ECPR Methods Summer School: Big Data Analysis in the Social Sciences

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/ECPR-SC104

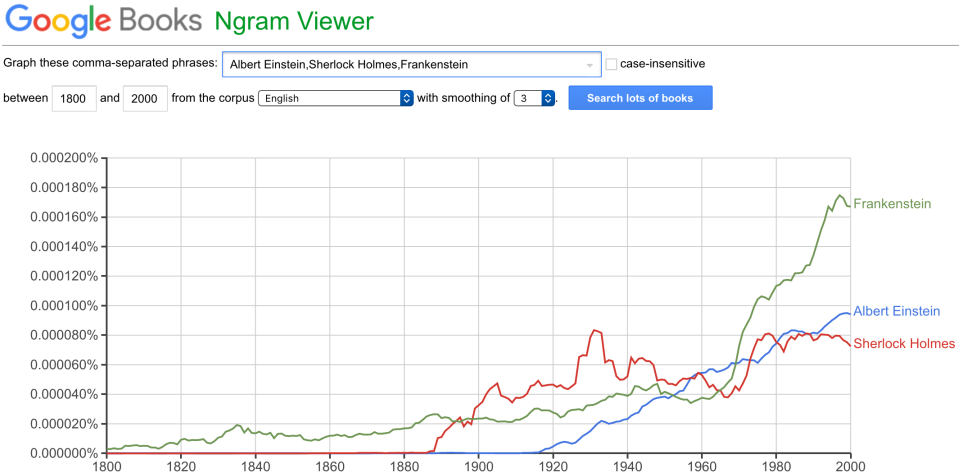
Data is everywhere


243.26
TOMACCO
JUICE


Rx

NBC/Instagram

XING
Gefällt mir
Instagram
twitter
LinkedIn
facebook
Pinterest
flickr
skype
YouTube


Google Books Ngram Viewer
Graph these comma-separated phrases:
Albert Einstein,Sherlock Holmes,Frankenstein
case-insensitive
between
1800
and
2000
from the corpus
English
with smoothing of
3
Search lots of books
0.000200%
0.000180%
0.000160%
0.000140%
0.000120%
0.000100%
0.000080%
0.000060%
0.000040%
0.000020%
0.000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Frankenstein
Albert Einstein
Sherlock Holmes


Strongly agree
Agree
Disagree
disagree

The Data revolution in election campaigns

CNN Tech » Gadgets | Cyber Security | Innovation Nation   Live TV   U.S. Edition +   menu

How Obama's data crunchers helped him win

By Michael Scherer
Updated 11:45 AM ET, Thu November 8, 2012

President Obama's campaign manager hired an analytics department five times as large as that of the 2008 operation.

Top stories

Top US commander warns Russia, Syria

Is NBC's Olympics coverage really that bad?

The Data revolution in election campaigns

Data Analyst

APPLY FOR THIS JOB

BROOKLYN, NY ANALYTICS FULL-TIME

We are looking for Data Analysts, at both the junior and senior levels, to join our team at our Brooklyn, NY headquarters. The Analyst will play a pivotal role in developing data-driven strategies for key primary and battleground states. They will be responsible for designing and building tools to guide strategies at all levels of the campaign. By utilizing their statistical expertise, our Analysts will dissect large datasets, synthesize results and present findings to team leaders.

2016

Trump's secret data reversal

Having once dismissed the importance of campaign tech, the mogul is now rushing to catch up with Clinton.

By **KENNETH P. VOGEL** and **DARREN SAMUELSOHN** | 06/28/16 05:22 AM EDT

Donald Trump has dismissed political data operations as “overrated,” but his campaign is now bolstering its online fundraising and digital outreach by turning to GOP tech specialists who previously tried to stop him from winning the party’s nomination.

Data Journalism

theguardian

election 2016 | US | world | opinion | sports | soccer | tech | arts | lifestyle | fashion | business | travel | environment | all sections

home

datablog

Support for leaving EU likely to be overstated in polls, analysis suggests

One of the only major pollsters to correctly call the UK general election in 2015 says that problems still remain in the way polls are being done

2,489

Will Brexit mean a Brit expat exodus?

