ECPR Methods Summer School: Automated Collection of Web and Social Data

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:
github.com/pablobarbera/ECPR-SC103

Social event

Today!

Wednesday Aug. 2nd, 6pm

Location TBA


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

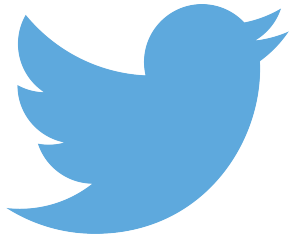
Why should we care about social media?

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)

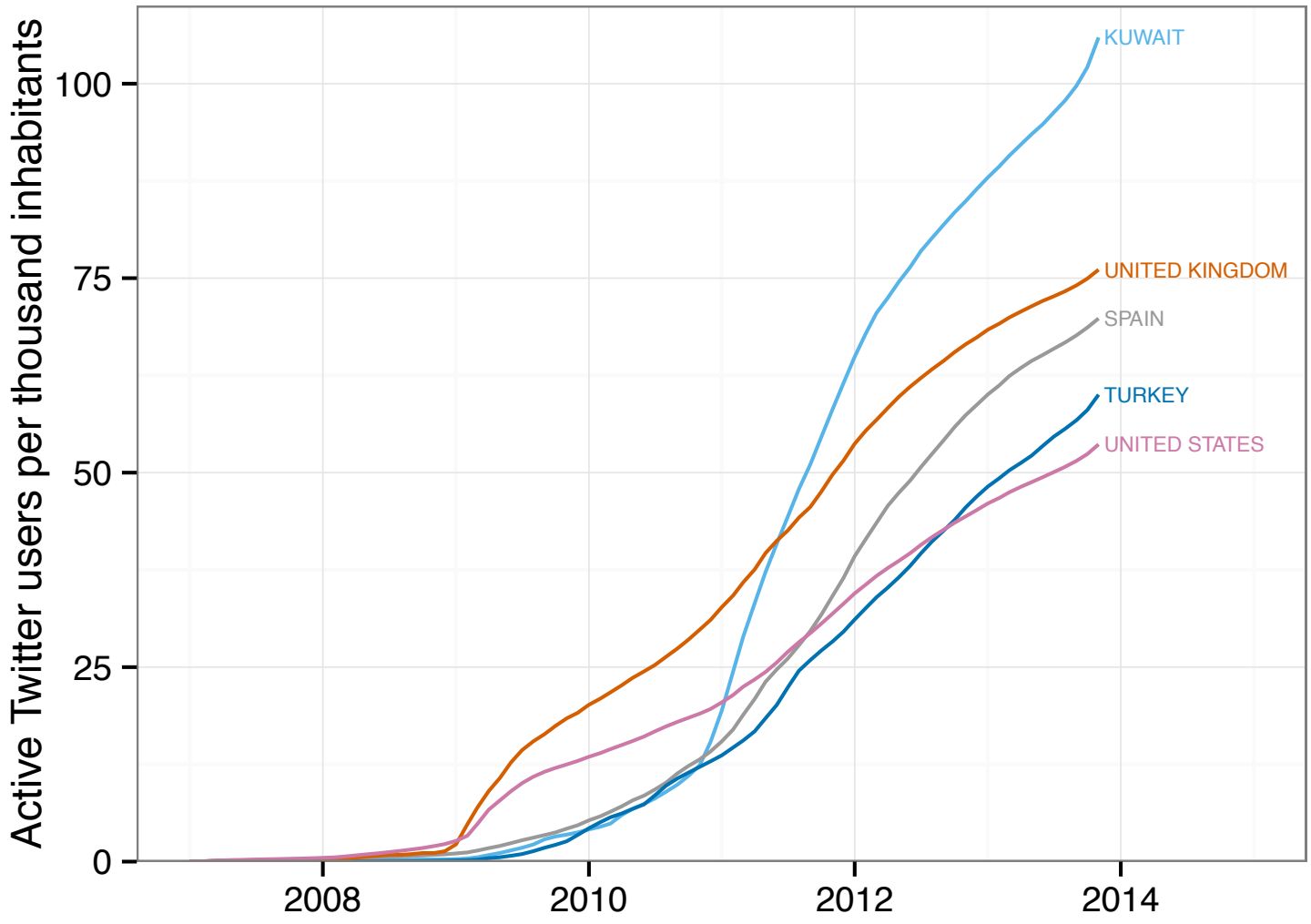


(Sources: Pew Research Center (2014), Twitter and Facebook official statistics)

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, ISQ 2017)

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

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, ISQ 2017; Pew Research Center)

Social media as a new campaign tool:

