RECSM Summer School:
Social Media and Big Data Research

Pablo Barberá

School of International Relations
University of Southern California
`pablobarbera.com`

Networked Democracy Lab
`www.netdem.org`

Course website:

github.com/pablobarbera/big-data-upf

George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares

George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares

Bon Alimagno
@karma_thief

Follow

I need a hug. I have never been so traumatized by a television show. #gameofthrones

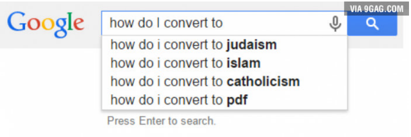
Reply · Retweet · Favorite · More

RETWEETS 356 · FAVORITES 110

10:06 PM - 2 Jun 2013

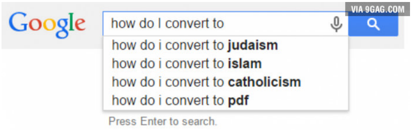

George Takei
March 28 at 10:10pm ·
Who's with me.
HOW I PLAN ON SPENDING
MY FRIDAY NIGHT...
Like · Comment · Share
408,735 people like this.
66,990 shares


Bon Alimagno
@karma_thief
Follow
I need a hug. I have never been so traumatized by a television show. #gameofthrones
Reply Retweet Favorite More
RETWEETS
356
FAVORITES
110
10:06 PM - 2 Jun 2013


VIA 9GAG.COM
Google
how do I convert to
how do i convert to judaism
how do i convert to islam
how do i convert to catholicism
how do i convert to pdf
Press Enter to search.


George Takei
March 28 at 10:10pm ·
Who's with me.
HOW I PLAN ON SPENDING
MY FRIDAY NIGHT...
Like · Comment · Share
408,735 people like this.
66,990 shares


Bon Alimagno
@karma_thief
Follow
I need a hug. I have never been so traumatized by a television show. #gameofthrones
Reply Retweet Favorite More
RETWEETS 356
FAVORITES 110
10:06 PM - 2 Jun 2013


VIA 9GAG.COM
Google
how do I convert to
how do i convert to **judaism**
how do i convert to **islam**
how do i convert to **catholicism**
how do i convert to **pdf**
Press Enter to search.


Justin Bieber
@justinbieber
Follow
I make music. I love music.
Reply Retweet Favorite More
RETWEETS 54,213
FAVORITES 59,205
10:09 PM - 7 Apr 2014

dustin curtis
@dcurtis

Follow

"At any moment, Justin Bieber uses 3% of our infrastructure. Racks of servers are dedicated to him. - A guy who works at Twitter

RETWEETS	FAVORITES
1,528	267

8:56 PM - 6 Sep 2010

Dmitry Medvedev
@MedvedevRussiaE

Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply Retweet Favorite More

RETWEETS	FAVORITES
144	57

10:39 AM - 21 Mar 2014

Dmitry Medvedev
@MedvedevRussiaE

Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply Retweet Favorite More

RETWEETS 144 FAVORITES 57

10:39 AM - 21 Mar 2014

The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

Like · Comment · Share 57

262 people like this. Top Comments

Dmitry Medvedev
@MedvedevRussiaE

Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply · Retweet · Favorite · More

RETWEETS 144 · FAVORITES 57

10:39 AM - 21 Mar 2014

The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

Like · Comment · Share · 57

262 people like this. · Top Comments

Elizabeth Warren shared a link.
January 16

I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.

Warren: This is the moment to back on economy
www.msnbc.com

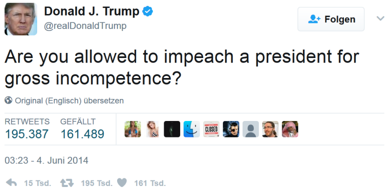
President Obama faces one huge problem with his effort to improve the economy: an opposition party