1,496

Data Journalism

TheUpshot

THE 2016 RACE

50 Years of Electoral College Maps: How the U.S. Turned Red and Blue

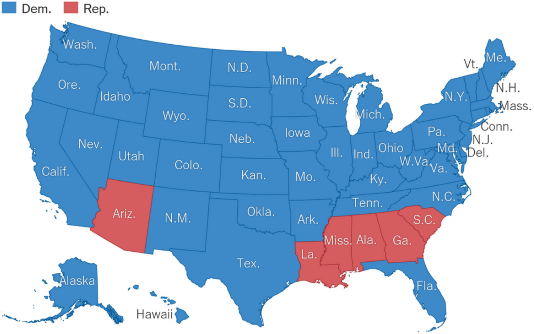
A brief guided tour: Understanding the history of modern American politics means reckoning with the effect of race.

6h ago • By TONI MONKOVIC

The 1964 Election: Johnson Defeats Goldwater

In the popular vote, Lyndon Johnson defeated Barry Goldwater, 61.1 to 38.5.

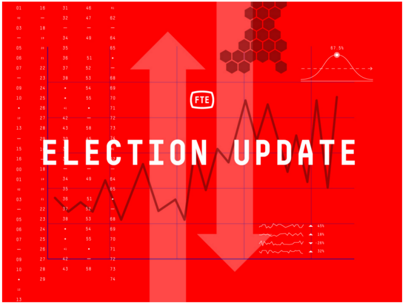
The blue states reflect a total of 486 electoral votes for Lyndon Johnson



Data Journalism

FiveThirtyEight ESPN

Politics | Sports | Science & Health | Economics | Culture | All week: Rio 2016 coverage



2016 ELECTION

National Polls Show The Race Tightening — But State Polls Don't

By Nate Silver

THE LATEST

8:55 AM
Significant Digits For Monday, Aug. 22, 2016

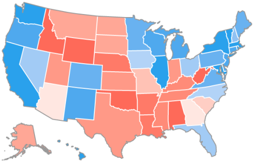
AUG 21
Election Update: National Polls Show The Race Tightening — But State Polls Don't

AUG 19
Winning An Olympic Gold Medal Hasn't Been This Difficult Since 1896

AUG 19
Let Caster Run! We Should Celebrate Semenya's Extraordinary Talent

INTERACTIVES

2016 Election Forecast
UPDATED 4 HOURS AGO



See polls and forecasts

MLB Predictions
UPDATED 15 HOURS AGO

Upcoming games

Nationals def. Orioles	50%
Pirates def. Astros	56%

Non-profit sector

Development data
Datablog

Data without borders: why I want to change the world

Data scientist **Jake Porway** wants to hook up developers with charities and the developing world. Here he explains why

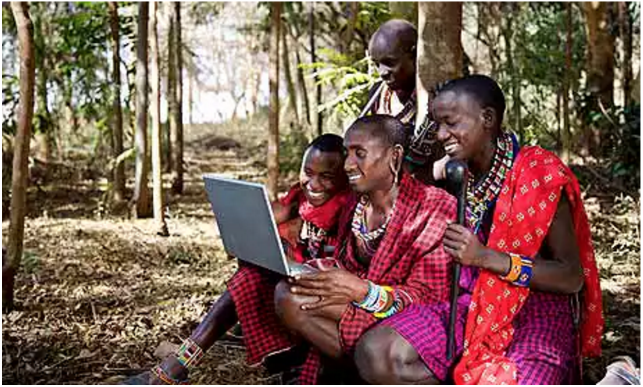
Jake Porway

Thursday 23 June 2011
11.10 EDT

Shares
8

Comments
0

Save for later



Data without borders: Men on the Samburu National Reserve, Kenya, using a laptop Photograph: Scott Stulberg/Corbis

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

How can we analyze *Big Data* to answer social science questions?