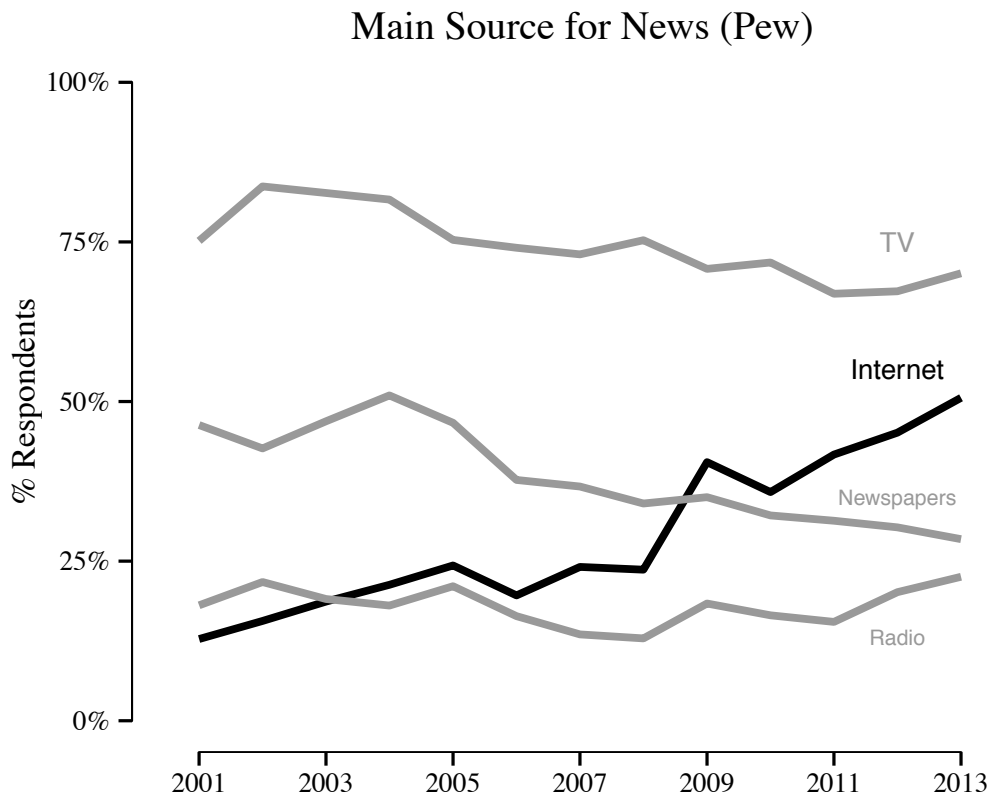
> *"Let me tell you about Twitter. I think that maybe I wouldn't be here if it wasn't for Twitter. [...] Twitter is a wonderful thing for me, because I get the word out... I might not be here talking to you right now as president if I didn't have an honest way of getting the word out."*
>
> ***Donald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

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.