Like · Comment · Share · 15,483 · 720 · 1,041

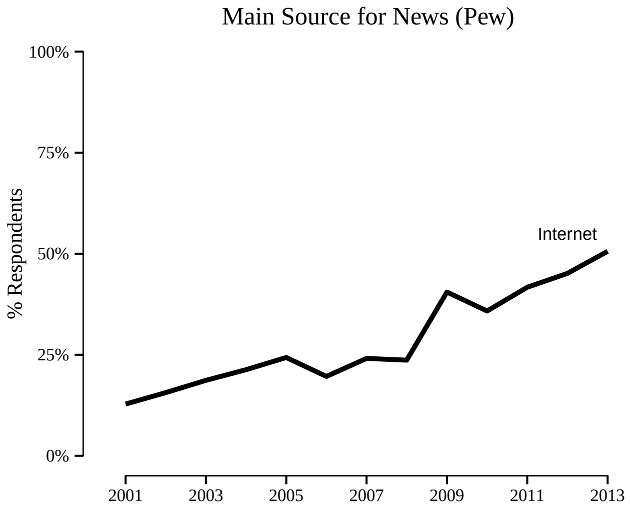

Dmitry Medvedev
@MedvedevRussiaE
Follow
The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government
Reply Retweet Favorite More
RETWEETS 144
FAVORITES 57
10:39 AM - 21 Mar 2014


The New York Times
April 2
"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I
Like · Comment · Share
57
262 people like this.
Top Comments


Elizabeth Warren shared a link.
January 16
I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.
Warren: This is the moment to back on economy
www.msnbc.com
President Obama faces one huge problem with his effort to improve the economy: an opposition party
Like · Comment · Share
15,483
720
1,041

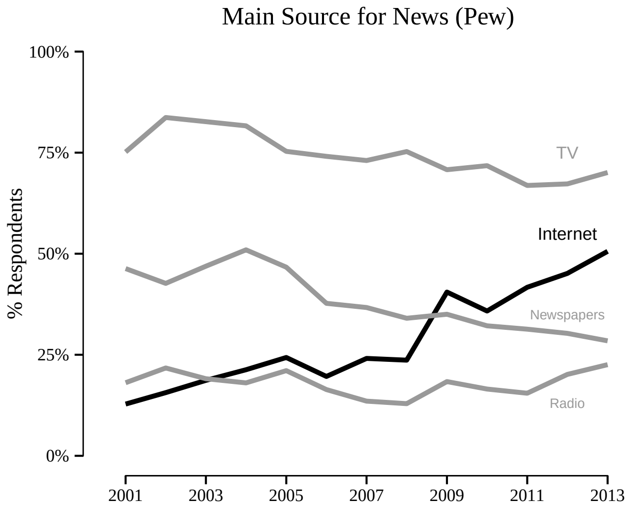

Donald J. Trump
@realDonaldTrump
Folgen
Are you allowed to impeach a president for gross incompetence?
Original (Englisch) übersetzen
RETWEETS 195.387
GEFÄLLT 161.489
03:23 - 4. Juni 2014
15 Tsd.
195 Tsd.
161 Tsd.

Sources of Political Information



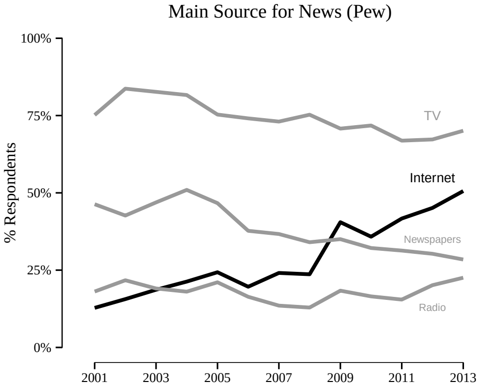
Data: Pew Research Center. Respondents were allowed to name up to two sources.

Sources of Political Information



Data: Pew Research Center. Respondents were allowed to name up to two sources.

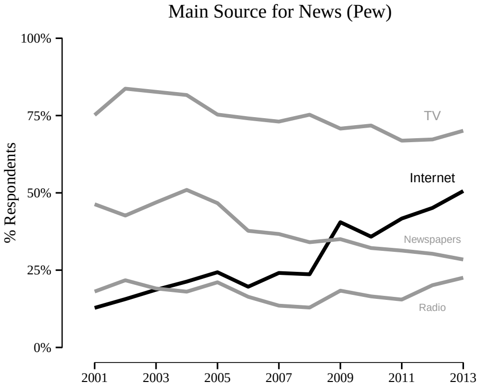
Sources of Political Information



Data: Pew Research Center. Respondents were allowed to name up to two sources.

