Collecting and Analyzing Social Media Data with R

Pablo Barberá
Politics – SMaPP Lab
New York University

slides and code:
github.com/pablobarbera/social-media-workshop

February 2nd, 2015

LinkedIn
skype
foursquare
LiVEJOURN
vimeo
YouTube
Insta
facebook
g+
my Blogger
twitter
flickr
Be
tumblr.
Like
yelp

Why should we care about social media?

1. Social media usage is widespread

Widespread use of social media sites

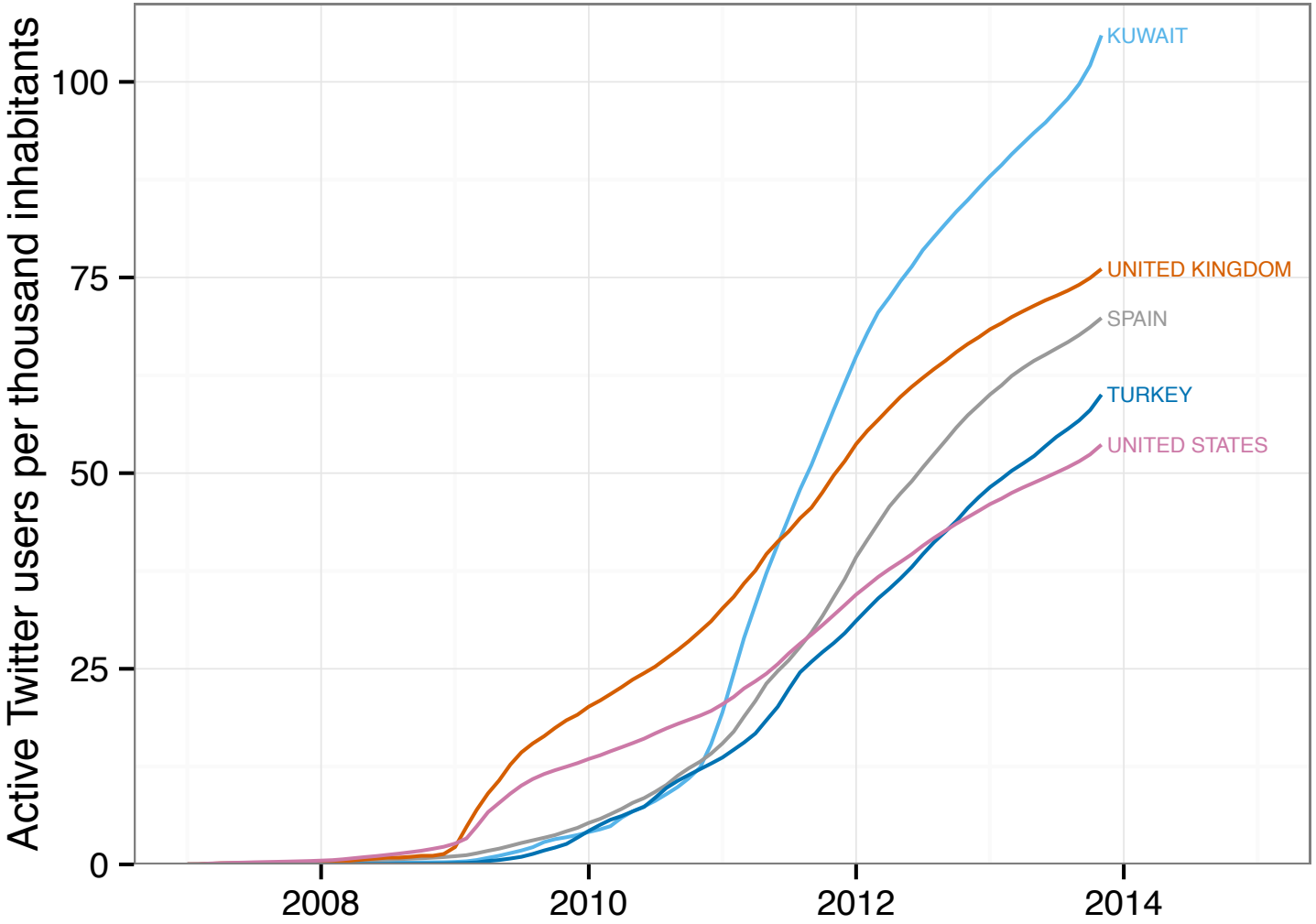
- One in every ten people in the world logged onto Facebook yesterday.
- 71% of online adults in the US use Facebook (84% use among ages 18–29)
- 400+ million tweets are sent everyday by 200+ million active users worldwide
- 23% of online adults in the US use Twitter (31% use among ages 18–29)
- Instagram has 300+ million active users (26% of online adults in the US)

(Sources: Pew Research Center (2014), Twitter and Facebook official statistics)

Why should we care about social media?

1. Social media usage is widespread
2. Social media usage is increasing

Social media usage is increasing



(Source: Zeitzoff and Barberá, MPSA 2014)

Why should we care about social media?