Course outline

1. Efficient data analysis in R
   - Writing more efficient code
   - Parallel computing
2. Text classification at scale
   - Supervised machine learning
   - Applications to text
3. Topic discovery in text
   - Dimensionality reduction
   - Topic models
4. Finding communities in massive networks
   - Cluster detection
   - Latent-space modeling
5. Cloud computing
   - SQL for data manipulation
   - Large-scale data processing in the cloud

Hello!

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Prof. at Univ. of Southern California
- Director, Networked Democracy Lab, `netdem.org`
- 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`

Juraj Medzihorsky

- Post-doc at the V-Dem Institute at the University of Gothenburg as of August 2017
- Currently post-doc at CEU
- PhD in political science, CEU (2015)
- Research interests:
  - Mixture models, categorical data analysis, measurement models, Bayesian statistics
  - Elections and assemblies
- Contact:
  - juraj.medzihorsky@gmail.com

Your turn!

1. Name?
2. Affiliation?
3. Research interests?
4. Previous experience with R?
5. Why are you interested in this course?

Course philosophy

How to learn the techniques in this course?

- Lecture approach: not ideal for learning how to code
- You can only learn by doing.

→ We will cover each concept three times during each session

1. Introduction to the topic (20-30 minutes)
2. Guided coding session (30-40 minutes)
3. Coding challenges (30 minutes)

- You're encouraged to continue working on the coding challenges after class. Solutions will be posted the following day.
- Additional questions? We can arrange one-on-one meetings after class

Course logistics

ECTS credits:

- Attendance: 2 credits (pass/fail grade)
- Submission of at least 3 coding challenges: +1 credit
  - Due before beginning of following class via email
  - Only applies to challenge 2 of the day
  - Graded on a 100-point scale
- Submission of class project: +1 credit
  - Due by August 27th
  - Goal: collect and analyze data from the web or social media
  - 10 pages max (including code) in Rmarkdown format
  - Graded on a 100-point scale

If you wish to obtain more than 2 credits, please indicate so in the attendance sheet

Social event

Save the date:

