POIR 613: Computational Social Science

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Course website:

pablobarbera.com/POIR613/

Today

1. Project milestones
   - Nov 25 (Monday): full draft
   - Dec 4 (Wednesday): 8-minute presentations
   - Dec 18 (Tuesday): submission
2. Other announcements
   - Next week: informal Q&A on methods job market + industry opportunities?
   - Dec 18: happy hour after class?
3. Plan for today:
   - Social network analysis: diffusion dynamics
   - Collecting Twitter data
   - Review of SQL materials
   - Solutions to challenge 9

Social network analysis:
diffusion dynamics

Diffusion dynamics

Diffusion via social ties are key mechanisms explaining how **diseases, information, and behavior spreads.**

Diffusion dynamics

Two types of diffusion processes:

1. Simple contagion
   - One contact is enough for contagion (adopting behavior, receiving information, etc)
   - Example: spread of diseases
2. Complex contagion
   - Multiple and/or diverse contacts are necessary for contagion
   - *Threshold* models: adopt behavior if x% of your ties have already adopted it
   - Examples: online memes, technology or social media adoption, collective action, public opinion change, etc.
   - Most common mechanism in social processes

Example from NetLogo

Contagion dynamics

Why does it matter? Interaction between network properties and diffusion dynamics:

- In highly clustered networks, complex contagion is unlikely to reach the entire network
- Simple contagion will be faster if it reaches a node with degree centrality
- In contrast, individuals with high betweenness centrality are key if contagion is complex

Social network analysis: tie strength

Tie strength

Not all ties are created equal:

- Strong ties: family, partner, close friends...
- Weak ties: distant relative, acquaintances, co-workers...

Where tie **strength** can be defined in terms of:

- Frequency of interaction
- Potential to persuade, trust
- Shared traits
- Many mutual contacts

The strength of weak ties

Granovetter (1973, AJS):

- Random sample of recent job changers in Boston
- “How often did you see the contact around the time they passed job information?” (measure of tie strength)
- Key finding: 55.6% saw contact only occasionally
- The strength of weak ties – Why?
  1. Less influential, but strength in numbers
  2. Bridges across loosely connected network components

The strength of weak ties

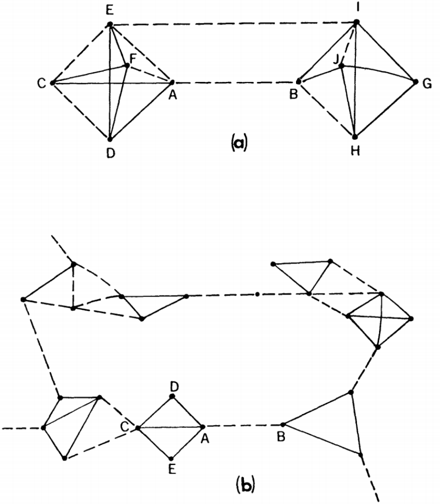


FIG. 2.—Local bridges. *a*, Degree 3; *b*, Degree 13. ——— = strong tie; – – – – = weak tie.

Source: Granovetter (1973, AJS):

Digital weak ties

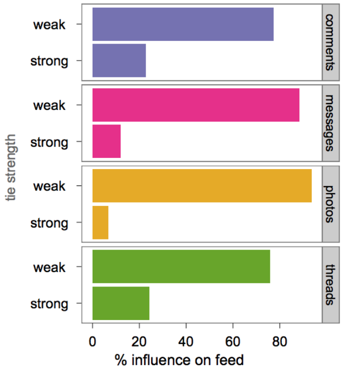


Figure 7: Weak ties are collectively more influential than strong ties. Panels show the percentage of information spread by strong and weak ties for all four measurements of tie strength. Although the probability of influence is significantly higher for those that interact frequently, most contagion occurs along weak ties, which are more abundant.

