# POIR 613: Measurement Models and Statistical Computing

**Pablo Barberá**

School of International Relations
University of Southern California
pablobarbera.com

Course website:
pablobarbera.com/POIR613/

# Today

1. Computational social science research: challenges and opportunities
2. Discussion: ethics of Big Data research.
   - Kramer et al 2014 (and “Editorial Expression of Concern”)
   - Tufekci 2014
3. Scraping data from the web

# Logistics

1. Referee reports:
   - You should all have already signed-up
   - Due day before class at 8pm
2. Class project:
   - One-paragraph idea due September 15
3. No office hours tomorrow

# Computational Social Science

## Shift in communication patterns

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Behavior, opinions, and latent traits

- Digital footprints: check-ins, conversations, geolocated pictures, likes, shares, retweets, ...

→ Non-intrusive measurement of behavior and public opinion

Toole et al (2015): "Tracking employment shocks using mobile phone data"
Beauchamp (2016): "Predicting and Interpolating State-level Polls using Twitter Textual Data"

# Behavior, opinions, and latent traits

- Digital footprints: check-ins, conversations, geolocated pictures, likes, shares, retweets, ...

→ Non-intrusive measurement of behavior and public opinion

→ Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, ...

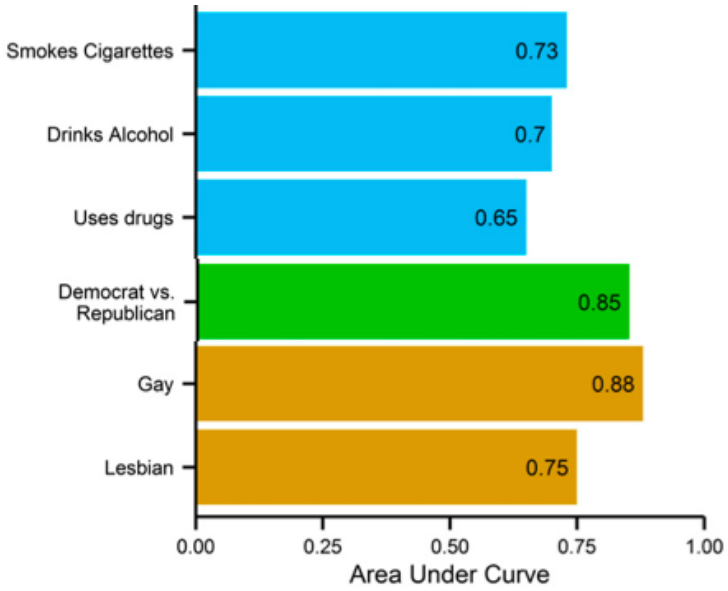



**Fig. 2.** Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

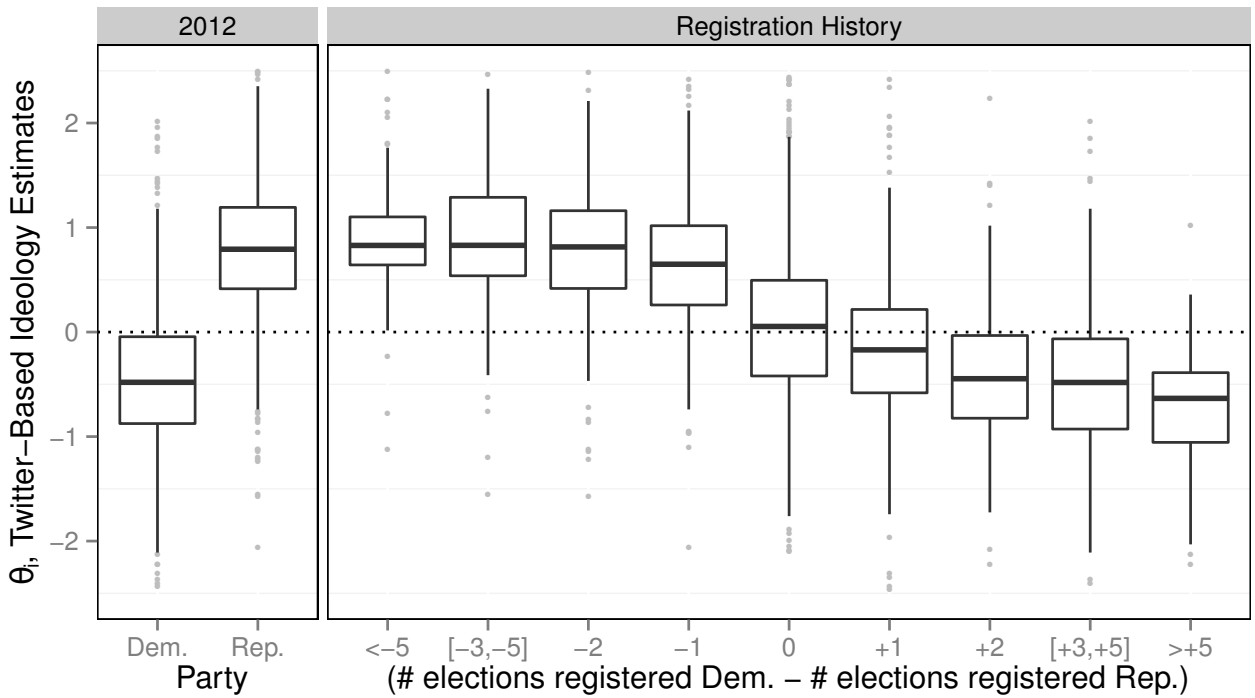
Kosinki et al, 2013, "Private traits and attributes are predictable from digital records of human behavior", *PNAS* (also personality, PNAS 2015)

# Behavior, opinions, and latent traits

- Digital footprints: check-ins, conversations, geolocated pictures, likes, shares, retweets, . . .