- 62% of Americans gets news on social media (Pew)

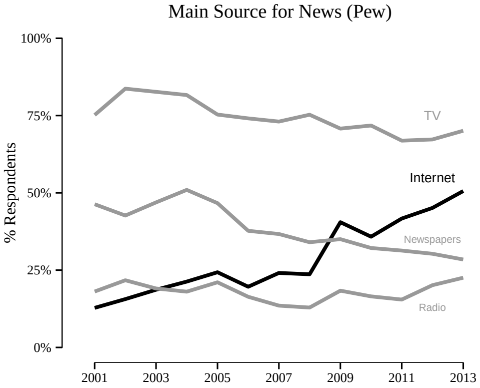
Sources of Political Information



Data: Pew Research Center. Respondents were allowed to name up to two sources.

- 62% of Americans gets news on social media (Pew)
- 27% of online EU citizens use social media to get news on national political matters (Eurobarometer, Fall 2012)

Sources of Political Information



Data: Pew Research Center. Respondents were allowed to name up to two sources.

- 62% of Americans gets news on social media (Pew)
- 27% of online EU citizens use social media to get news on national political matters (Eurobarometer, Fall 2012)
- Social media: top source of news for U.S. young adults (Pew)

$\longrightarrow$

Shift in communication patterns

Shift in communication patterns

Digital footprints of human behavior

This course

1. Research opportunities and challenges
   - New and old social science questions
   - Limits of Big Data
2. Data collection
   - Webscraping
   - Twitter, Facebook
3. Data analysis
   - Large-scale network and text datasets

Hello!

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:

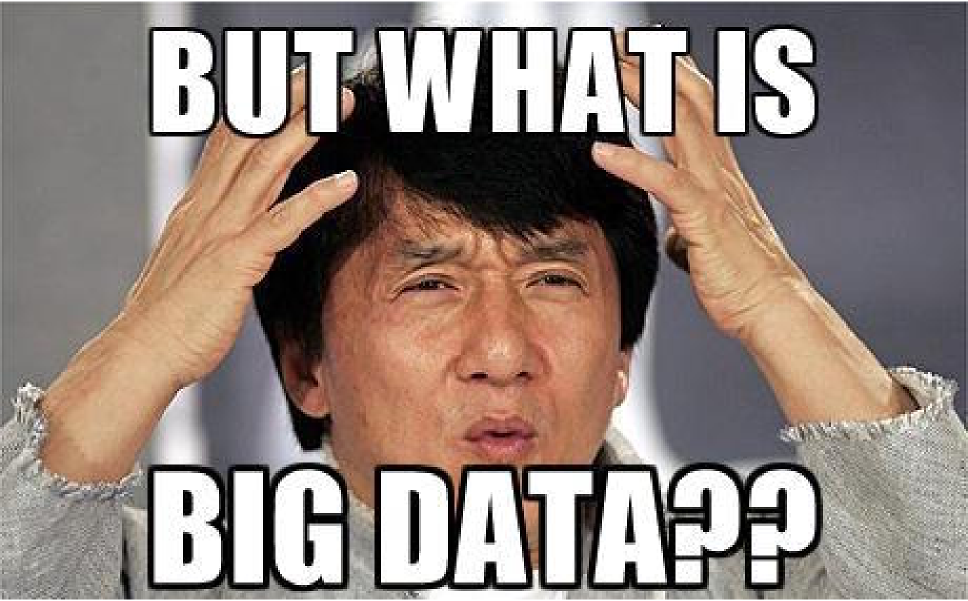
About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:
  - `pbarbera@usc.edu`

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Professor at University of Southern California
- PhD in Politics, New York University (2015)
- Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior, corruption and accountability
  - Social network analysis, Bayesian statistics, text as data methods
  - Author of R packages to analyze data from social media
- Contact:
  - `pbarbera@usc.edu`
  - `www.pablobarbera.com`

Big Data: Opportunities and Challenges


BUT WHAT IS
BIG DATA??

The Three V's of Big Data

Dumbill (2012), Monroe (2013):

1. Volume: 6 billion mobile phones, 1+ billion Facebook users, 500+ million tweets per day...

The Three V's of Big Data

Dumbill (2012), Monroe (2013):

1. Volume: 6 billion mobile phones, 1+ billion Facebook users, 500+ million tweets per day...
2. Velocity: personal, spatial and temporal granularity.

