computational social media

lecture 3: tweeting

part 3

daniel gatica-perez
announcements

reading #3 will be presented today
Z. Tufekci, Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls, in Proc. AAAI ICWSM 2014

assignment #2 is due today (7pm)

projects: please contact me about your HREC submission if you have not done so yet.
this lecture

a human-centric view of twitter

1. introduction
2. twitter users & uses
3. understanding large-scale human behavior
4. inferring real-world events & trends
5. spreading information in the real world
spreading information in the real world

1. who says what to whom on twitter
2. cascading behavior in networks
3. structural virality of online diffusion
4. twitter and the news
1. who says what to whom on Twitter

Harold Lasswell (1948):
“who says what to whom in what channel with what effect”

“difficult to examine information flow in large populations”

"communication channels may have different effects"

credit: Photo by John Schnobrich on Unsplash https://unsplash.com/photos/2FPjIAyMQTA

the goal of media communication research
three models of communication

mass communication:
“one-way message transmission from one source to a large, relatively undifferentiated and anonymous audience”

interpersonal communication:
“two-way message exchange between two or more individuals”

two-step flow of communication:
“mass media influence the public only indirectly”
“the critical intermediate layer are media-savvy individuals – the opinion leaders”

who is on twitter?

<table>
<thead>
<tr>
<th>Communication Type</th>
<th>User Category Examples</th>
<th>User Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass-personal</td>
<td>Celebrities, Bloggers</td>
<td>Barack Obama, Lady Gaga</td>
</tr>
<tr>
<td>Personal</td>
<td>Others (the rest of us)</td>
<td>Jack Ma, Elon Musk</td>
</tr>
</tbody>
</table>
goals of the study

who?
user categories

who listens to whom?
information flow & consumption

who says what?
information production
quick detour: what is the "full" dataset of users?

Q1. all people living in a country?  
Q2. all Twitter user accounts?

A1: exact number unknown  
A2: exact number known only to Twitter

estimates for each case might exist  
(with varying levels of uncertainty)

more often than not, we work with  
partial data, with samples
sampling

assume that $X$ is a random variable with distribution $p(X)$

Monte Carlo: sampling $p(X)$ generates a finite number of samples that can be used to approximate functions of $X$ (e.g. expected value)

a random sample of $X$: $(X_1,\ldots,X_N)$ is representative in this sense

Image source: https://www.slideshare.net/kohta/particle-filter-tracking-in-python
sampling in the social sciences

access to full populations is impossible or impractical

$X$ is a vector of individual attributes: age group, zip code, etc.

how to obtain representative population samples has been studied in depth in the social sciences

non-probabilistic sampling techniques exist, e.g. convenience sampling, known to be non-representative of the population

**bias**: systematic error arising from many factors, including but not limited to the lack of representativeness of the sample
Twitter data samples (up to 2020)

<table>
<thead>
<tr>
<th>Developer</th>
<th>Use cases</th>
<th>Products</th>
<th>Docs</th>
<th>More</th>
<th>Labs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STANDARD</strong>&lt;br&gt;Our free, standard APIs are great for getting started, testing an integration, or validating a concept.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Featuring:&lt;br&gt;Foundational, free APIs&lt;br&gt;Basic query complexity&lt;br&gt;Forum access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PREMIUM</strong>&lt;br&gt;Our premium APIs offer scalable access to Twitter data for those looking to grow, experiment, and innovate.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Featuring:&lt;br&gt;Scalable access to increased data&lt;br&gt;Free sandbox and flexible month-to-month contracts&lt;br&gt;Forum access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ENTERPRISE</strong>&lt;br&gt;Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Featuring:&lt;br&gt;Enterprise-grade APIs&lt;br&gt;Tailored packages and annual contracts&lt;br&gt;Dedicated account managers and technical support</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

fully random sampling: impossible unless you are Twitter or pay for data: Twitter API - Enterprise category

convenience sample: Twitter API - Standard category

https://developer.twitter.com/en/pricing
# Twitter data samples for academic research (today)

<table>
<thead>
<tr>
<th>Academic Research product track</th>
<th>Enterprise API</th>
<th>Curated datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>This specialized track on the new Twitter API grants free access to full-archive search and other v2 endpoints, a higher Tweet volume cap, as well as enhanced features and functionality to get more precise, complete, and unbiased data for analyzing the public conversation. To use this product track, you need to submit your use case through the Academic Research application.</td>
<td>This provides managed, paid services to access advanced reliability and scalability features on the v1.1 Twitter API. While the Academic Research track will serve all research needs, we are still building access elevations in the v2 API. Until this work is complete, the Enterprise API can serve institutions, organizations, and labs that require custom access and Twitter support.</td>
<td>These free, no-code datasets are intended to make it easier for developers and academics to study topics that are of frequent interest to the research community. They are purpose-built, predefined, comprehensive datasets of all public Tweets related to a specific topic.</td>
</tr>
</tbody>
</table>

back to the main topic: datasets

1. follower graph [Kwak et al, WWW 2010] collected Jul 2009, 42M users, 1.5B edges median number of followers < 100 few users have millions of followers