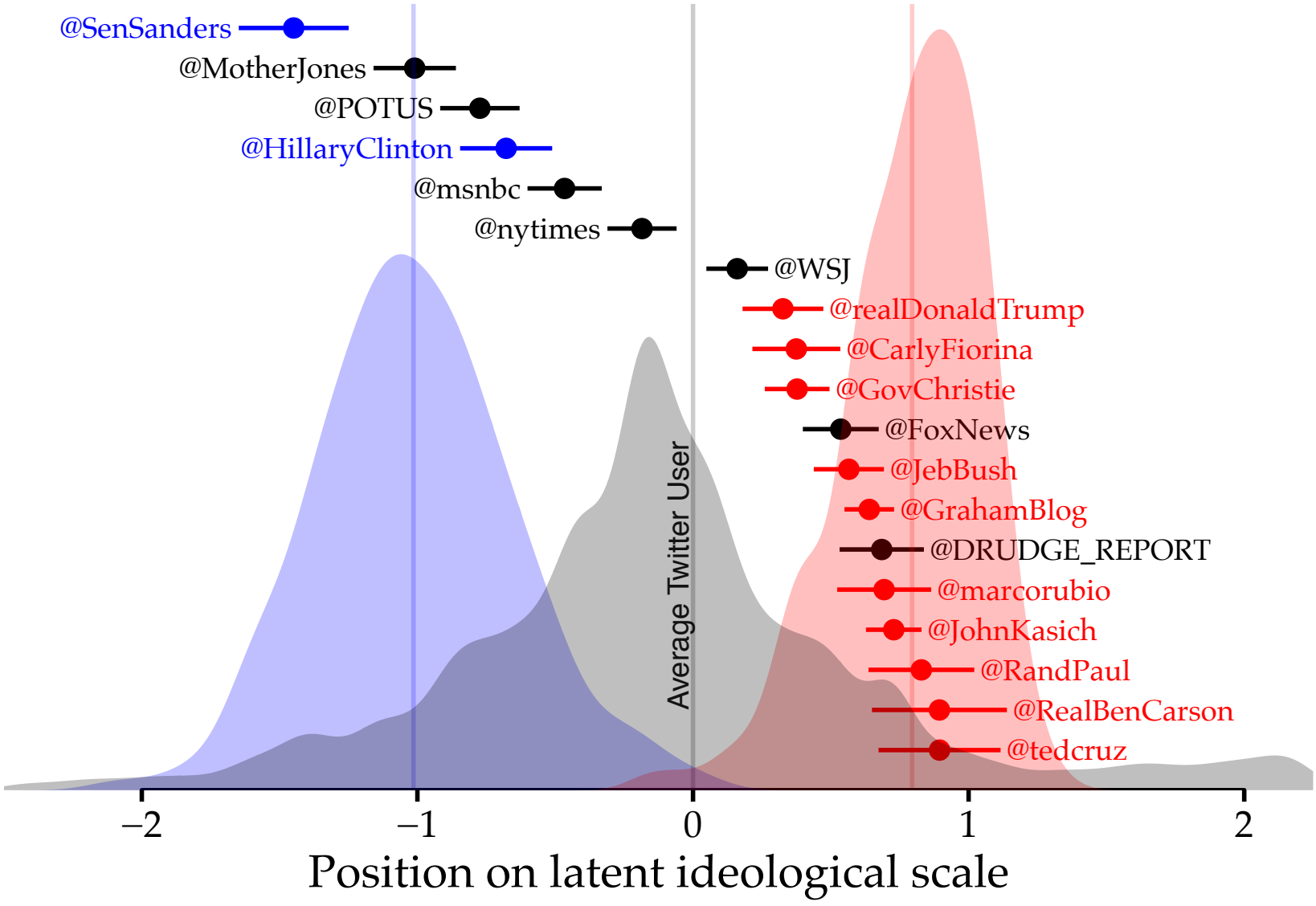
- 41% of Americans see news on social media every day
- 27% of online EU citizens use social media to get news on national political matters (Eurobarometer, Fall 2012)
- Social media: top source of news for U.S. young adults

Social media research

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

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

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

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)*

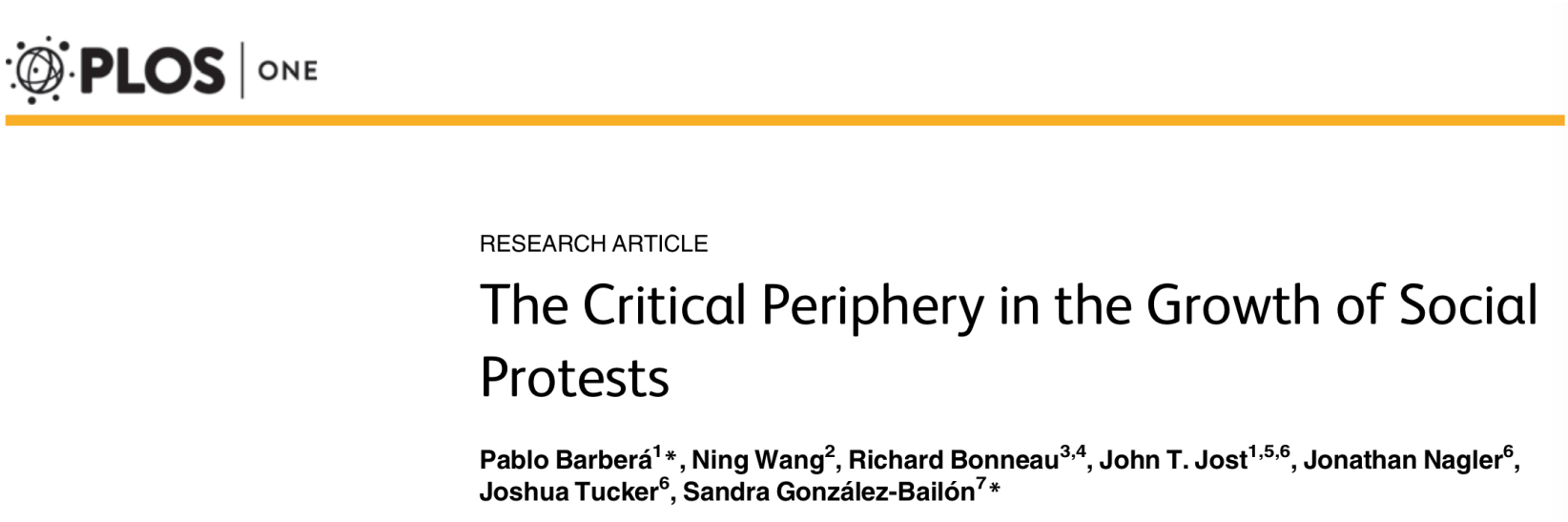
You can't simply join a revolution any time you want, contribute a comma to a random revolutionary decree, rephrase the guillotine manual, and then slack off for months. Revolutions prize centralization and require fully committed leaders, strict discipline, absolute dedication, and strong relationships.