Wednesday Aug. 9th, 6pm

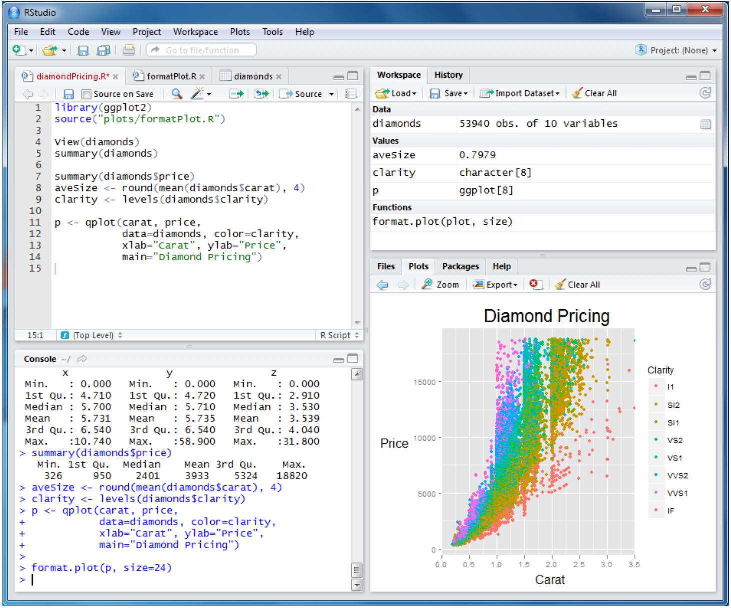
Location TBA

Why we're using R

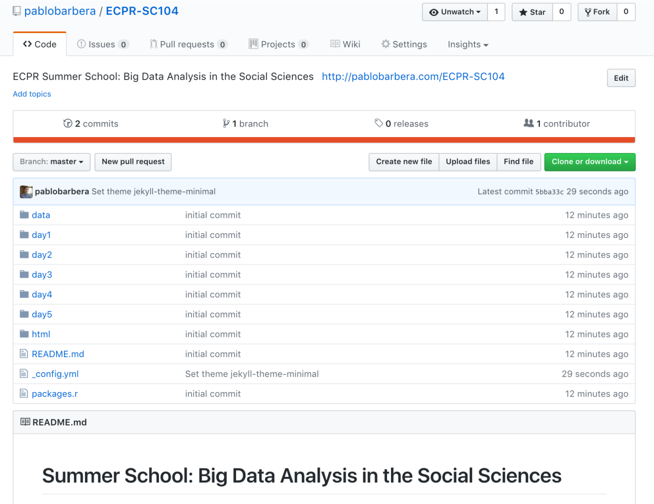
- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It's free and open-source
- Flexible and extensible through *packages* (over 10,000 and counting!)
- Powerful tool to conduct automated text analysis, social network analysis, and data visualization, with packages such as quanteda, igraph or ggplot2.
- Excellent libraries to do machine learning, efficiently read large files, and interact with online databases.
- Command-line interface and scripts favors reproducibility.
- Outstanding documentation and online help resources.

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

RStudio Server

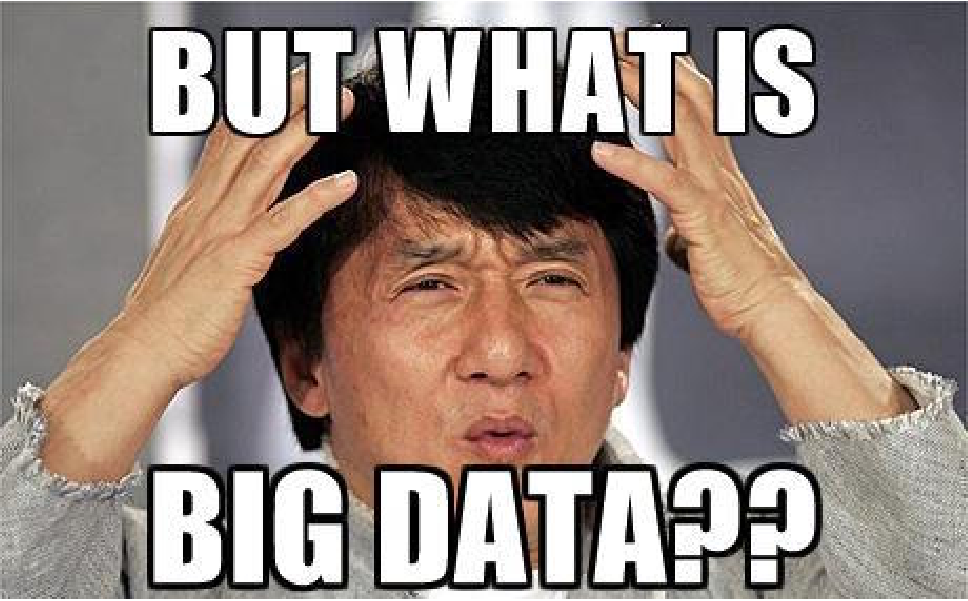


Course website



github.com/pablobarbera/ECPR-SC104

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...
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:

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, ...

→ 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, ...

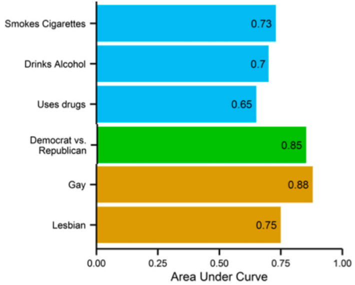


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

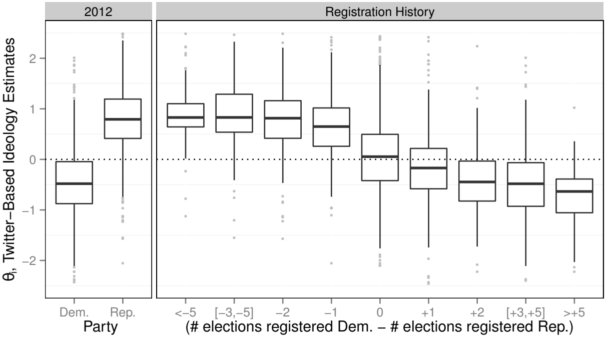
Kosinki et al, 2013, “Private traits and attributes are predictable from digital records of human behavior”, *PNAS* (also personality, PNAS 2015)