1. Social media usage is widespread
2. Social media usage is increasing
3. Political content on social media

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

Justin Amash
December 10, 2014 ·

Sec. 309 of the new Intelligence Authorization Act permits the U.S. government to acquire, retain, and disseminate nonpublic telephone or electronic communications to or from a U.S. person. I demanded a roll call vote on the bill, and I will be voting NO.

Like · Comment · Share · 2,534 120 673

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

Donald J. Trump
@realDonaldTrump

Follow

This very expensive GLOBAL WARMING bullshit has got to stop. Our planet is freezing, record low temps,and our GW scientists are stuck in ice

Reply · Retweet · Favorite · More

RETWEETS 2,007 FAVORITES 1,145

7:39 PM - 1 Jan 2014

Social media and politics

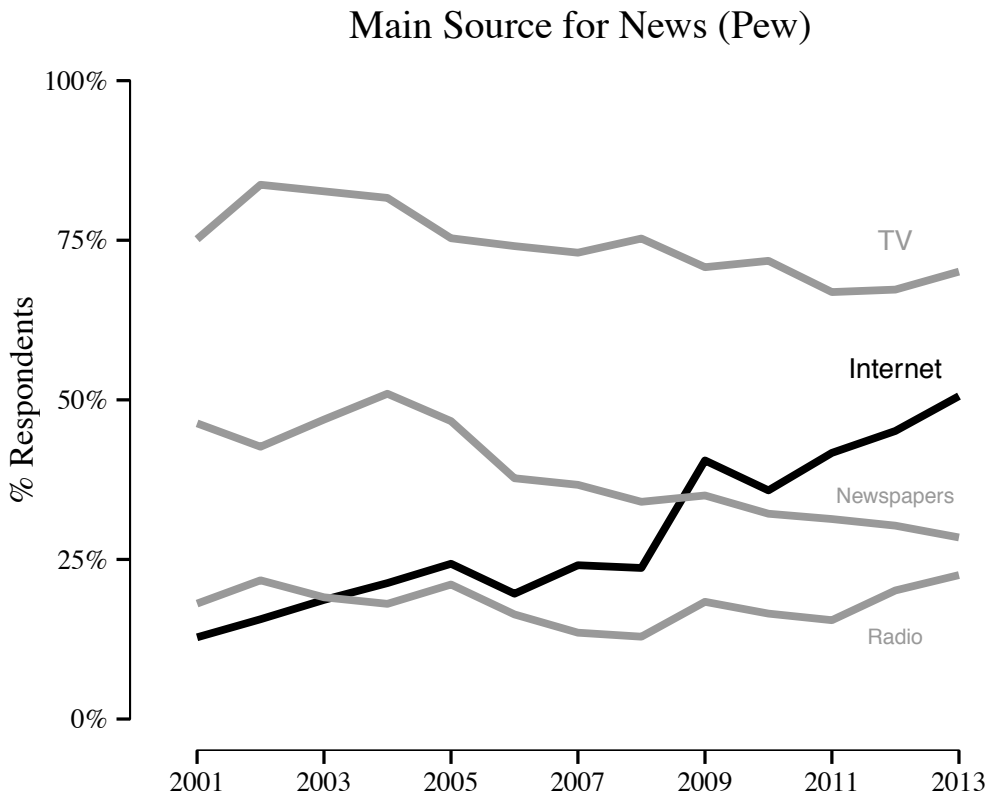
- 99% of Members of the US Congress have an active social media account
- 80% of governments have a presence on Twitter
- "Traditional" media outlets rely on social media to promote their content
- 50% of social media users in U.S. share information about news stories, images or videos about current events
- 46% have discussed a news issue or event on social media

(Sources: Electionista; Zeitzoff and Barberá, MPSA 2014; Pew Research Center)

Why should we care about social media?

1. Social media usage is widespread
2. Social media usage is increasing
3. Political content on social media
4. Social media is a primary source of political information

- Large changes in citizens' news consumption habits



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

- 41% of Americans see news on social media every day (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)

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Behavior, opinions, and latent traits

- Digital footprint: check-ins, conversations, geolocated pictures, likes, shares, retweets, …

→ Non-intrusive measurement of behavior and opinion

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

Behavior, opinions, and latent traits

- Digital footprint: check-ins, conversations, geolocated pictures, likes, shares, retweets, …

→ Non-intrusive measurement of behavior and opinion

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*

Behavior, opinions, and latent traits

- Digital footprint: check-ins, conversations, geolocated pictures, likes, shares, retweets, …