→ Non-intrusive measurement of behavior and public opinion

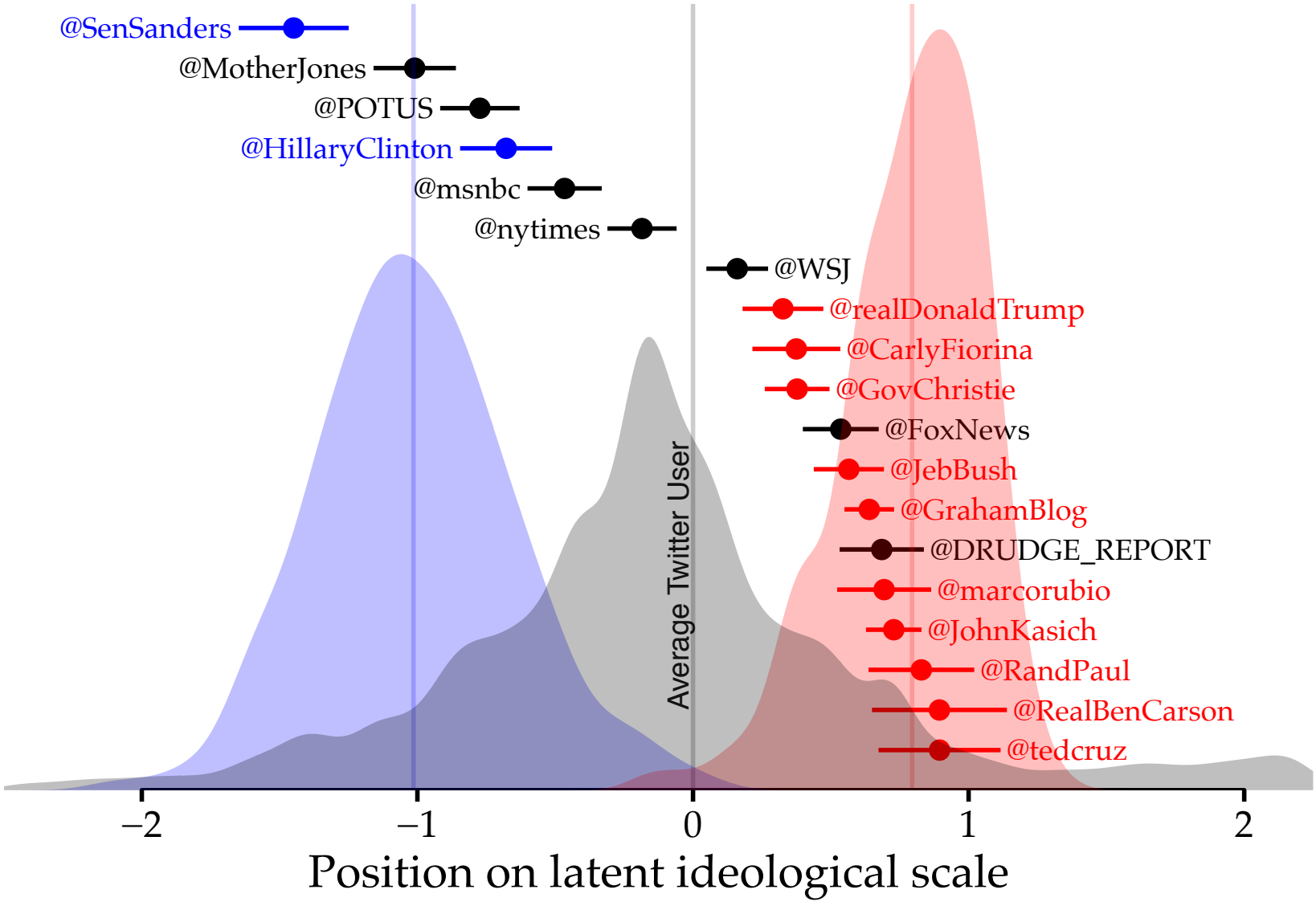
→ Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, . . .




Data: 2,360 Twitter accounts, matched with Ohio voter file.

Barberá, 2015, "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data", *Political Analysis*

# Estimating political ideology using Twitter networks




Barberá "Who is the most conservative Republican candidate for president?" *The Monkey Cage / The Washington Post*, June 16 2015