Behavior, opinions, and latent traits

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

→ Non-intrusive measurement of behavior and public opinion

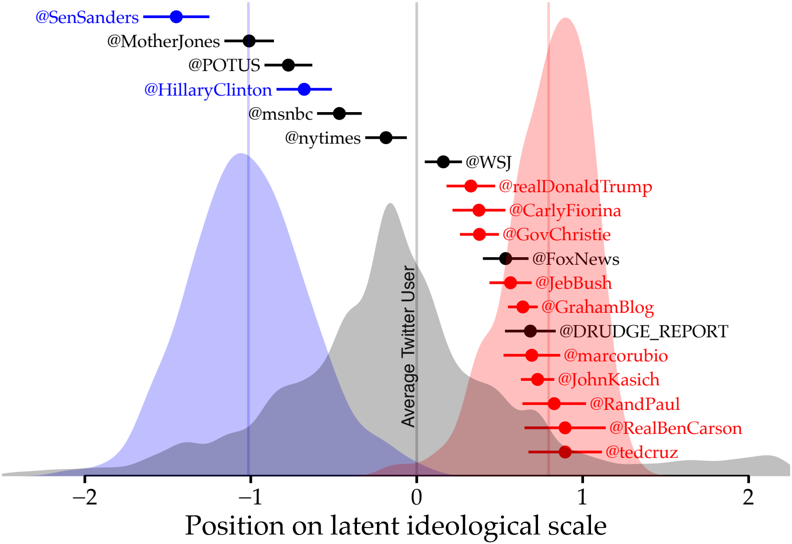
→ Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, ...



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

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).

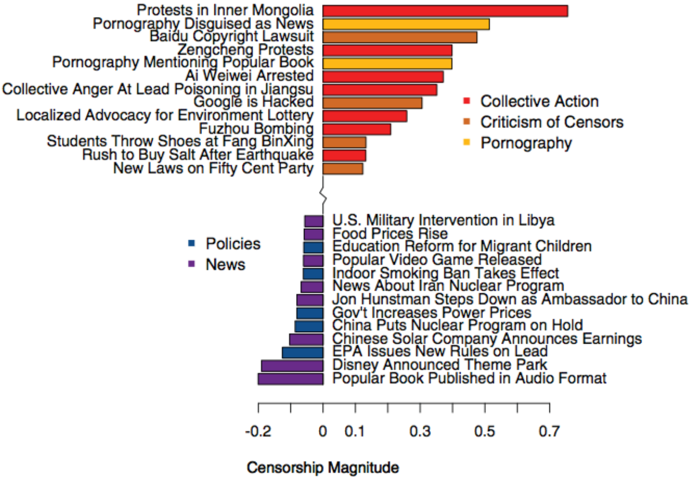
Jones et al, 2013, "Inferring Tie Strength from Online Directed Behavior", *PLOS One*

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

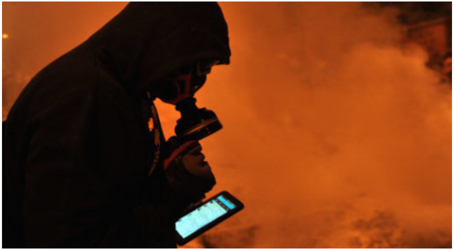
For members of 114th Congress, partisan criticism ruled on Facebook

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
LOVE



#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.