→ Non-intrusive measurement of behavior and opinion

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

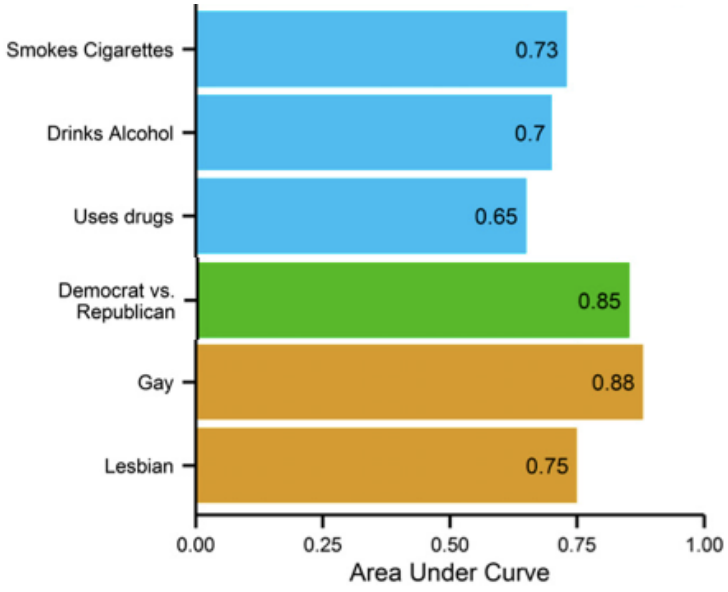


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

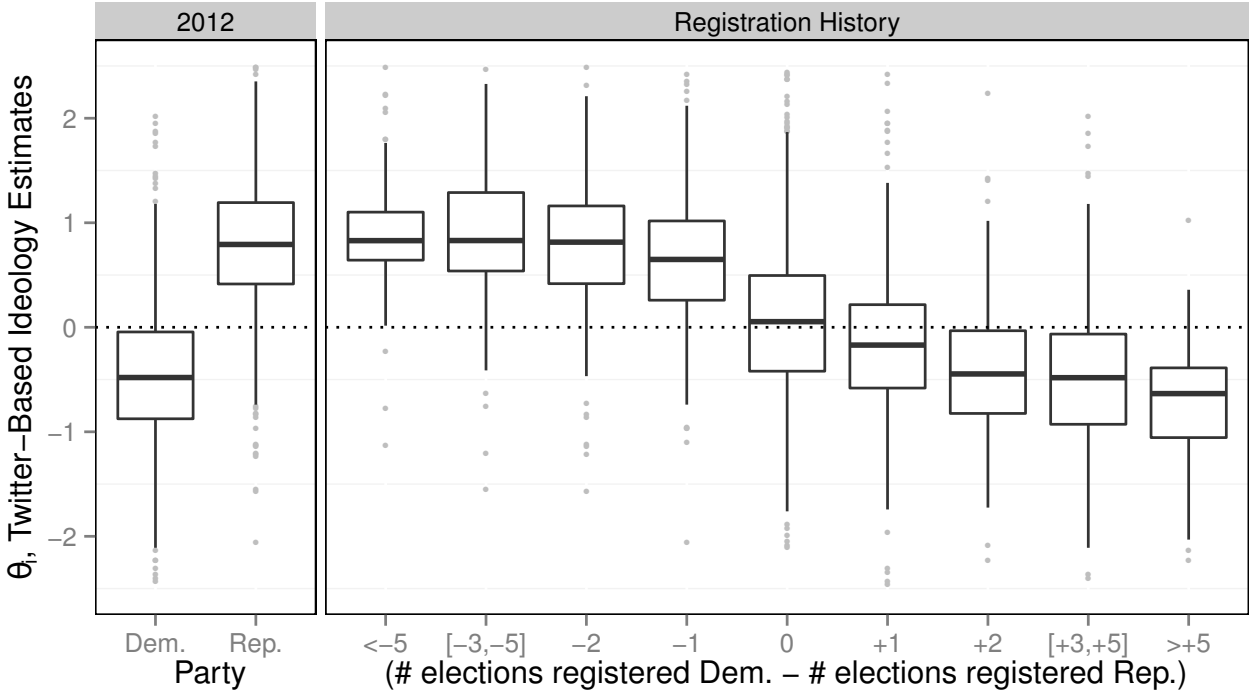
Kosinki et al, 2013, "Private traits and attributes are predictable from digital records of human behavior", *PNAS* (also personality, PNAS 2015)

Behavior, opinions, and latent traits

- Digital footprint: check-ins, conversations, geolocated pictures, likes, shares, retweets, …

→ Non-intrusive measurement of behavior and 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*

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

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).

