

Chapter 4: Studies of small communities

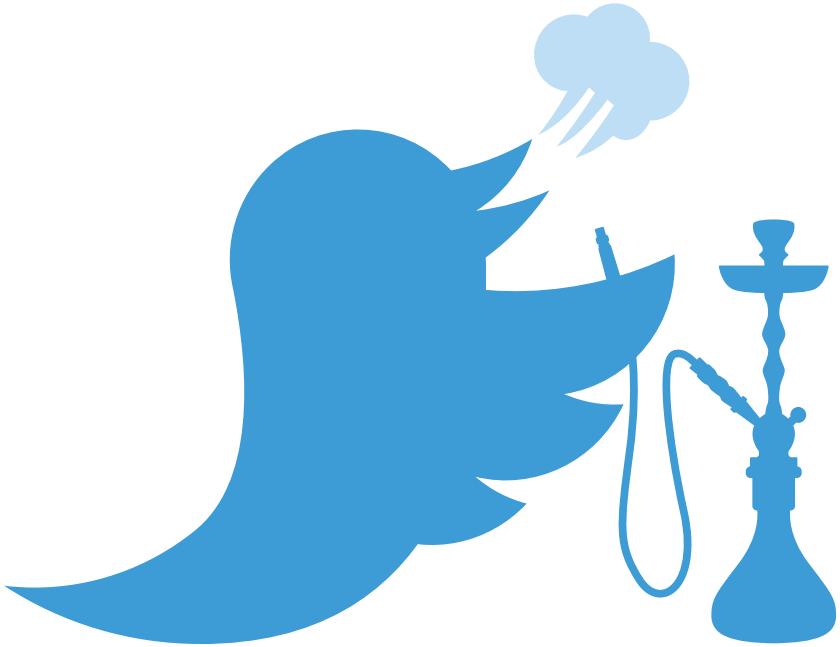
Lessons:

- Study interesting communities
- Studies need to involve, and contribute to, the studied to be ethical

Contributions:

- Demonstration of ethical digital media research
- Exploratory research for psychopharmaceutical literature
- Domain-relevant application of topic modeling

Chapter 5: Social media for applied research and interventions



- Waterpipe Tobacco Smoking (WTS), i.e. hookah/shisha, on the rise among young people
- Young people more likely to be on Twitter
- Health risks largely unknown and underappreciated to this population
- Previous public health messaging in print, television, billboards
- Why not use Twitter as well?
- Social media marketing-style approach

Chapter 5: Social media for applied research and interventions

Work in progress. (With Jason B. Colditz, Kar-Hai Chu, Brian Primack, & Anind Dey.)

- With University of Pittsburgh's Center for Research on Media, Technology, and Health
- I contribute: (1) correct use of social media data, (2) correct machine learning procedures, (3) explaining use of machine learning

Chapter 5: Social media for applied research and interventions

1) Correct use of social media:
   - Changed from basic research to monitoring
   - Streaming API as mimicking application setting
2) Correct use of machine learning:
   - Labeling two weekends of tweets to labeling tweets randomly selected over 6 months
   - Test on temporally held-out data
   - Use trained classifier for active learning
3) Explain use of machine learning:
   - "Supervised learning" better than hand-coding everything
   - Domain-specific meanings (e.g., 💨, "dash") need domain coding
   - Estimation not interesting, since *n*-grams are not causal
   - Random forest performs better than 'usual' social science classifiers

Chapter 5: Social media for applied research and interventions

Lesson:

- Platform-specific interventions avoid problems of biases

Contributions:

- Demonstrating an appropriate use of Twitter
- Aiding public health researchers in machine learning
- Aiding WTS researchers

Chapter 6: Sensor-based study design

- How can we use large-scale trace data not (just) opportunistically, but purposively?
- Sensors provide opportunities for careful study design
- Sensor studies so far: not connected to theoretical questions, statistical tools

Chapter 6: Sensor-based study design

Work in progress. (With Michael Merrill, Afsaneh Doryab, & Anind Dey.)