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Interpersonal networks

- Political behavior is social, strongly influenced by peers

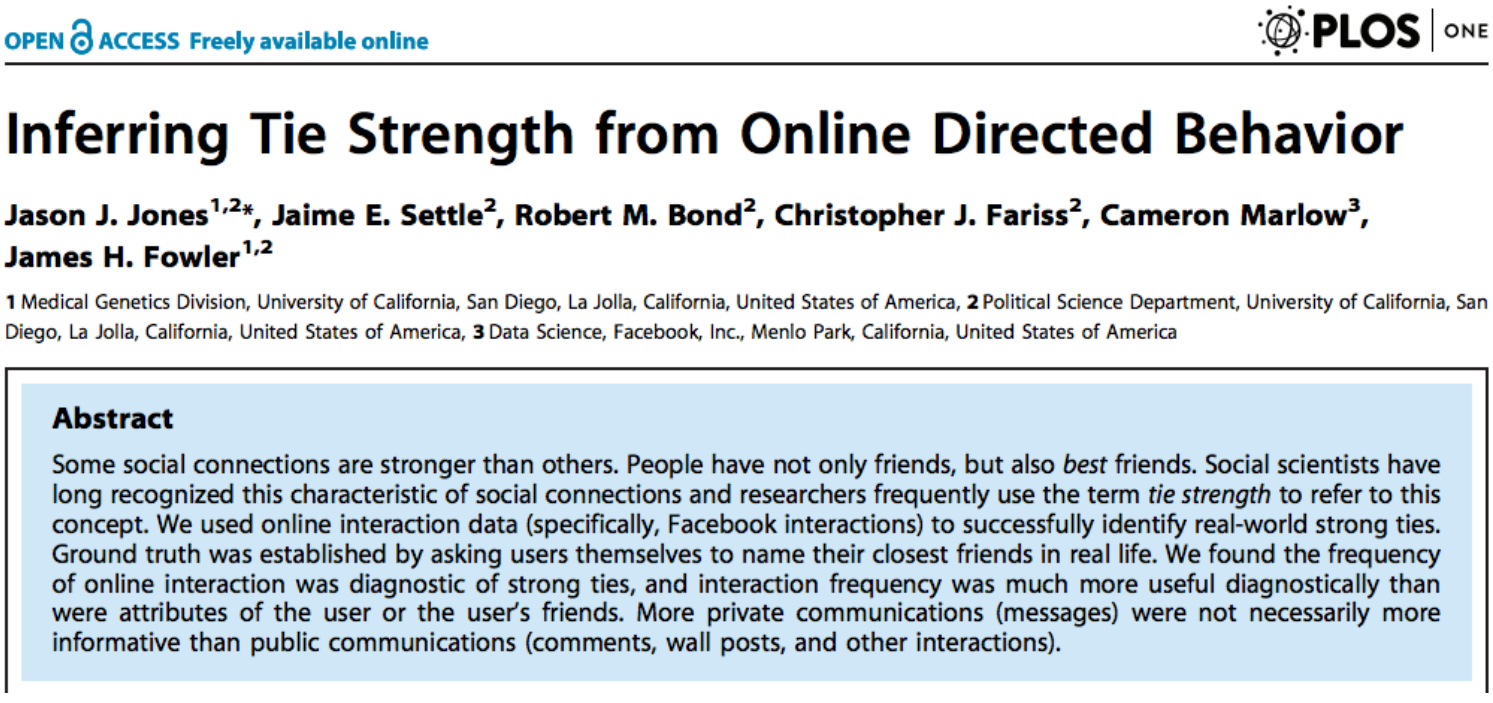

Bond et al, 2012, "A 61-million-person experiment in social influence and political mobilization", *Nature*

# Interpersonal networks

- Political behavior is social, strongly influenced by peers
- Costly to measure network structure

# Interpersonal networks

- Political behavior is social, strongly influenced by peers
- Costly to measure network structure
- High overlap across online and offline social networks



OPEN ACCESS Freely available online

PLOS ONE

## Inferring Tie Strength from Online Directed Behavior

**Jason J. Jones[1,2]*, Jaime E. Settle[2], Robert M. Bond[2], Christopher J. Fariss[2], Cameron Marlow[3], James H. Fowler[1,2]**

**1** Medical Genetics Division, University of California, San Diego, La Jolla, California, United States of America, **2** Political Science Department, University of California, San Diego, La Jolla, California, United States of America, **3** Data Science, Facebook, Inc., Menlo Park, California, United States of America

**Abstract**

Some social connections are stronger than others. People have not only friends, but also *best* friends. Social scientists have long recognized this characteristic of social connections and researchers frequently use the term *tie strength* to refer to this concept. We used online interaction data (specifically, Facebook interactions) to successfully identify real-world strong ties. Ground truth was established by asking users themselves to name their closest friends in real life. We found the frequency of online interaction was diagnostic of strong ties, and interaction frequency was much more useful diagnostically than were attributes of the user or the user's friends. More private communications (messages) were not necessarily more informative than public communications (comments, wall posts, and other interactions).

Jones et al, 2013, "Inferring Tie Strength from Online Directed Behavior", *PLOS One*

# Interpersonal networks

- Political behavior is social, strongly influenced by peers
- Costly to measure network structure
- High overlap across online and offline social networks
- Application: social networks, contagion, and diffusion (Fowler, Centola, Aral)

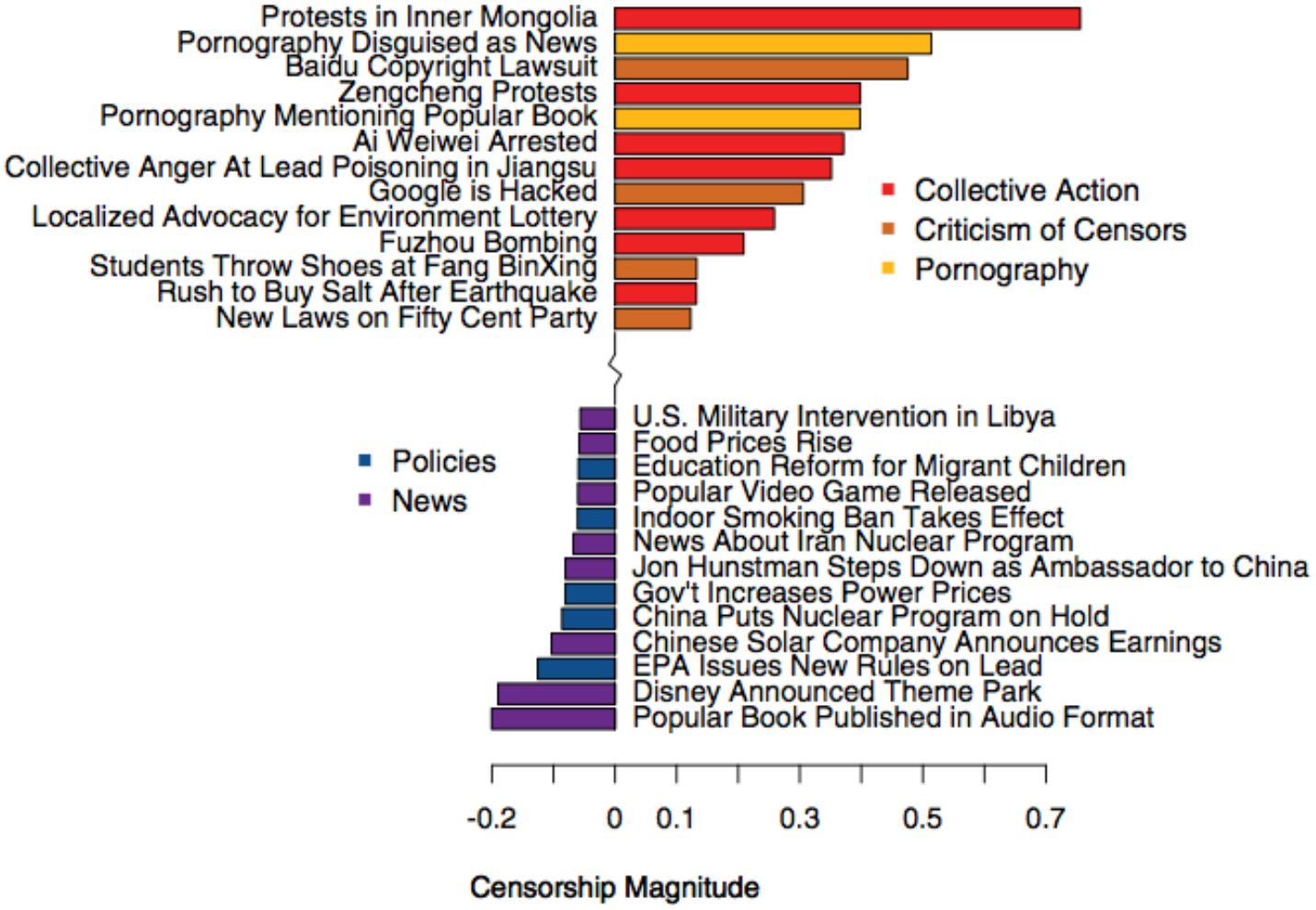
# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Elite behavior