The Three V's of Big Data

Dumbill (2012), Monroe (2013):

1. Volume: 6 billion mobile phones, 1+ billion Facebook users, 500+ million tweets per day...
2. Velocity: personal, spatial and temporal granularity.
3. Variability: images, networks, long and short text, geographic coordinates, streaming...

The Three V's of Big Data

Dumbill (2012), Monroe (2013):

1. Volume: 6 billion mobile phones, 1+ billion Facebook users, 500+ million tweets per day...
2. Velocity: personal, spatial and temporal granularity.
3. Variability: images, networks, long and short text, geographic coordinates, streaming...

Big data: data that are so large, complex, and/or variable that the tools required to understand them must first be invented.

Computational Social Science

"We have life in the network. We check our emails regularly, make mobile phone calls from almost any location ... make purchases with credit cards ... [and] maintain friendships through online social networks ... These transactions leave 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** et al (2009) Science*

Two different approaches to the study of big data and social sciences:

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinion, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Behavior, opinions, and latent traits

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

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”

Behavior, opinions, and latent traits

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

$\rightarrow$ 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, ...

$\rightarrow$ Non-intrusive measurement of behavior and public opinion

$\rightarrow$ Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, ...

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, ...

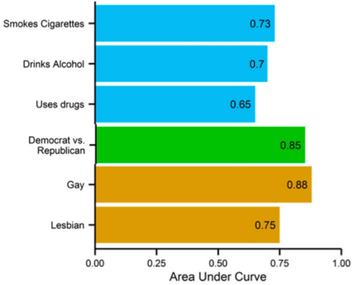


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

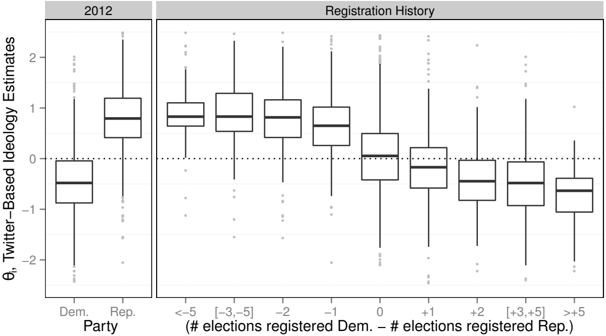
Kosinki et al, 2013, “Private traits and attributes are predictable from digital records of human behavior”, *PNAS* (also personality, PNAS 2015)

Behavior, opinions, and latent traits

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

$\rightarrow$ Non-intrusive measurement of behavior and public opinion

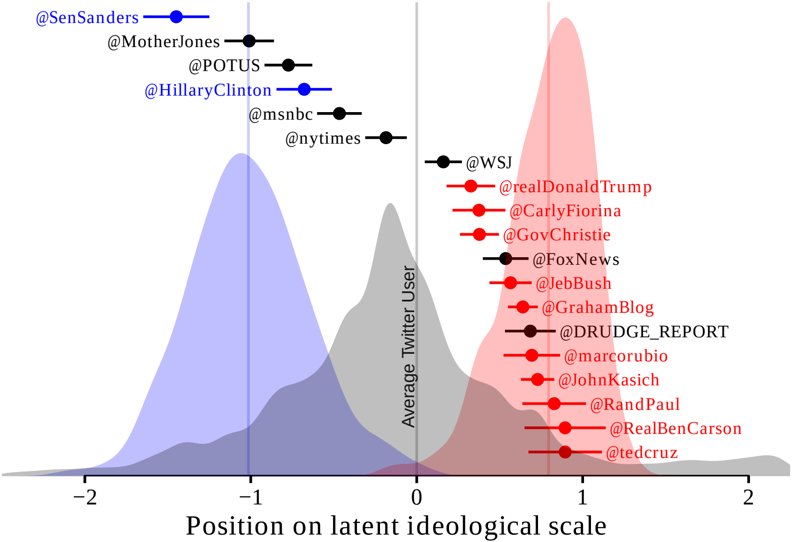
$\rightarrow$ Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, ...



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

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

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).