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*

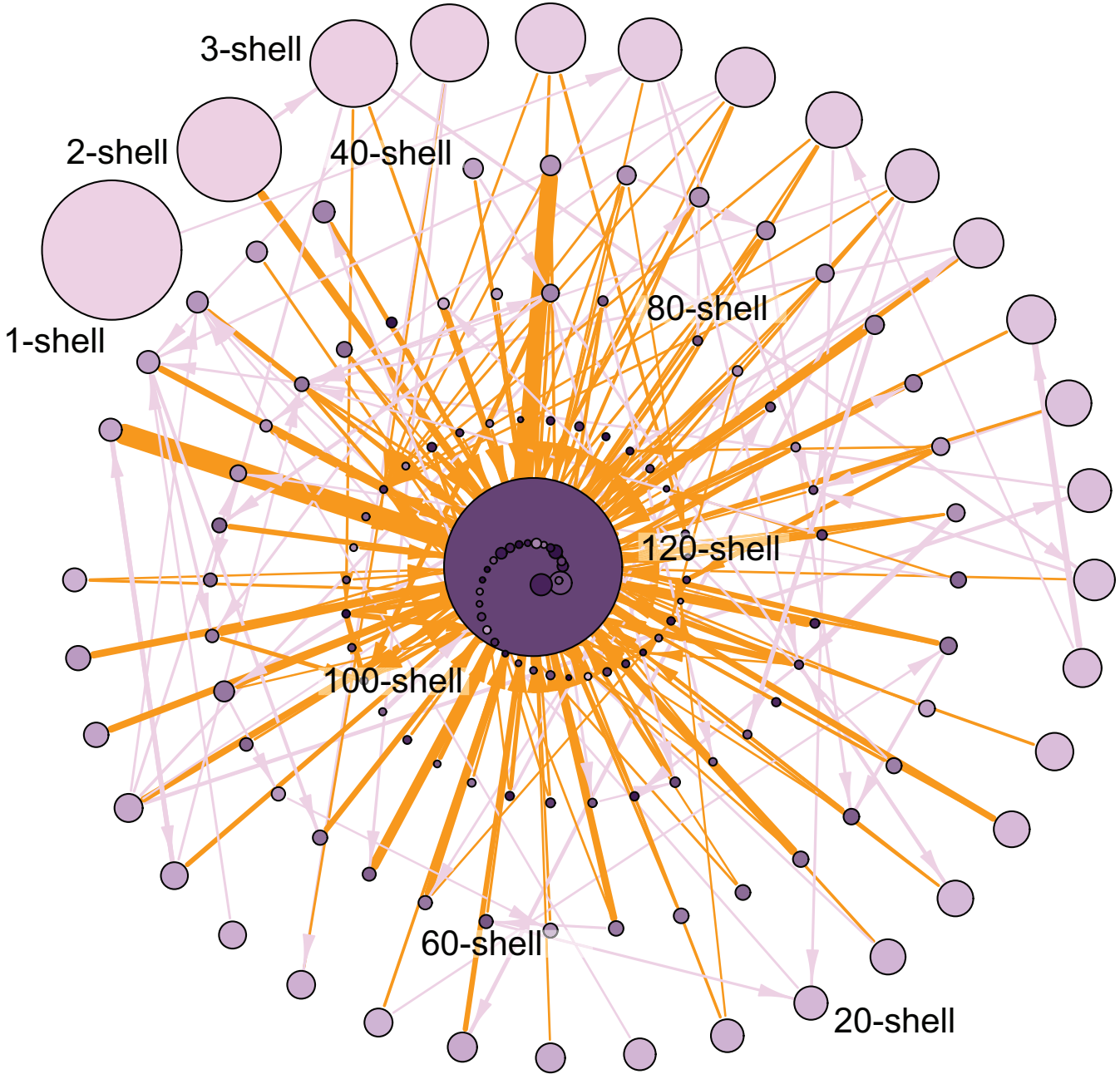
parody or reality?