2. twitter firehose (full stream) 223 days (Jul 2009 – Mar 2010) 5B tweets 260M tweets with bit.ly URL links restriction to URLs motivated by easier to track content give access to richer content
lists: feature to groups users

allows to organize users into sets

list names are meaningful labels to describe listed users

→ user categorization

snowball user sample: exploiting lists of popular users

$u_0$: manual seed users (4 categories)

Check lists that seed users belong to & manually select keywords

$l_0$: crawl all lists where seed users appear in

Prune lists to keep those lists that contain keywords

$u_1$: crawl all users in pruned lists

Repeat

Media, Celebrities, Organizations, Blogs

Media (news, news-media), Celebrities (stars, celebs), Organizations (company, ngo, brand), Blogs (blog, blogger)
the concept of elite users: top 5000 users (ranked by how frequently they are listed in each category)

<table>
<thead>
<tr>
<th>category</th>
<th># of users</th>
<th>% of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>celeb</td>
<td>82,770</td>
<td>15.8%</td>
</tr>
<tr>
<td>media</td>
<td>216,010</td>
<td>41.2%</td>
</tr>
<tr>
<td>org</td>
<td>97,853</td>
<td>18.7%</td>
</tr>
<tr>
<td>blog</td>
<td>127,483</td>
<td>24.3%</td>
</tr>
<tr>
<td>total</td>
<td>524,116</td>
<td>100%</td>
</tr>
</tbody>
</table>

top 5 users per category (ranked by #lists in that category)

<table>
<thead>
<tr>
<th>Celebrity</th>
<th>Media</th>
<th>Org</th>
<th>Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>aplusk</td>
<td>cnnbrk</td>
<td>google</td>
<td>mashable</td>
</tr>
<tr>
<td>ladygaga</td>
<td>nytimes</td>
<td>Starbucks</td>
<td>problogger</td>
</tr>
<tr>
<td>TheEllenShow</td>
<td>asahi</td>
<td>twitter</td>
<td>kibeloco</td>
</tr>
<tr>
<td>taylorswift13</td>
<td>BreakingNews</td>
<td>joined</td>
<td>naosalvo</td>
</tr>
<tr>
<td>Oprah</td>
<td>TIME</td>
<td>ollehkt</td>
<td>dooce</td>
</tr>
</tbody>
</table>

counts of URLs initiated by each category composed of 5000 elite users

<table>
<thead>
<tr>
<th>category</th>
<th># of URLs</th>
<th># of URLs per-capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>celeb</td>
<td>139,058</td>
<td>27.81</td>
</tr>
<tr>
<td>media</td>
<td>5,119,739</td>
<td>1023.94</td>
</tr>
<tr>
<td>org</td>
<td>523,698</td>
<td>104.74</td>
</tr>
<tr>
<td>blog</td>
<td>1,360,131</td>
<td>272.03</td>
</tr>
<tr>
<td>ordinary</td>
<td>244,228,364</td>
<td>6.10</td>
</tr>
</tbody>
</table>
elite users: how do they relate to ordinary users?

start with 100K ordinary (non-elite) users

celebrities dominate: users get 25% of their tweets from the top 1000 celebrities

average fraction of tweets for an ordinary user that are accounted for by the top K elite users that the ordinary user follows
who listens to whom?

"Ordinary users receive their information from thousands of distinct sources, many of which are not the media."

“Audiences are increasingly fragmented.”

“Only ~15% of tweets received by ordinary users are received directly from the media”

"20K elite users attract ~50% of all attention"

→ add values for k=5000 for 4 categories

photo credit: John Schnobrich on Unsplash: https://unsplash.com/photos/FIPc9_VocJ4
who listens to whom among the 4 categories?

tweets (with URL) received

Category of Twitter Users

A receive tweets from B

<table>
<thead>
<tr>
<th></th>
<th>Celeb</th>
<th>Media</th>
<th>Org</th>
<th>Blog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celeb</td>
<td>38.27</td>
<td>6.23</td>
<td>1.55</td>
<td>3.98</td>
</tr>
<tr>
<td>Media</td>
<td>3.91</td>
<td>26.22</td>
<td>1.66</td>
<td>5.69</td>
</tr>
<tr>
<td>Org</td>
<td>4.64</td>
<td>6.41</td>
<td>8.05</td>
<td>8.70</td>
</tr>
<tr>
<td>Blog</td>
<td>4.94</td>
<td>3.89</td>
<td>1.58</td>
<td>22.55</td>
</tr>
</tbody>
</table>
**two-step flow of information**

media has an indirect influence over the public via an **intermediate** layer of opinion leaders (Katz 1955)

* figure from authors presentation

information on Twitter passes through intermediaries via
(1) retweets
(2) tweets of URLs
two-step flow of information (2)

for 1M random ordinary users, 46% of received URLs generated by top 5000 media users were received via intermediaries

intermediaries: pass along content to at least one other user
  * 99% are ordinary users, not elite
  * exposed to more media than ordinary users (9100 vs. 1300 URLs)
  * more active (543 vs. 34 followers; 180 vs. 7 tweets)
what to remember

who talks to whom on Twitter

- **fragmented** audiences: no longer ruled by classical media
- **concentrated** attention: 20K elite users get half the attention
- **homophily**: celebrities follow celebrities; media follows media
- **two-step information flow**: half of media URLs pass via intermediaries
questions?

daniel.gatica-perez@epfl.ch