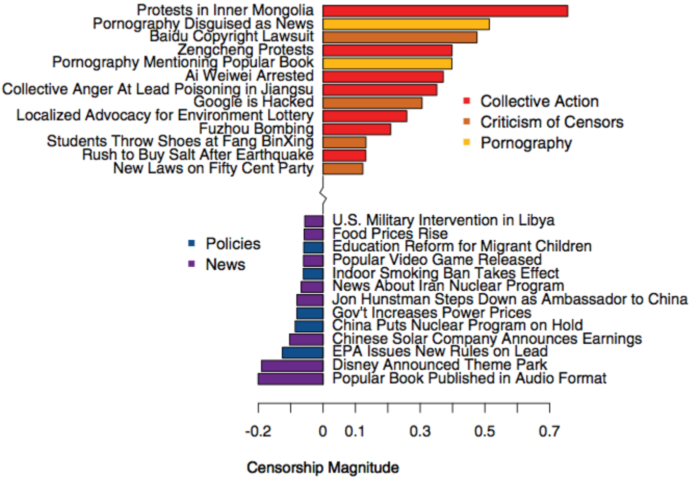
Jones et al, 2013, "Inferring Tie Strength from Online Directed Behavior", *PLOS One*

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

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

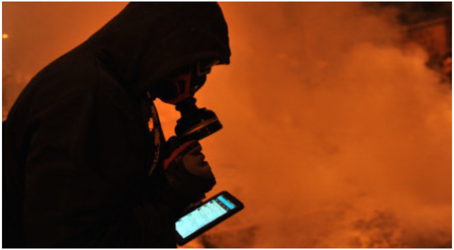
For members of 114th Congress, partisan criticism ruled on Facebook

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.

Two different approaches to the study of big data and social sciences:

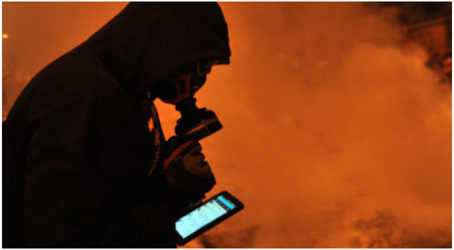
1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization


MY BODY
NOT YOURS
AND
PROUD



#OccupyGezi

#Euromaidan



#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)*

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.

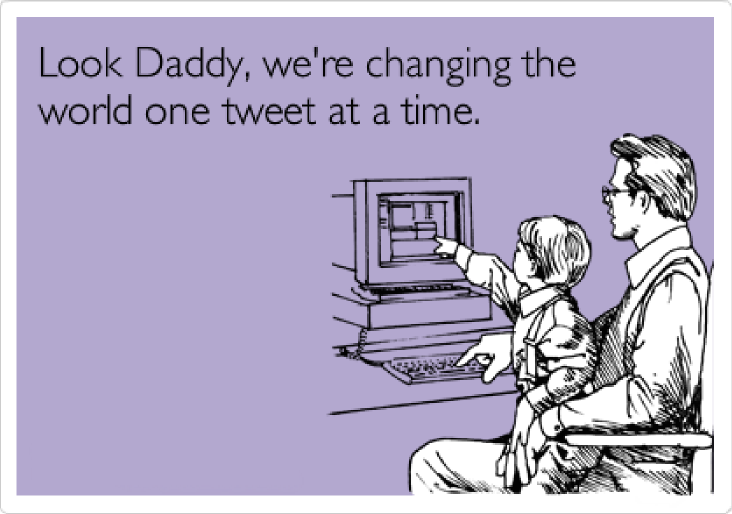
***Gladwell**, Small Change (New Yorker)*

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*

parody or reality?



the critical periphery

PLOS | ONE

RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- Structure of online protest networks:

the critical periphery

PLOS | ONE

RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- Structure of online protest networks:
  1. Core: committed minority of resourceful protesters

the critical periphery

PLOS | ONE

RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- Structure of online protest networks:
  1. Core: committed minority of resourceful protesters
  2. Periphery: majority of less motivated individuals

the critical periphery

PLOS | ONE

RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- 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

the critical periphery

PLOS | ONE

RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- 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)