- Authoritarian governments' response to threat of collective action




King et al, 2013, "How Censorship in China Allows Government Criticism but Silences Collective Expression", *APSR*

# Elite behavior

- Authoritarian governments' response to threat of collective action
- Estimation of conflict intensity in real time

Journal of Conflict Resolution
55(6) 938-969

DOI: 10.1177/0022002711408014
http://jcr.sagepub.com
SAGE

**Using Social Media to Measure Conflict Dynamics: An Application to the 2008–2009 Gaza Conflict**

**Thomas Zeitzoff**[1]

# Elite behavior

- Authoritarian governments' response to threat of collective action
- Estimation of conflict intensity in real time
- How elected officials communicate with constituents

FEBRUARY 23, 2017

## For members of 114th Congress, partisan criticism ruled on Facebook


Facebook posts from members of the 114th Congress attracted more attention when they contained disagreement with the opposing party than when they expressed bipartisanship, according to a Pew Research Center analysis of over 100,000 posts.

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Affordable online experiments

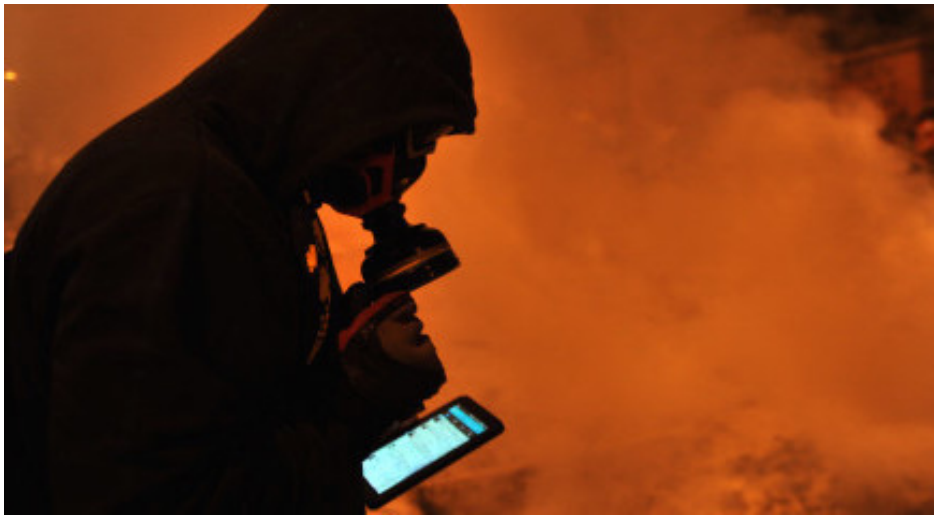

Google Scholar: 23,600 articles mentioned AMT.

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior


MY BODY
NOT YOURS
NASTY AND PROUD



#OccupyGezi

#OccupyWallStreet

#Euromaidan


#Indignados

slacktivism?

# why the revolution will not be tweeted

*When the sit-in movement spread from Greensboro throughout the South, it did not spread indiscriminately. It spread to those cities which had preexisting "movement centers" – a core of dedicated and trained activists ready to turn the "fever" into action.*

*The kind of activism associated with social media isn't like this at all. [. . .] Social networks are effective at increasing participation – by lessening the level of motivation that participation requires.*

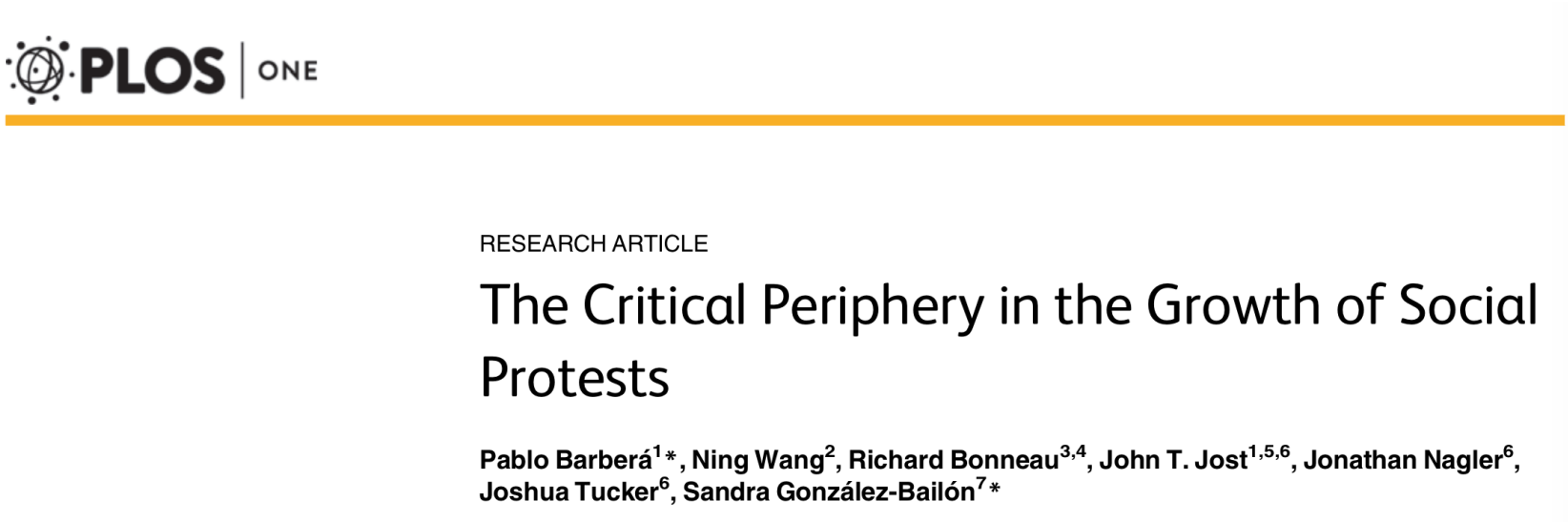
***Gladwell**, Small Change (New Yorker)*

*You can't simply join a revolution any time you want, contribute a comma to a random revolutionary decree, rephrase the guillotine manual, and then slack off for months. Revolutions prize centralization and require fully committed leaders, strict discipline, absolute dedication, and strong relationships.*