the critical periphery



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

k-core decomposition of #OccupyGezi network



periphery

core

activity
(no. of tweets)

in Taksim

max 18%

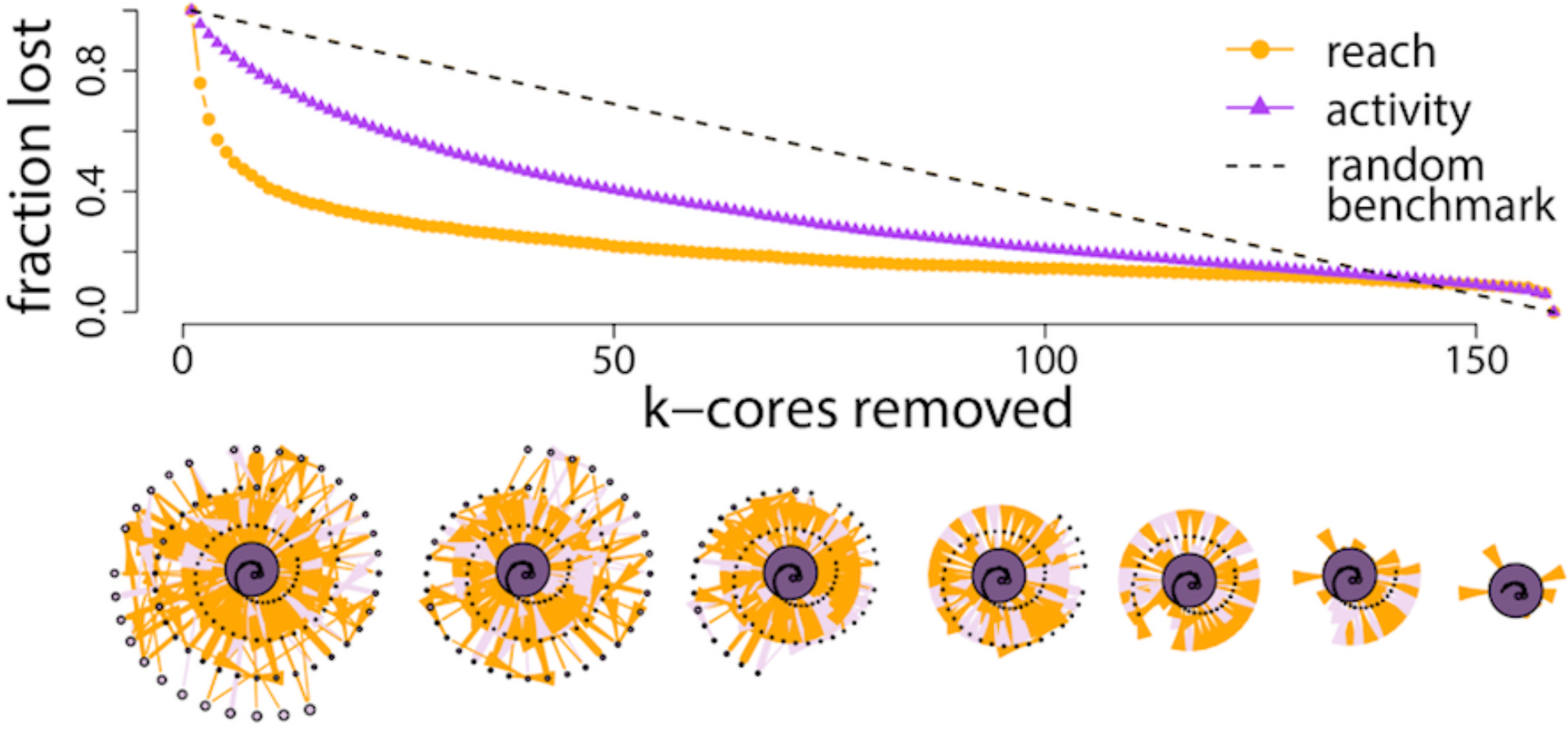
min .25%

RTs

periphery to core

periphery to periphery

Relative importance of core and periphery

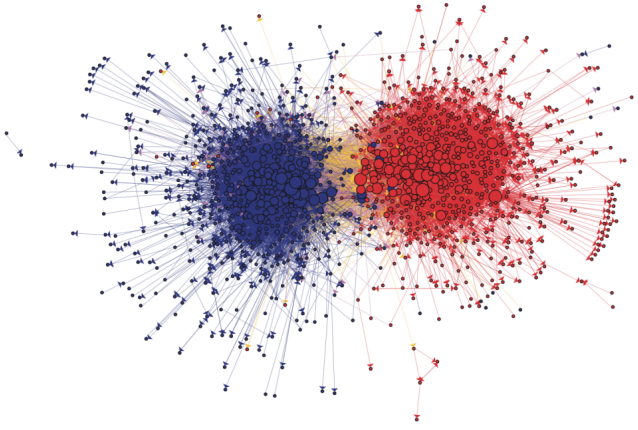