the critical periphery

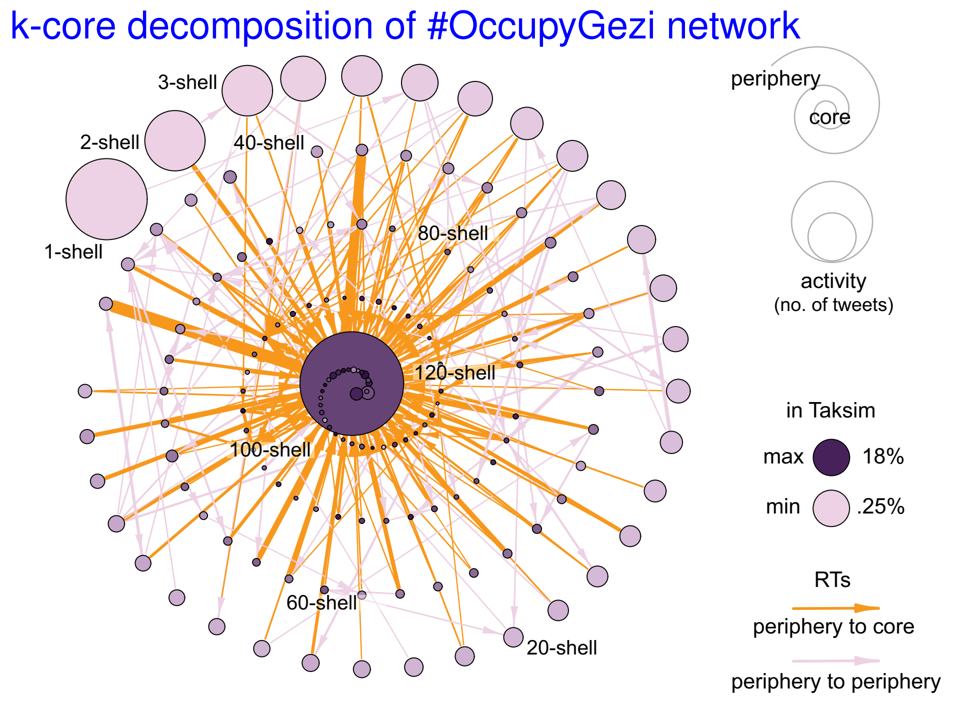
PLOS | ONE

RESEARCH ARTICLE

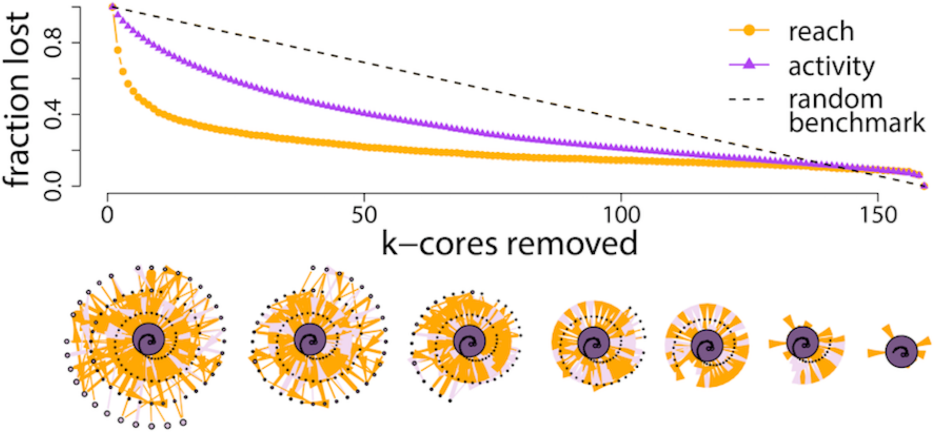
The Critical Periphery in the Growth of Social Protests

Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- 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
3-shell
2-shell
40-shell
1-shell
80-shell
120-shell
100-shell
60-shell
20-shell
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)

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Barack Obama
@BarackObama

Follow

Four more years.

RETWEETS 756,411 FAVORITES 288,867

11:16 PM - 6 Nov 2012

Sections ≡ The Washington Post Search Sign In

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)*

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

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)

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

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

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

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)

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

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

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
    - Behavior, opinions, and latent traits
    - Interpersonal networks
    - Elite behavior
2. How big data and social media affect social behavior
    - Mass protests
    - Political persuasion
    - Social capital
    - Political polarization

Social capital

- Social connections are essential in democratic societies, but online interactions do not facilitate creation and strengthening of social capital (Putnam, 2001)

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

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

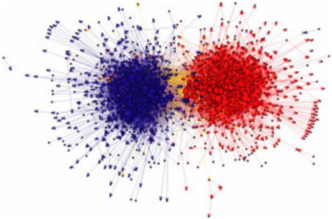
Javier Sajuria[1], Jennifer vanHeerde-Hudson[1], David Hudson[1], Niheer Dasandi[1], and Yannis Theocharis[2]