*When every node on the network can send a message to all other nodes, confusion is the new default equilibrium.*

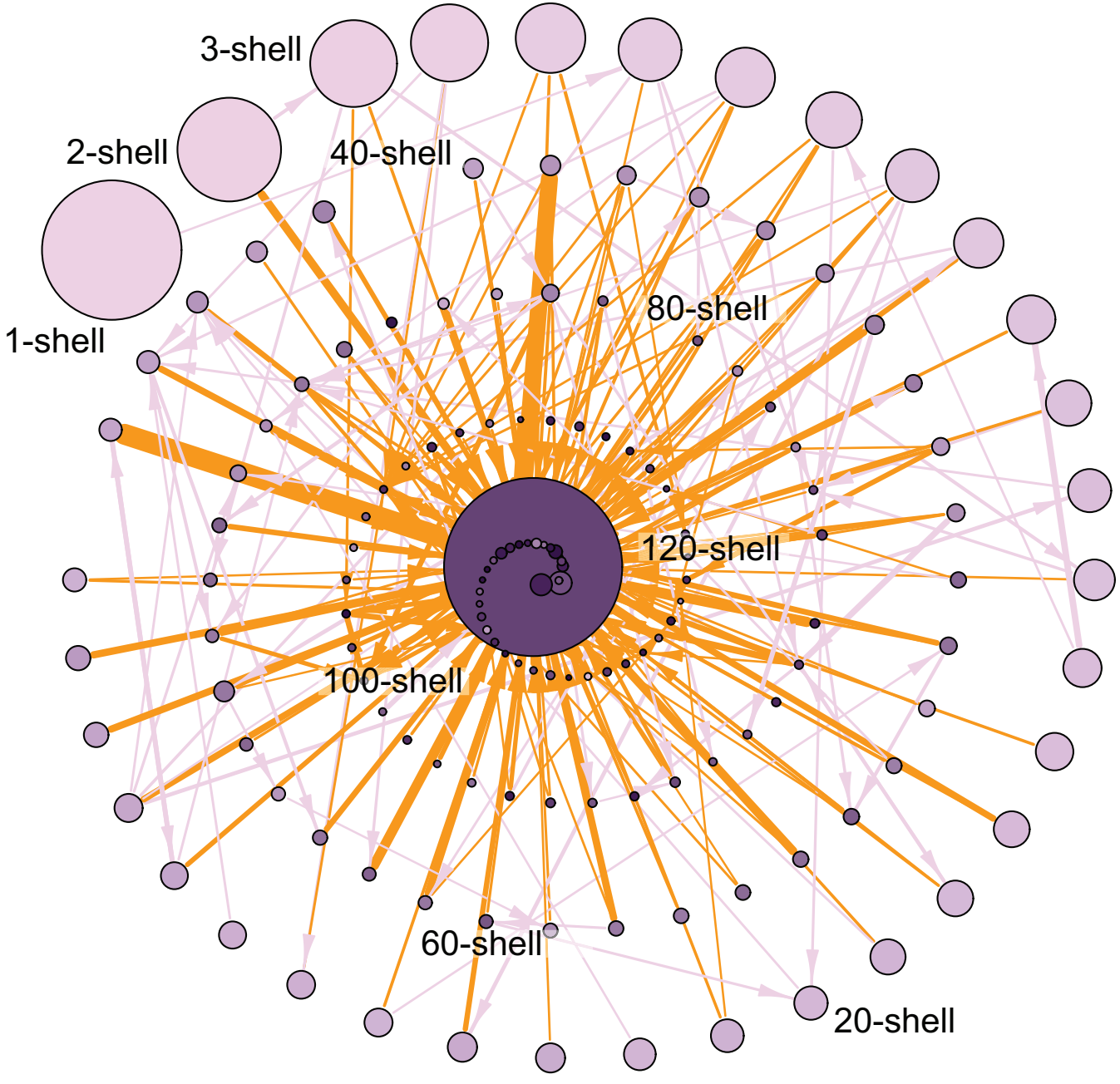
***Morozov**, The Net Delusion: The Dark Side of Internet Freedom*

# the critical periphery




- Structure of online protest networks:
  1. Core: committed minority of resourceful protesters
  2. Periphery: majority of less motivated individuals
- Our argument: key role of peripheral participants
  1. Increase reach of protest messages (positional effect)
  2. Large contribution to overall activity (size effect)

# k-core decomposition of #OccupyGezi network




periphery

core

activity
(no. of tweets)