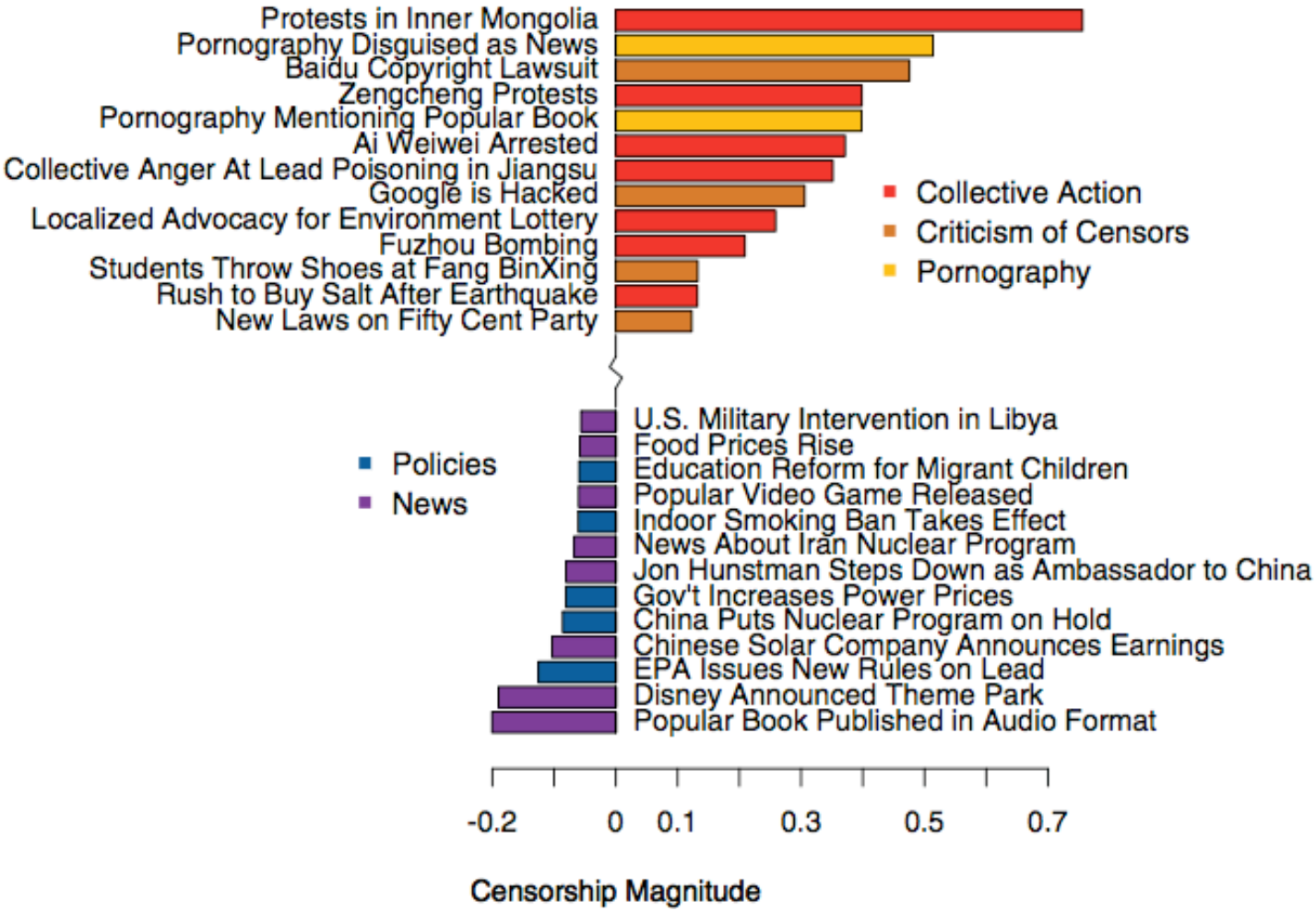
Jones et al, 2013, "Inferring Tie Strength from Online Directed Behavior", *PLOS One*

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Elite behavior

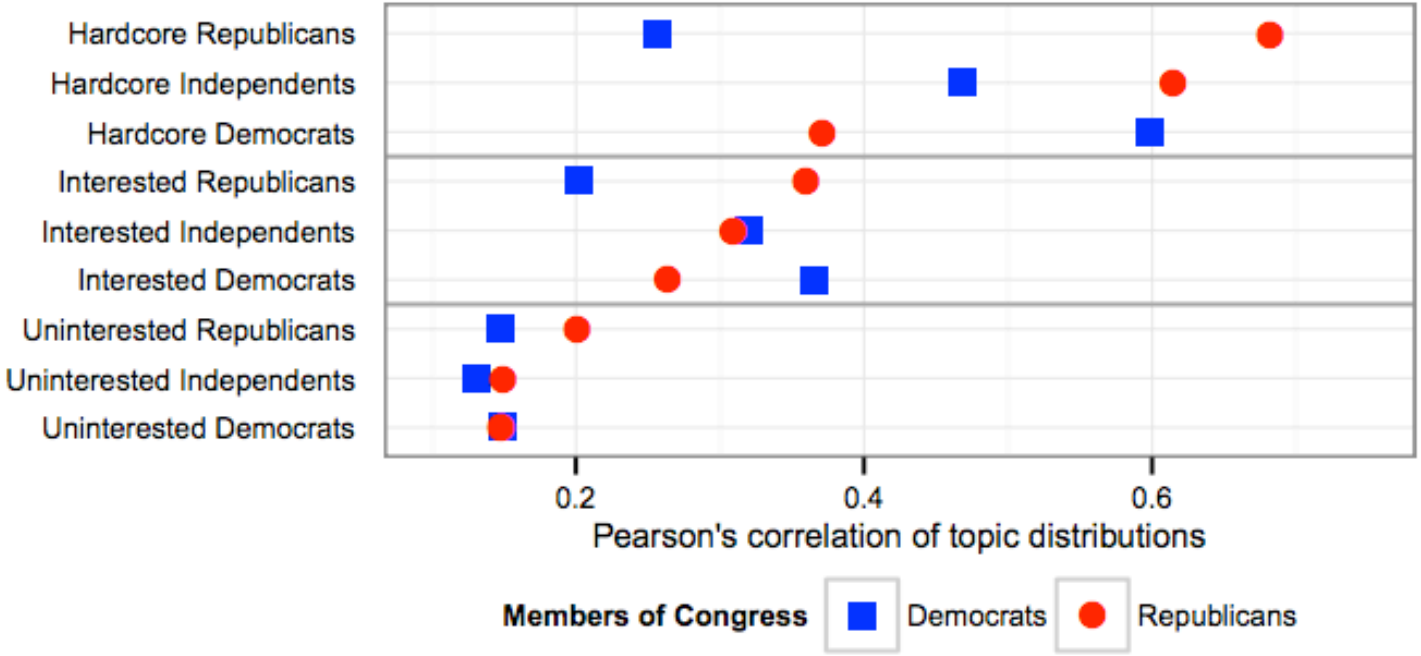
- Authoritarian governments' response to threat of collective action