Two different approaches to the study of big data and social sciences:

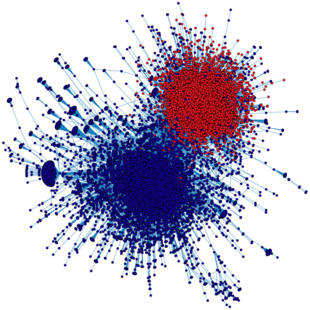
1. Big data as a new source of information
    - Behavior, opinions, and latent traits
    - Interpersonal networks
    - Elite behavior
2. How big data and social media affect social behavior
    - Mass protests
    - Political persuasion
    - Social capital
    - Political polarization

Social media as echo chambers?

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



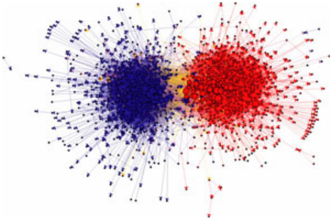
Adamic and Glance (2005)



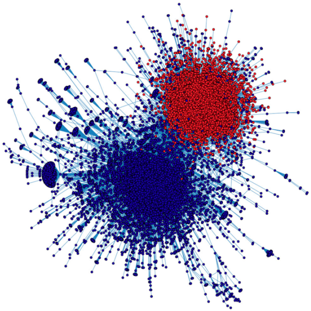
Conover et al (2012)

Social media as echo chambers?

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



Adamic and Glance (2005)

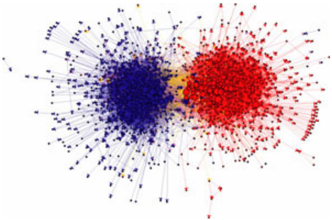


Conover et al (2012)

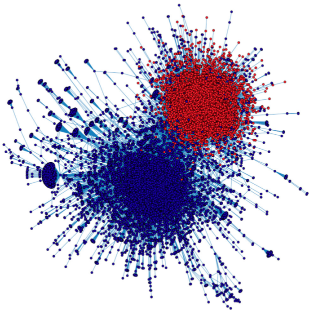
- ...generates selective exposure to congenial information
- ...reinforced by ranking algorithms – “filter bubble” (Parisier)

Social media as echo chambers?

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



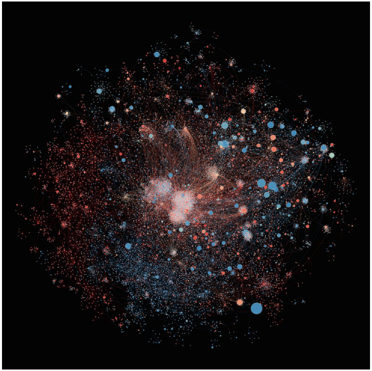
Adamic and Glance (2005)



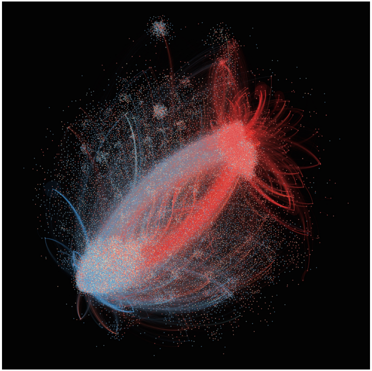
Conover et al (2012)

- ...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

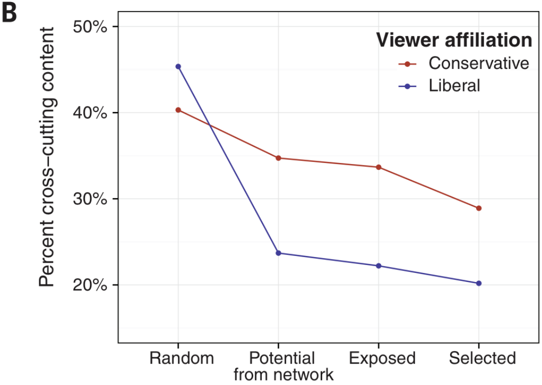


2012 Election

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.*

Two different approaches to the study of big data and social sciences:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. How big data and social media affect social behavior
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Big data and social science: challenges

1. Big data, big bias?
2. The end of theory?
3. Spam and bots
4. Ethical concerns

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. 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

4. 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

4. Ethical concerns

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

4. 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)

Collecting Big Data: First Steps

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages*, able to interact with databases, machine learning libraries, etc.