in Taksim

max 18%

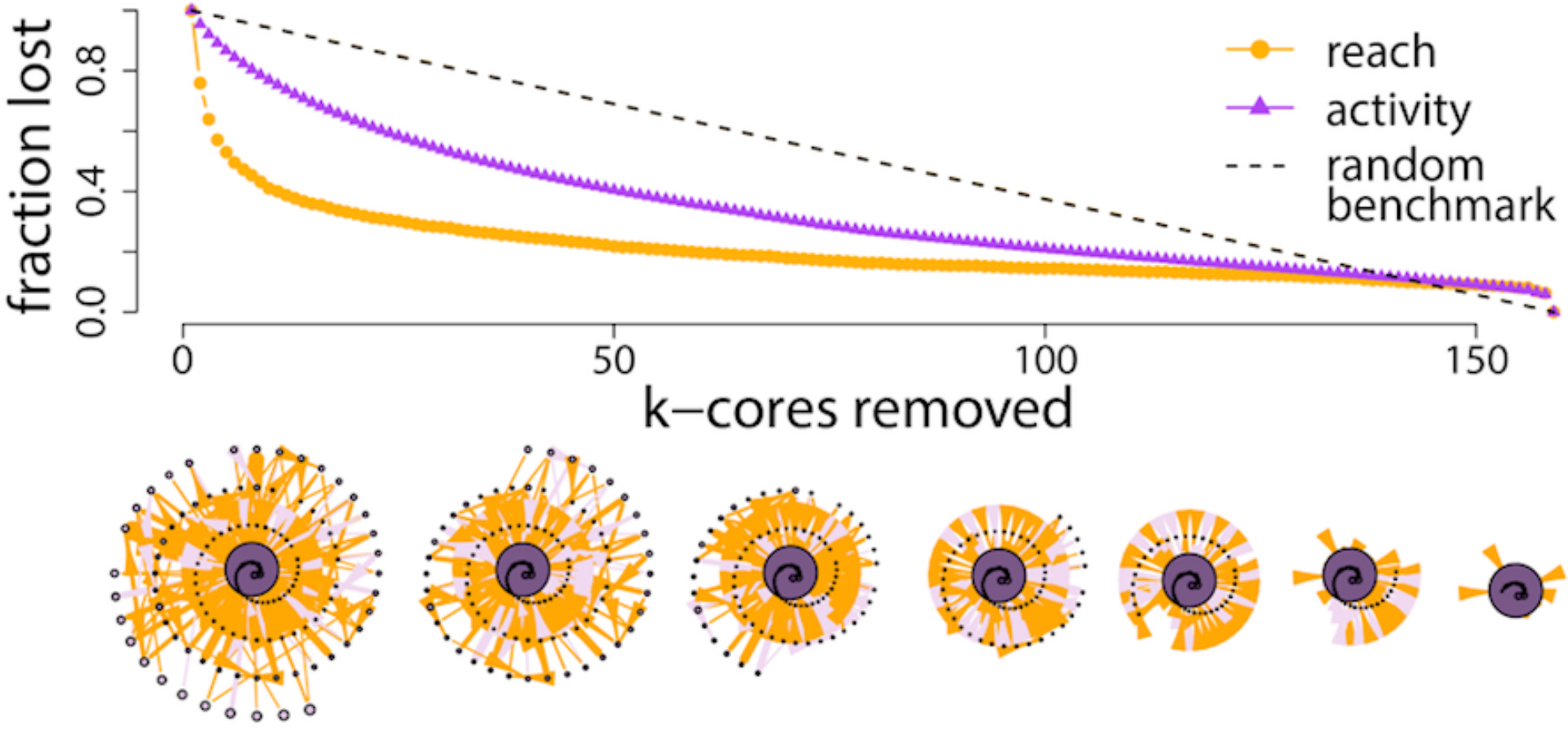
min .25%

RTs

periphery to core

periphery to periphery

# Relative importance of core and periphery




reach: aggregate size of participants' audience

activity: total number of protest messages published (not only RTs)

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

**Barack Obama**
@BarackObama

Follow

# Four more years.

RETWEETS
**756,411**

FAVORITES
**288,867**


11:16 PM - 6 Nov 2012

Post Politics

# By the end of the 2012 campaign, every Mitt Romney tweet had to be approved by 22 people

# Political persuasion

Social media as a new campaign tool:

> *"Let me tell you about Twitter. I think that maybe I wouldn't be here if it wasn't for Twitter. [...] Twitter is a wonderful thing for me, because I get the word out... I might not be here talking to you right now as president if I didn't have an honest way of getting the word out."*
>
> ***Donald Trump**, March 16, 2017 (Fox News)*

- Diminished gatekeeping role of journalists
  - Part of a trend towards citizen journalism (Goode, 2009)
- Information is contextualized within social layer
  - Messing and Westwood (2012): social cues can be as important as partisan cues to explain news consumption through social media
- Real-time broadcasting in reaction to events
  - e.g. *dual screening* (Vaccari et al, 2015)
- Micro-targeting
  - Affects how campaigns perceive voters (Hersh, 2015), but unclear if effective in mobilizing or persuading voters

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Social capital

- Social connections are essential in democratic societies, but online interactions do not facilitate creation and strengthening of social capital (Putnam, 2001)
- Online networking sites facilitate and transform how social ties are established

American Politics Research
1–31

DOI: 10.1177/1532673X14557942
apr.sagepub.com
SAGE

**Tweeting Alone? An Analysis of Bridging and Bonding Social Capital in Online Networks**

**Javier Sajuria[1], Jennifer vanHeerde-Hudson[1], David Hudson[1], Niheer Dasandi[1], and Yannis Theocharis[2]**

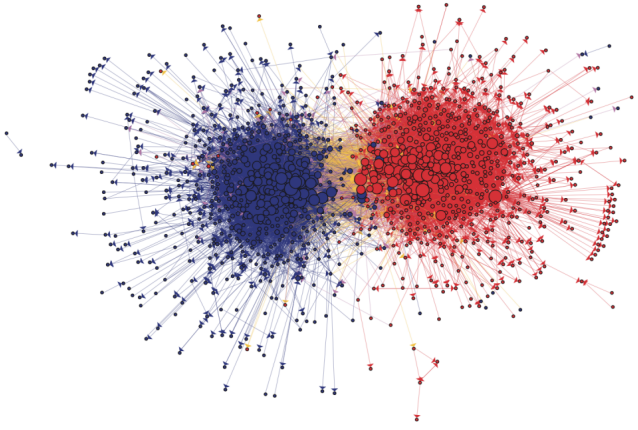
# Computational Social Science

Two different approaches in the growing field of computational social science:

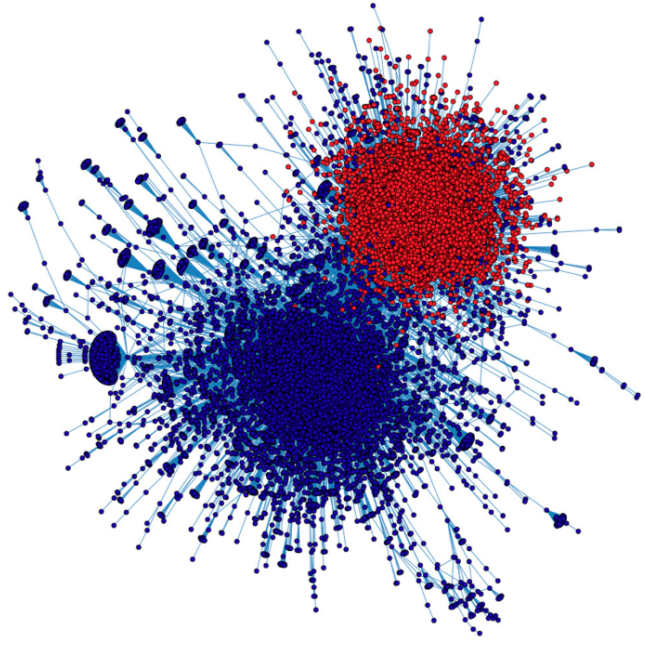
1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

# Social media as echo chambers?

- communities of like-minded individuals (homophily, influence)