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
- How legislators in democracies are responsive to their constituents in how they communicate publicly



Barberá et al, 2014, "Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data", *APSA*

Elite behavior

- Authoritarian governments' response to threat of collective action
- How legislators in democracies are responsive to their constituents in how they communicate publicly
- Estimation of conflict intensity in real time

Journal of Conflict Resolution
55(6) 938-969
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
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]

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

THE NEW YORKER

ANNALS OF INNOVATION

SMALL CHANGE

Why the revolution will not be tweeted.

by Malcolm Gladwell

OCTOBER 4, 2010

Social media can't provide what social change has always required.

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - 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

Post Politics

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

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Social capital

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

---

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

American Politics Research
1–31
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1532673X14557942
apr.sagepub.com
SAGE

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

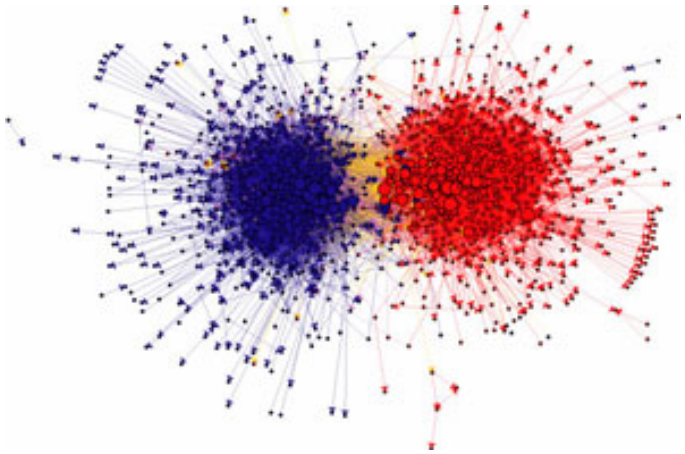
Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

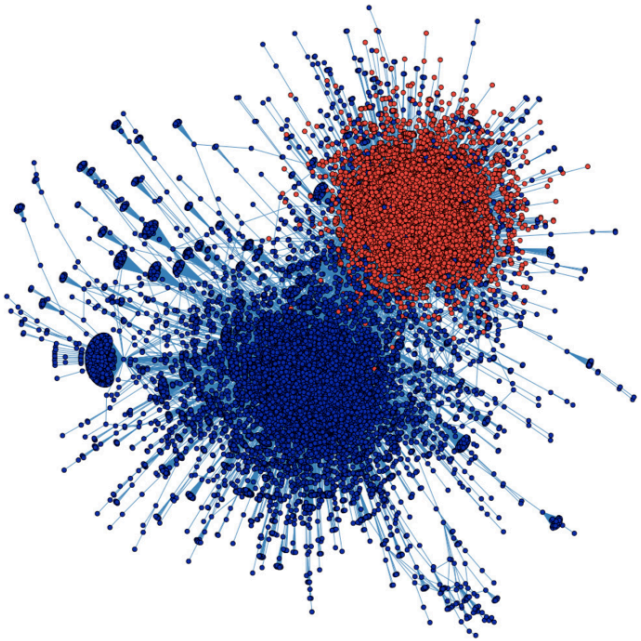
Political polarization

Social media as *echo chambers* or *filter bubbles*:

- communities of like-minded individuals (homophily)



Adamic and Glance (2005)



Conover et al (2012)

- ...generates selective exposure to congenial information
- ...increases political polarization (Sunstein, Prior)

Political polarization

Social media usage induces political moderation (Barberá, 2015)

1. Inadvertent exposure to political messages
   - "Your friends deliver the news" (Messing and Westwood, 2014)
   - Less selective exposure
2. More frequent interactions beyond immediate personal network
   - "The strength of weak ties" in providing novel information (Granovetter, 1973; Bakshy et al, 2012)

...increases exposure to dissonant views

...and therefore mass political polarization decreases.

Two different approaches to the study of social media and politics:

1. Social media as data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
2. Social media as a variable
   - Mass protests
   - Political persuasion
   - Social capital
   - Political polarization

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Collecting Social Media Data

Two different methods:

1. Screen scraping: extract data from source code of website
2. Web APIs (application programming interface): use a set of structured https requests that return JSON or XML files

Types of APIs:

1. RESTful APIs: queries for static information in current moment (e.g. user profiles, posts, etc.)
2. Streaming APIs: changes in users' data in real time (e.g. new messages, deletions, etc.)