reach: aggregate size of participants' audience

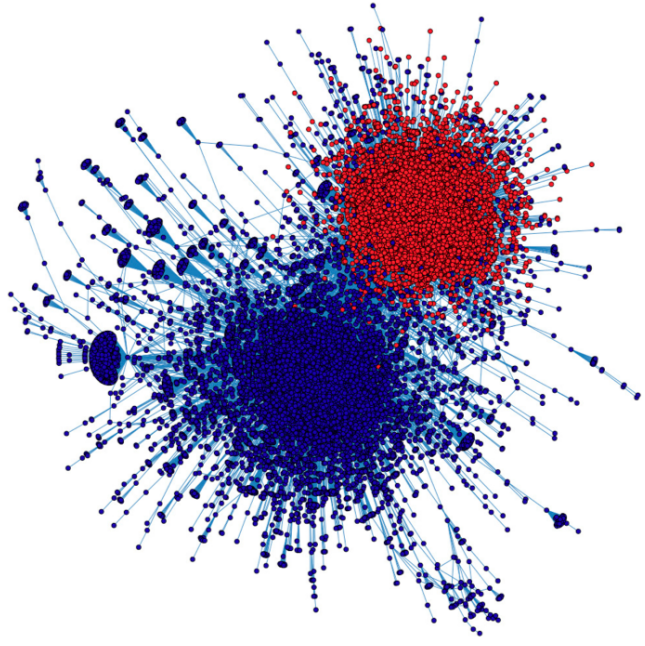
activity: total number of protest messages published (not only RTs)

Social media as echo chambers?

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



Adamic and Glance (2005)

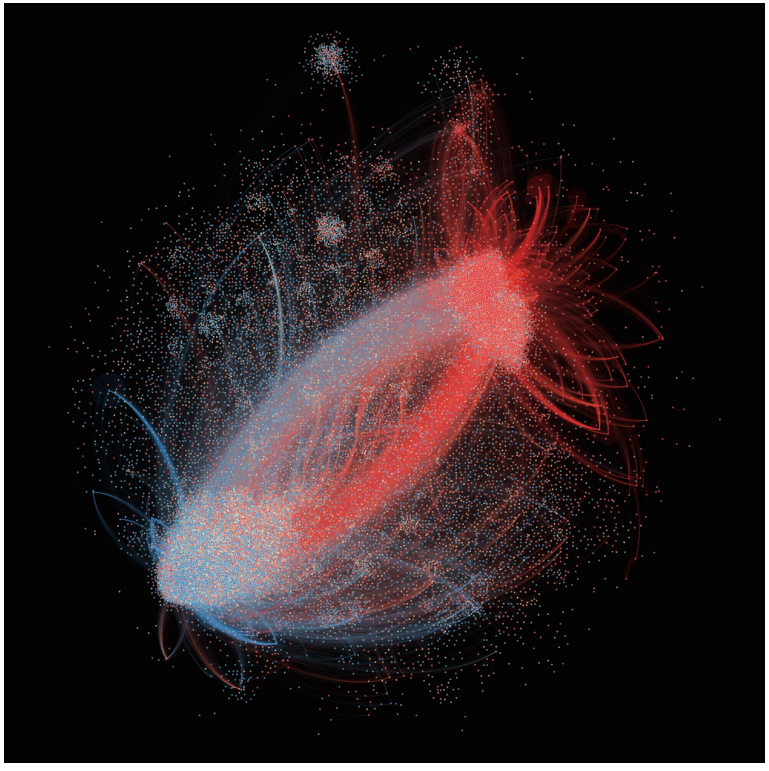


Conover et al (2012)