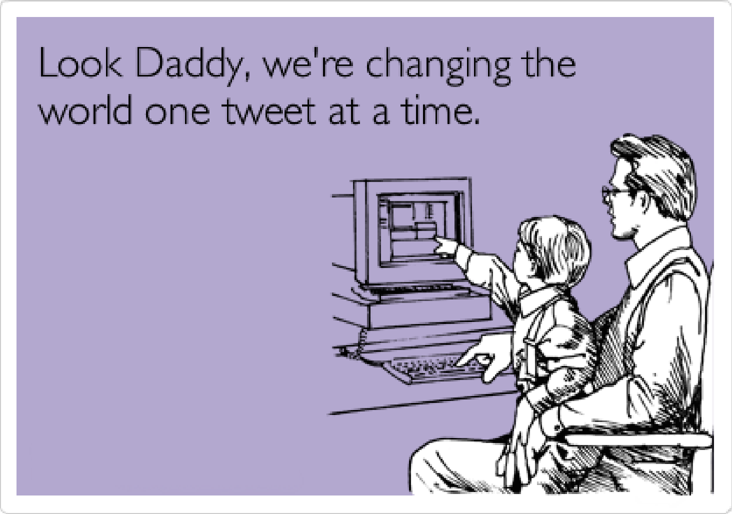
Gladwell, *Small Change (New Yorker)*

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*

parody or reality?



the critical periphery

PLOS | ONE

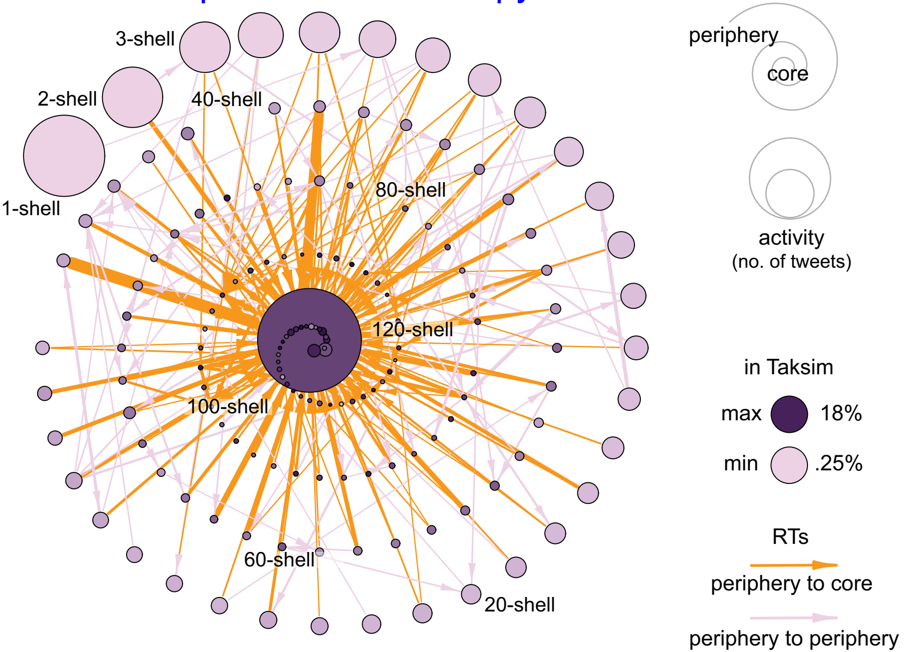
RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

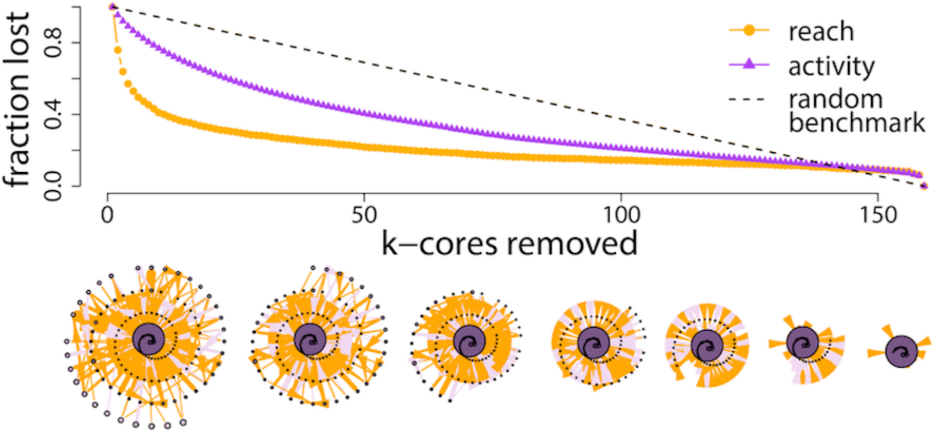
Pablo Barberá[1]*, Ning Wang[2], Richard Bonneau[3,4], John T. Jost[1,5,6], Jonathan Nagler[6], Joshua Tucker[6], Sandra González-Bailón[7]*