Potential issues

1. Rate limits: restrictions on number of API calls by user and period of time (APIs are expensive!)
2. Ongoing debate on replication of social science research using social media data

Connecting with an API

Constructing a REST API call

- Baseline URL: http://graph.facebook.com/
- Parameters: ?ids=barackobama,johnmccain

Response often in JSON format. (example)

Authentication

- Most common is an open standard called OAuth
- Connections without sharing username and password, only temporary tokens that can be refreshed
- `httr` package in R implements most cases (examples)

Interacting with social media APIs

R packages

- Twitter: twitteR for REST, streamR for Streaming
- Facebook: Rfacebook
- Instagram: instaR (not on CRAN)

Why R? Most common programming language in data science, 5,000+ packages, great documentation, “it just works”

Other APIs: CRAN Task View for Web Technologies and Services

Equivalent libraries for python, java, ruby... whatever works for you!

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Fork my repo! github.com/pablobarbera/social-media-workshop

Code:

- Twitter
  - 01-twitter-data-collection.r
  - 02-twitter-data-analysis.r
- Facebook
  - 03-facebook-data-collection.r
  - 04-facebook-data-analysis.r
- Instagram
  - 05-instagram-data.r

Slides: slides/social-media-workshop.pdf
Data: backup/

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
   - R library: twitteR
2. Streaming API:
   - Connect to the "stream" of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location
     - 2.3 Sample stream: 1% random sample of tweets
   - R library: streamR

Important limitation: tweets can only be downloaded in real time (exception: user timelines, $\sim$ 3,200 most recent tweets are available)

Anatomy of a tweet

Anatomy of a tweet

Tweets are stored in JSON format:


```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
       Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 54873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 756411,
  "favorite_count": 288867,
  "lang": "en"
}
```

Collecting Twitter Data

The R script 01-twitter-data-collection.r shows how to:

- Create an OAuth token to authenticate
- Extract basic user information
- Search tweets that contain a given keyword (REST)
- Collect tweets filtering by keywords and location (Streaming)
- Collect a random sample of tweets
- Download all tweets sent by a given user

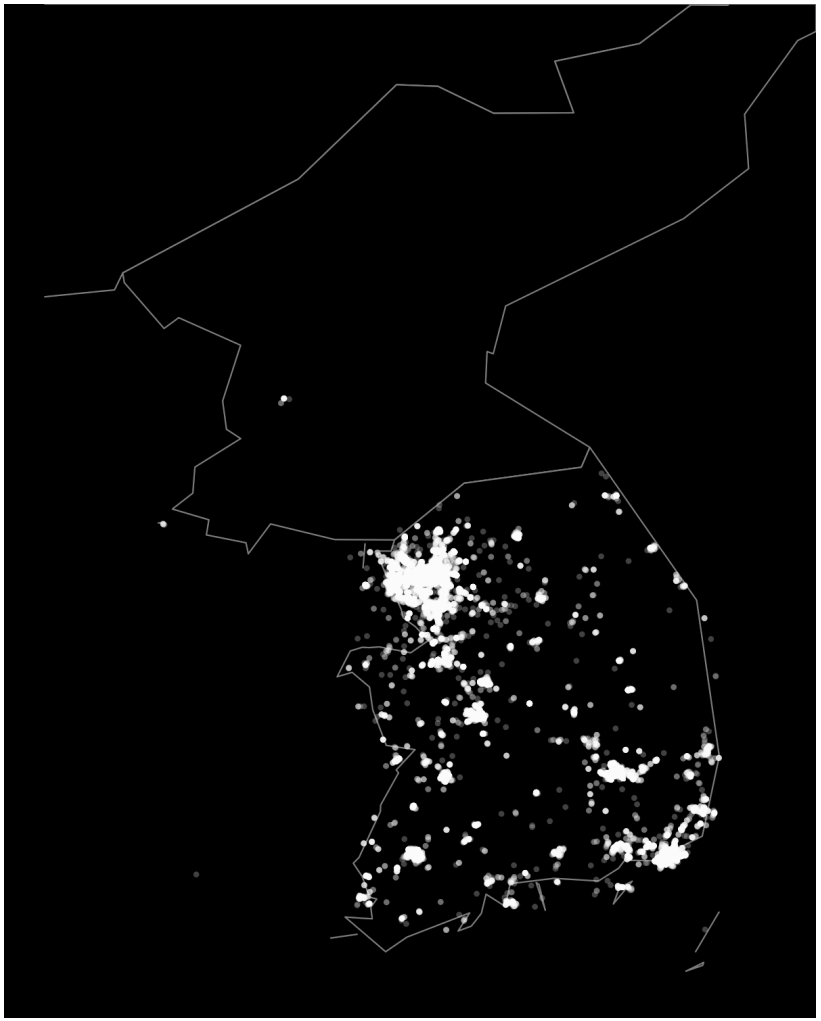
Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer (1% of all tweets)
  - Stream connections tend to die spontaneously. Restart regularly.
  - Lots of invalid content in stream. If it can't be parsed, drop it.
- My workflow:
  - Amazon EC2 Ubuntu micro instance (free tier)
  - Cron jobs to restart R scripts every hour.
  - Save tweets in .json files or in MongoDB.
  - For large .json files, preprocess with python (see: github.com/pablobarbera/pytwools)