- ...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

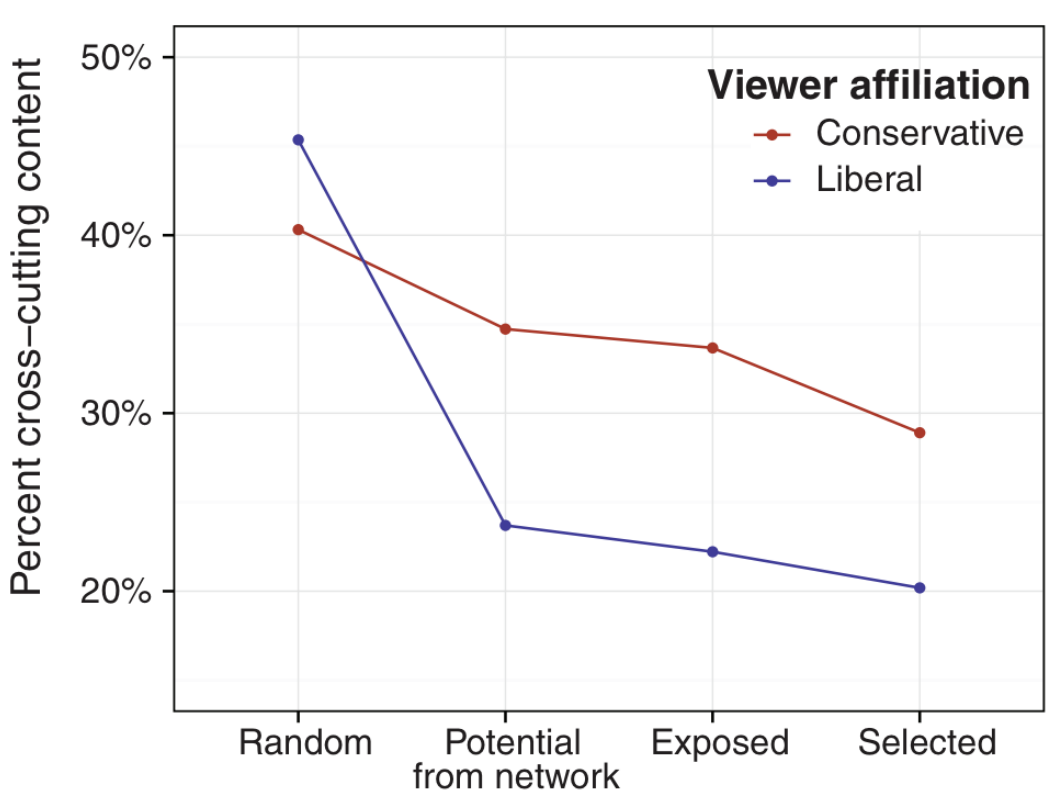


2012 Election

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.*

Twitter data

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
   - R library: netdemR (also twitteR, rtweet)
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"
}
```

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - 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, cloud computing
  - Cron jobs to restart R scripts every hour.
  - Save tweets in .json files, one per day.
  - For large .json files, preprocess with python (see: github.com/pablobarbera/pytwools)

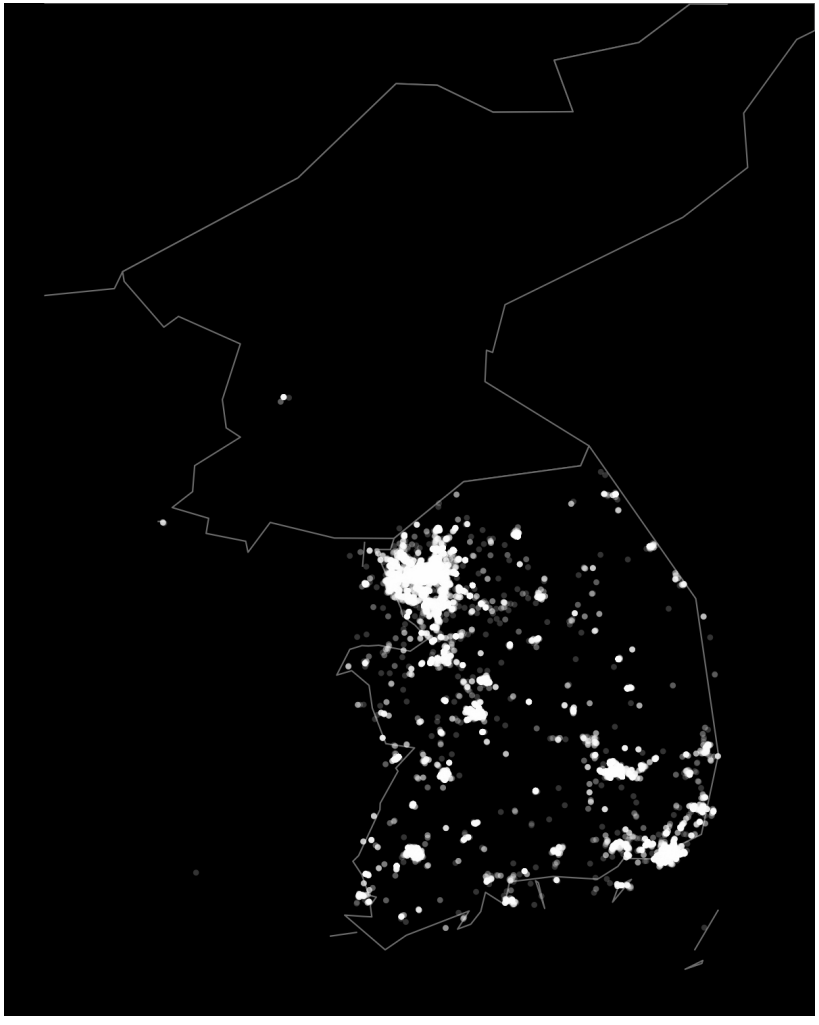
Sampling bias?

Morstatter et al, 2013, *ICWSM*, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose":

- 1% random sample from Streaming API is not truly random
- Less popular hashtags, users, topics... less likely to be sampled