- Structure of online protest networks:
  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



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	FAVORITES
756,411	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

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)
- 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

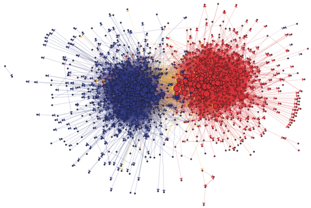
Javier Sajuria[1], Jennifer vanHeerde-Hudson[1], David Hudson[1], Niheer Dasandi[1], and Yannis Theocharis[2]

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

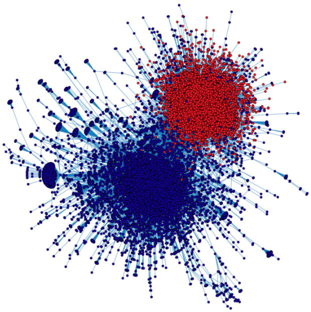
1. Big data as a new source of information
    - Behavior, opinions, and latent traits
    - Interpersonal networks
    - Elite behavior
2. How big data and social media affect social behavior
    - Mass protests
    - Political persuasion
    - Social capital
    - Political polarization

Social media as echo chambers?

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



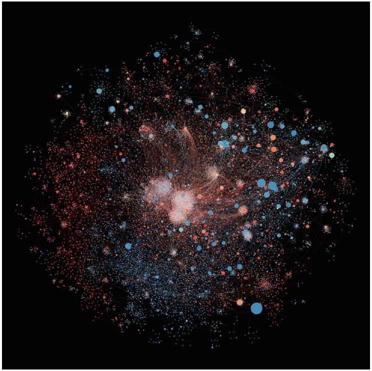
Adamic and Glance (2005)



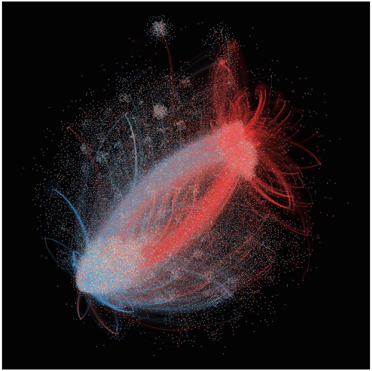
Conover et al (2012)

- ...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?

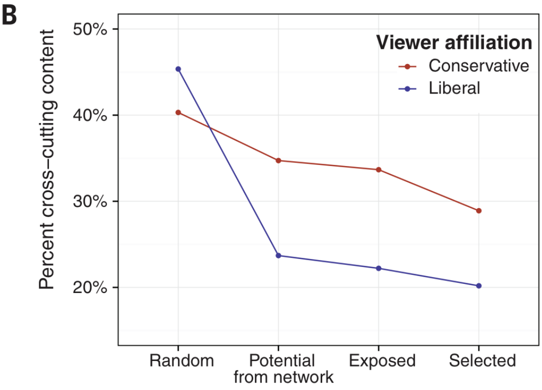


Fig. 3. Cross-cutting content at each stage in the diffusion process. (**A**) Illustration of how algorithmic ranking and individual choice affect the proportion of ideologically cross-cutting content that individuals encounter. Gray circles illustrate the content present at each stage in the media exposure 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

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