Bakshy et al (2012):

- Weak ties are responsible for most propagation of novel information on Facebook
- Strong ties provide redundant information
- Suggests contagion processes on Facebook may be more likely to be simple rather than complex

Twitter data

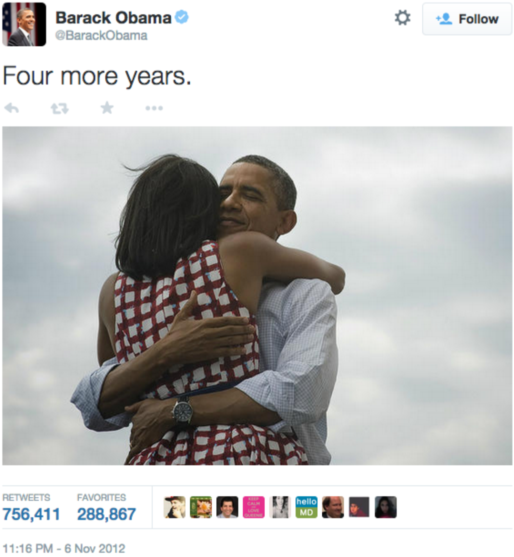
Twitter APIs

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: tweetscores (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



Barack Obama
@BarackObama

Four more years.

RETWEETS	FAVORITES
756,411	288,867

11:16 PM - 6 Nov 2012

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
    - Good to restart stream connections regularly.
- My workflow:
    - Amazon EC2, cloud computing
    - Cron jobs to restart R scripts every hour.
    - Save tweets in .json files, one per day.

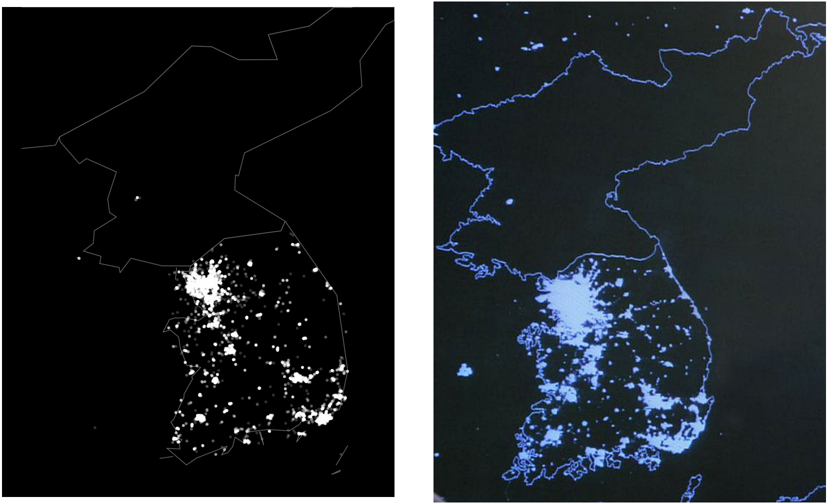
Sampling bias?

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

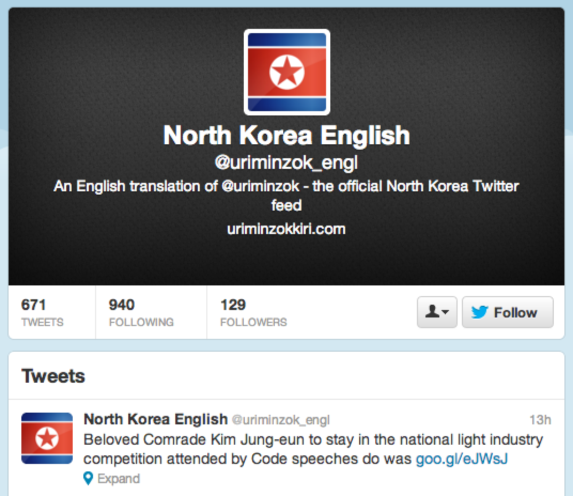
González-Bailón et al, 2014, *Social Networks*, “Assessing the bias in samples of large online networks”:

- 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?



North Korea English

@uriminzok_engl

An English translation of @uriminzok - the official North Korea Twitter feed

uriminzokkiri.com

671 TWEETS

940 FOLLOWING

129 FOLLOWERS

Tweets

North Korea English @uriminzok_engl 13h

Beloved Comrade Kim Jung-eun to stay in the national light industry competition attended by Code speeches do was goo.gl/eJWsJ

Twitter user: @uriminzok_engl