- But for keyword-based samples, bias is not as important

González-Bailón et al, 2014, *Social Networks*, "Assessing the bias in samples of large online networks":

- Small samples collected by filtering with a subset of relevant hashtags can be biased
- Central, most active users are more likely to be sampled
- Data collected via search (REST) API more biased than those collected with Streaming API

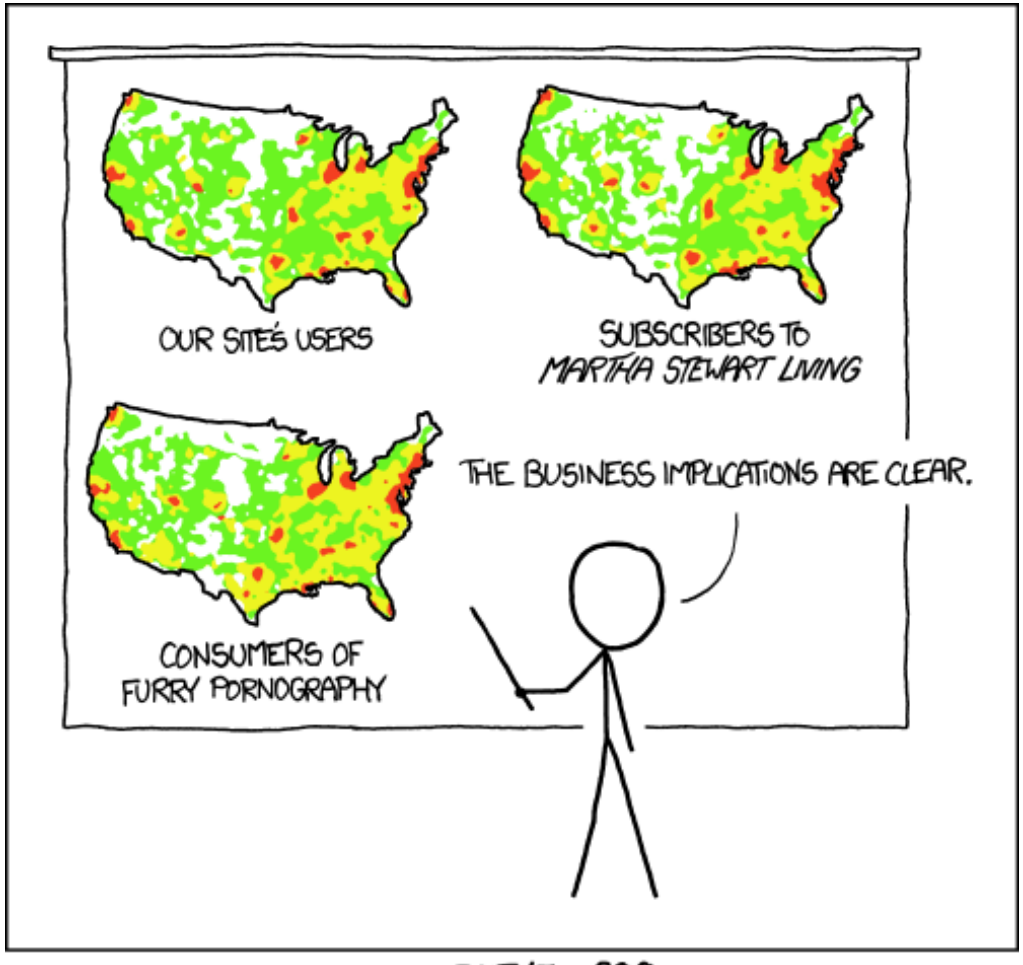


Tweets from Korea: 40k tweets collected in 2014 (left)
Korean peninsula at night, 2003 (right). Source: NASA.

Who is tweeting from North Korea?

Twitter user: @uriminzok_engl

But remember...



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

Facebook data

Collecting Facebook data

Facebook only allows access to public pages' data through the Graph API:

1. Posts on public pages and groups
2. Likes, reactions, comments, replies...

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