Analyzing Twitter Data

The R script 02-twitter-data-analysis.r shows:

1. How to work with geolocated tweets
   - Map distribution of tweets in the US
   - How many tweets from each state?
2. How to measure opinions on Twitter
   - Supervised sentiment analysis using a dictionary of positive and negative words

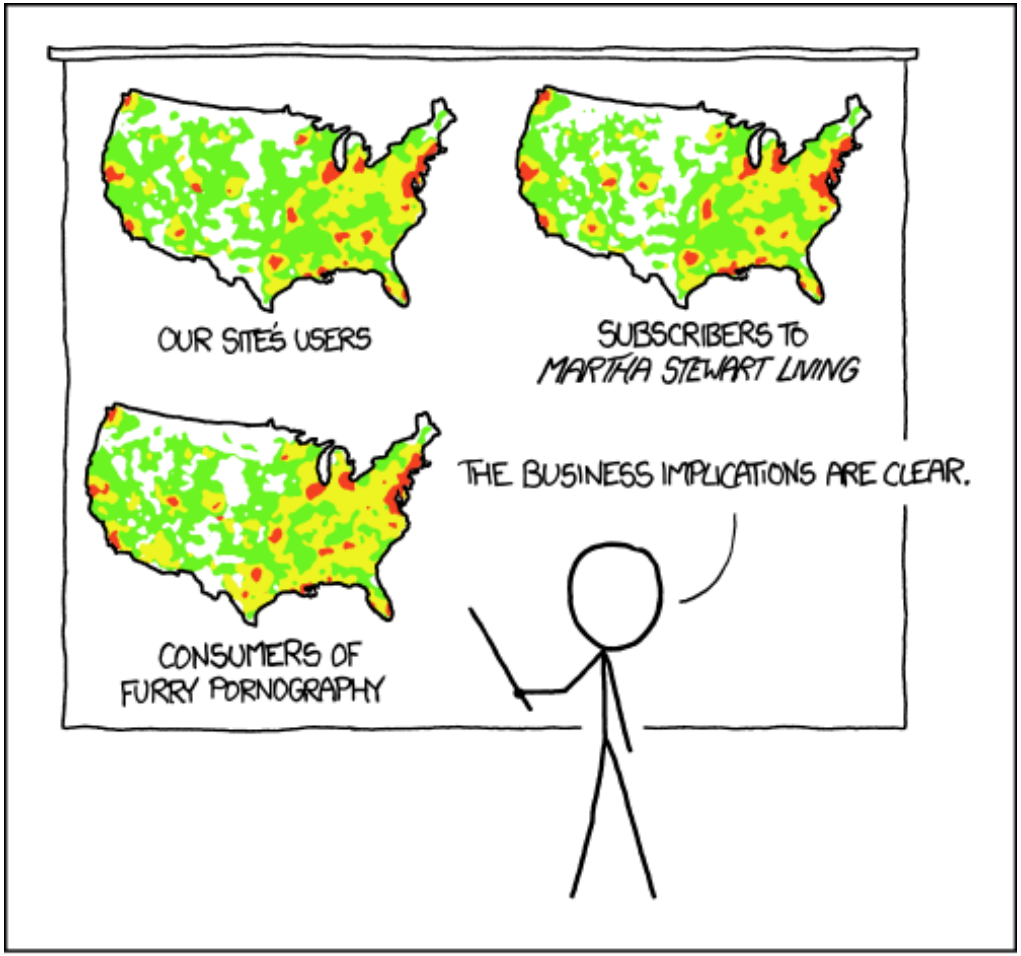


Tweets from Korea: 40k tweets collected in 2014 (left)
Korean peninsula at night, 2003 (right). Source: NASA.

Who is tweeting from North Korea?

Twitter user: @uriminzok_engl

But remember...



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

Sentiment analysis

Sentiment, tone, valence, affective score... = % positive messages about a subject

Two main approaches:

1. Dictionary methods
   - Dictionary of positive and negative words (issue-specific)
   - Count number of times they appear
2. Machine learning
   - Training data: random sample of posts manually labelled as positive or negative
   - *Classifier* learns from data and predicts sentiment of unseen posts
   - Many different classifiers: regularized logistic regression, Naive Bayes, SVM, BART, ensemble methods...
   - Accuracy, precision, recall, test data, cross-validation.