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages*, able to interact with databases, machine learning libraries, etc.
- Command-line interface favors reproducibility

Why we're using R

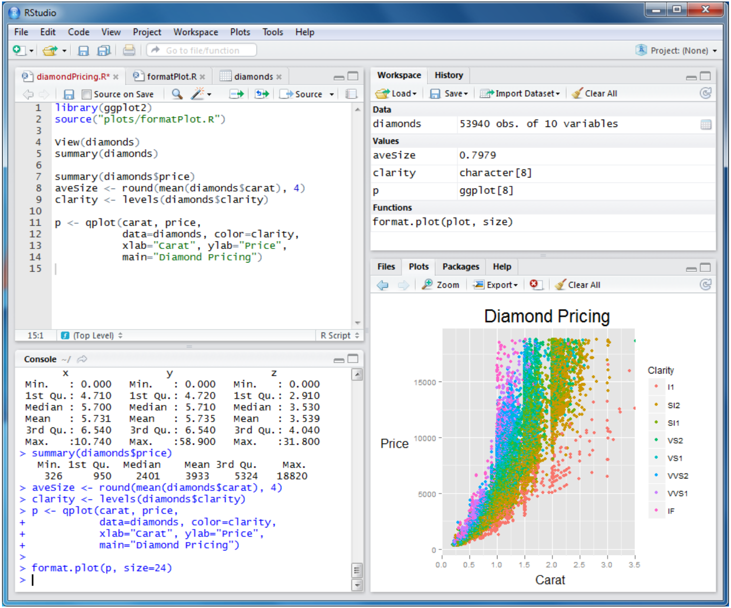
- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages*, able to interact with databases, machine learning libraries, etc.
- Command-line interface favors reproducibility
- Great for data visualization

Why we're using R

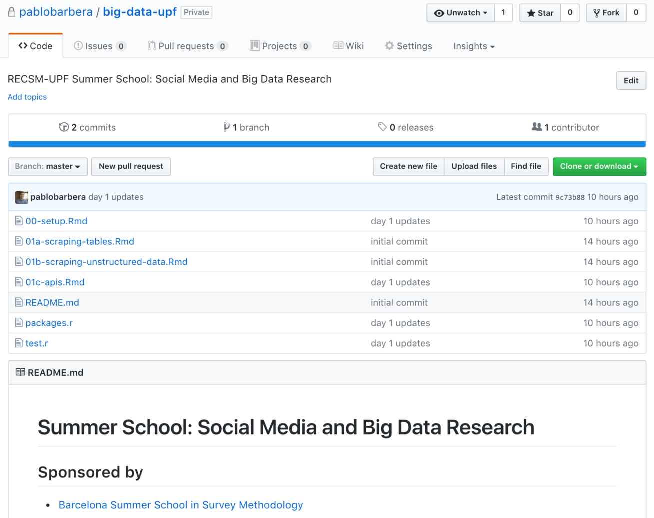
- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages*, able to interact with databases, machine learning libraries, etc.
- Command-line interface favors reproducibility
- Great for data visualization

R is also a full programming language; once you understand how to use it, you can learn other languages too.

RStudio Server



Course website



github.com/pablobarbera/big-data-upf

RStudio Server URL: `bigdata.pablobarbera.com`
Then enter user = userXX and password = passwordXX
where XX corresponds to the following number:

Aglamaz 03	Ansemil 04	Aznar 05	Belousova 06
Castro 07	Chan 08	Costas-Perez 09	Curto-Grau 10
Del Real 11	Djourelova 12	Ellingsen 13	Fabregas 14
Fonseca 15	Furlan 16	Grond 17	Hosseini 18
Huidobro 19	Ismailov 20	Macassi 21	Majo-Vazquez 22
Martini 23	Mavletova 24	Moreno 25	Muis 26
Nesena 27	Pinzon 28	Plaza 29	Rasic 30
Rodriguez 31	Rubal 32	Schoell 33	Serani 34
Staessens 35	Stein 36	Szewach 37	Tanovic 38
Trokhova 39	Vranceanu 40	Zhou 41	

RECSM Summer School: Social Media and Big Data Research

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/big-data-upf