Adamic and Glance (2005)

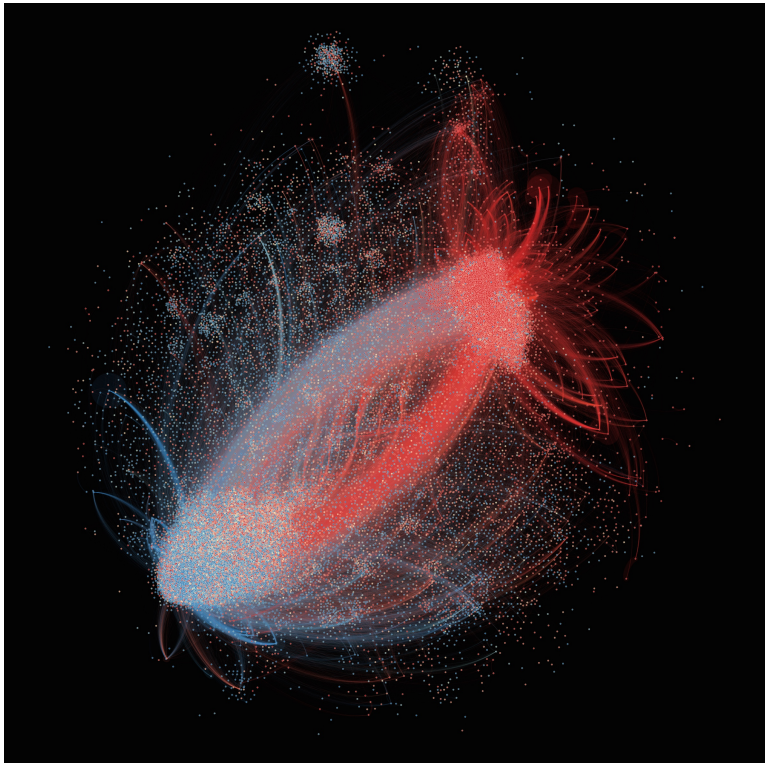


Conover et al (2012)

- ...generates selective exposure to congenial information
- ...reinforced by ranking algorithms – “filter bubble” (Parisier)
- ...increases political polarization (Sunstein, Prior)

# Social media as echo chambers?

2013 SuperBowl

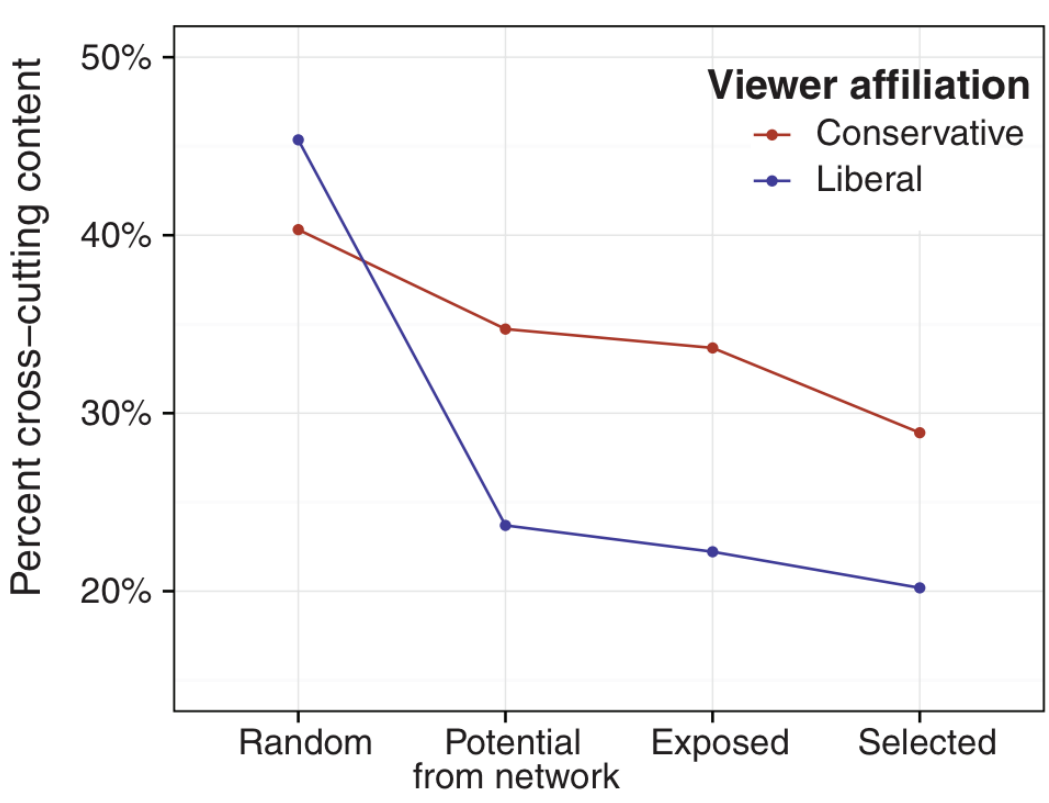


2012 Election

Barberá et al (2015) “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?” *Psychological Science*

# Social media as echo chambers?

**Fig. 3. Cross-cutting content at each stage in the diffusion process.** (**A**) Illustration of how algorithmic ranking and individual choice affect the proportion of ideologically cross-cutting content that individuals encounter. Gray circles illustrate the content present at each stage in the media exposure process. Red circles indicate conservatives, and blue circles indicate liberals. (**B**) Average ideological diversity of content (i) shared by random others (random), (ii) shared by friends (potential from network), (iii) actually appeared in users' News Feeds (exposed), and (iv) users clicked on (selected).




Bakshy, Messing, & Adamic (2015) "Exposure to ideologically diverse news and opinion on Facebook". *Science.*

# Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments
2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

What are the most important challenges when working with Big Data?