Challenges: sarcasm, intensity, subject identification.
In practice, accuracy over 80% is impossible (and that's ok)

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Collecting Facebook data

Facebook allows access to two different types of data through the public API:

1. Data from public Facebook pages (posts, likes, comments)
2. User's personal data (profile, checkins, likes...)

Some public user data (gender, location) was available through previous versions of the API (not anymore)

Access to other (anonymized) data used in published studies requires permission from Facebook

R library: Rfacebook

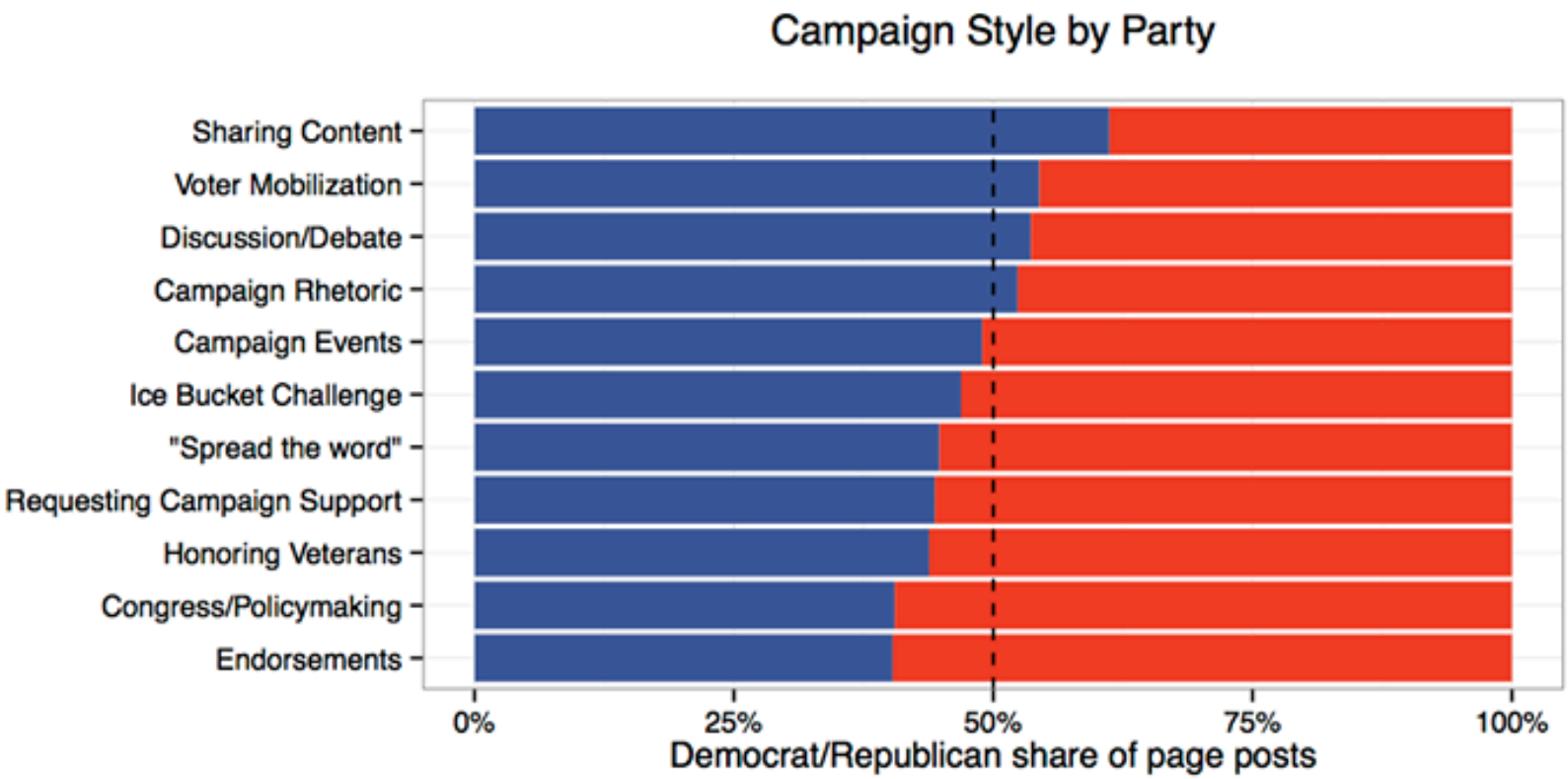
Collecting Facebook Data

The R script 03-facebook-data-collection.r shows how to:

- Use OAuth to authenticate
- Display your profile information
- Capture data from a Facebook page
- Collect likes and comments data from a public post

Facebook data analysis

What issues do Members of the U.S. Congress discuss on Facebook?



Messing et al, 2014, "Campaign Rhetoric and Style on Facebook in the 2014 U.S. Midterms" (Facebook Data Science Blog)

Analyzing Facebook Data

The R script 04-facebook-data-analysis.r shows how to:

- Find a list of Facebook pages for Members of Congress using the NYTimes API
- Collect Facebook page data for each member
- Parse and clean text from their posts
- Run a topic modeling technique (LDA) to estimate the amount of time they discuss each issue
- Interpret the results of this model

Collecting and Analyzing Social Media Data with R

1. Motivation
2. Overview of social media research
3. Social media APIs
4. Tools and applications:
   - 4.1 Twitter
   - 4.2 Facebook
   - 4.3 Instagram

Instagram data

What is available through Instagram API:

- Search and download pictures that mention a given hashtag on its caption, or that were sent from a specific location
- Collect information about these pictures (creation date, caption, author, filter, hashtags …)
- Download pictures sent by a given user
- Count number of photos that mention a specific hashtag

Examples: 05-instagram-data.r

Thanks! Questions?

materials: github.com/pablobarbera/social-media-workshop

website: pablobarbera.com

twitter: @p_barbera