- AWARE Framework for mobile phone data collection
- Instrumented 53 members of a 79-person fraternity for two months
- Combined with three waves of social network surveys

Chapter 6: Sensor-based study design

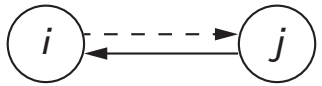
- What kind of model do we apply to these data?
- Considered: Network autocorrelation, MRQAP, logistic regression, p1, p2, p*/ERGMs, Latent Space Models
- Stochastic Actor-Oriented Models (SAOMs) appropriate
- Extract meaningful summaries of nodal and dyadic behaviors (e.g., co-location) and use in model

Chapter 6: Sensor-based study design

$$p_i(\tilde{x}|x,z) = \begin{cases} \exp(f_i(\tilde{x},z)) / \sum_{x' \in \mathcal{A}_i(x)} \exp(f_i(x',z)) & \text{if } \tilde{x} \in \mathcal{A}_i(x), \\ 0 & \text{otherwise.} \end{cases}$$

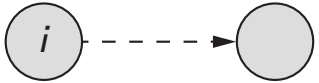
$f_i(x,z) = \sum_k \beta_k s_{ik}(x,z)$ is a linear combination of actor choice statistics, representing network and behavioral processes. E.g., for adjacency matrix A,

reciprocity:

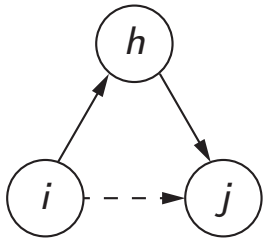


$$s_{i,recip} = \sum_j A_{ij} A_{ji}$$

homophily:



transitivity (friend-of-a-friend):



$$s_{i,trans} = \sum_{j,h} A_{ij} A_{ih} A_{hj}$$

Chapter 6: Sensor-based study design

Anticipated lessons:

- Findings on propinquity
- Demonstration of theory-driven sensor study design

Anticipated contributions:

- Link sensors to theoretical questions
- Application of SAOMs to sensor data
- Finding relevant to college fraternities and other similar groups

Conclusions

Conclusions

Thesis statement (breakdown), redux:

Social media and sensor data do not give unbiased, generalizable findings about human behavior.

- Inferences about constructs are complicated by:
  - Selection bias (Chapter 1)
  - Medium-specific norms and culture
  - Algorithmic user manipulation (Chapter 2)
- Raw measurements are of physical quantities rather than of causal underlying social constructs. (Chapter 3)

But by studying these forms of bias and the data-generating processes of such data and understanding their limitations, we can

- Establish proper scopes (Chapter 4-5)
- Create appropriate study designs (Chapter 6)

Conclusions

- Social media and sensor data do not give unbiased, generalizable findings about human behavior
  - Inferences about constructs are complicated by selection bias, medium-specific norms and culture (Chapter 1)
  - Algorithmic user manipulation (Chapter 2)
  - Raw measurements are of physical quantities rather than of causal underlying social constructs (Chapter 3)
- In light of these limitations, do
  - "Small-*n*"-style framing (Chapter 4)
  - Interventions (Chapter 5)
  - Primary data collection (Chapter 6)
- Altogether, I hope I have demonstrated challenges to generalizability when using large-scale trace data, and promising alternative approaches

Timeline

- April 15, 2017: Sensor study data collection completed!
- May 19, 2017: Propose
- (June-August, 2017: Internship)
- June 2017: Complete Twitter Waterpipe Tobacco Smoking paper
- July 2017: Complete Erowid analysis
- August 2017: Submit applied sensors paper to Ubicomp
- September 2017: Look at the possibility of doing another cohort study at CMU as a subset of an expanded "Life@CMU" study
- October-November 2017: Submit a sensors paper for journal *Social Networks* focusing on theory
- December 2017: Complete thesis writing
- January-February 2018: Defend

Thank you!