# Big data and social science: challenges

1. Big data, big bias?
2. The end of theory?
3. Spam and bots
4. The privacy paradox
5. Generalizing from online to offline behavior
6. Ethical concerns

# 1. Big data, big bias?

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]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadver-

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

*Proprietary algorithms for public data.* Platform-specific sampling problems, for example, the highest-volume source of pub-

The rise of "embedded researc searchers who have special rela with providers that give them ele cess to platform-specific data, al and resources) is creating a divic media research community. Such ers, for example, can see a platfor workings and make accommodat may not be able to reveal their c or the data used to generate their f

Ruths and Pfeffer, 2015, "Social media for large studies of behavior", *Science*

# Big data, big bias?

Sources of bias (Ruths and Pfeffer, 2015; Lazer et al, 2017)

- Population bias
  - Sociodemographic characteristics are correlated with presence on social media
- Self-selection within samples
  - Partisans more likely to post about politics (Barberá & Rivero, 2014)
- Proprietary algorithms for public data
  - Twitter API does not always return 100% of publicly available tweets (Morstatter et al, 2014)
- Human behavior and online platform design
  - e.g. *Google Flu* (Lazer et al, 2014)

# 1. Big data, big bias?

## 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)

**Issues in evaluating data from social media. Large-scale social media studies of human behavior should i address issues listed and discussed herein (further discussion in supplementary materials).**

Ruths and Pfeffer, 2015, "Social media for large studies of behavior", *Science*

# 2. The end of theory?

*Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.*

***Chris Anderson**, Wired, June 2008*

*Correlations are a way of catching a scientist's attention, but the models and mechanisms that explain them are how we make the predictions that not only advance science, but generate practical applications.*

***John Timmer**, Ars Technica, June 2008*

(Big) social media data as a complement - not a substitute - for theoretical work and careful causal inference.

# 3. Spam and bots


*"Follow your coordinators. We need to start tweeting, all at the same time, using the hashtag #ItsTimeForMexico. . . and don't forget to retweet tweets from the candidate's account..."*

***Unidentified PRI campaign manager***

*minutes before the May 8, 2012 Mexican Presidential debate*

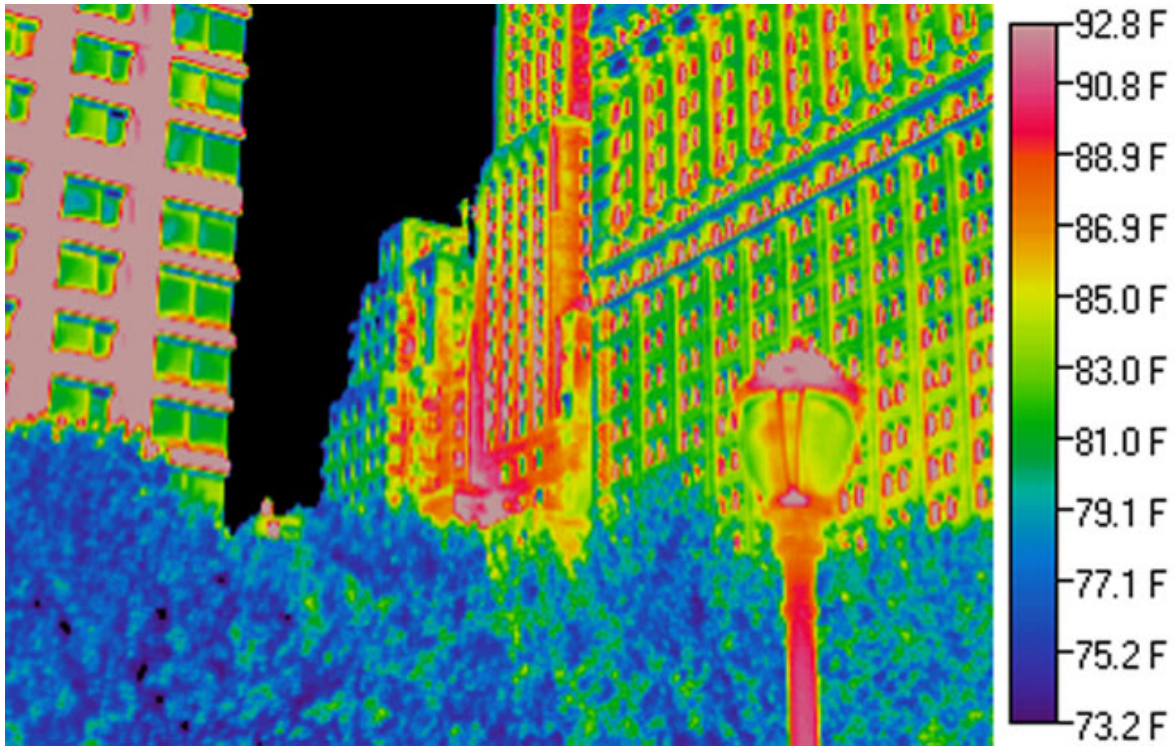
# 3. Spam and bots


Ferrara et al, 2016, *Communications of the ACM*

# 4. The privacy paradox

*Online data present a paradox in the protection of privacy: Data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms of providing the demographic background information needed by social scientists.*

***Golder & Macy**, Digital footprints, 2014*



# 5. Generalizing from online to offline behavior

What makes online behavior different:

- Platform affordances may distort behavior
- Tools extend innate capacities (e.g. Dunbar's number)
- Anonymity encourages vitriol

# 6. Ethical concerns

## 1. Shifting notion of *informed consent*

# Experimental evidence of massive-scale emotional contagion through social networks

**Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]**

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

**Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs**

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are

# 6. Ethical concerns

1. Shifting notion of *informed consent*
2. Most personal data can be de-anonymized

Ethics and Information Technology
December 2010, Volume 12, Issue 4, pp 313–325

## “But the data is already public”: on the ethics of research in Facebook

Authors | Authors and affiliations

Michael Zimmer

Article

**First Online:** 04 June 2010

**DOI**: 10.1007/s10676-010-9227-5

**Cite this article as:**
Zimmer, M. Ethics Inf Technol (2010) 12: 313. doi:10.1007/s10676-010-9227-5

# 6. Ethical concerns

1. Shifting notion of *informed consent*
2. Most personal data can be de-anonymized
3. Rise of “embedded researchers”

“Ethical concerns must be weighed against the value of social research with appropriate steps taken to protest individual privacy” (Shah et al, 2015)

# Today

1. Computational social science research: challenges and opportunities
2. Discussion: ethics of Big Data research.
   - Kramer et al 2014 (and “Editorial Expression of Concern”)
   - Tufekci 2014
3. Scraping data from the web

# For next week

1. Submit coding challenge via Blackboard
2. Readings for discussion:
   - Berinsky et al (2012)
   - Wang et al (2015)
   - Jager (2015)
   - Mullinix et al (2015)
3. No background readings