computational social media

lecture 03: tweeting

part 1

daniel gatica-perez

13.04.2018
a human-centric view of twitter

0. introduction
1. twitter users & uses
2. twitter-specific phenomena
3. twitter & large-scale human behavior
4. twitter & real-world events
0. introduction
## Twitter Basic Official Statistics

**Mission**

“to give everyone the power to create and share ideas and information instantly, without barriers”

<table>
<thead>
<tr>
<th>Category</th>
<th>2014</th>
<th>2016</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly active users</td>
<td>241M</td>
<td>320M</td>
<td>N/A</td>
</tr>
<tr>
<td>Tweets sent per day</td>
<td>500M</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Active users on mobile</td>
<td>76%</td>
<td>80%</td>
<td>N/A</td>
</tr>
<tr>
<td>Accounts outside the US</td>
<td>77%</td>
<td>79%</td>
<td>N/A</td>
</tr>
<tr>
<td>Supported languages</td>
<td>35+</td>
<td>35+</td>
<td>N/A</td>
</tr>
<tr>
<td>Employees</td>
<td>2700</td>
<td>3900</td>
<td>N/A</td>
</tr>
</tbody>
</table>
what is twitter?

"an echo chamber of random chatter"

“140 characters: somewhere between an SMS (with larger audience) an email (but less formal) a blog (but less cumbersome)"

280 characters (Nov 2017)

J. van Dijck The culture of connectivity, Oxford University Press, 2013
Jack Dorsey
@jack

just setting up my twttr

https://twitter.com/jack/status/20

https://about.twitter.com/milestones
before twitter...

古池や蛙飛び込む水の音
ふるいけやかわずとびこむみずのおと

old pond... 
a frog leaps in 
water’s sound

The Dinosaur

On waking, the dinosaur was still there.

Augusto Monterroso (20th century)

http://en.wikipedia.org/wiki/Haiku

http://es.wikipedia.org/wiki/Microrrelato
Dr. Enrico Fermi
Institute of Nuclear Studies
University of Chicago
Chicago, Illinois

RESERVATIONS MADE AT RITTENHOUSE CLUB, 1811 WALNUT STREET FOR BOTH EVENINGS.

HENRY B. ALLEN
THE FRANKLIN INSTITUTE
CHAINS REQUIRED
20 MILES AHEAD
EXPECT DELAYS
understanding twitter

1. twitter users & uses
2. twitter-specific phenomena
3. twitter & large-scale human behavior
4. twitter & real-world events
topic 1: understanding twitter users & uses
what is twitter made of?

@username

follow (2006)
users subscribe to other users’ tweets

#hashtag

hashtags # (2007, official 2009)
words articulating a topic or event
allow for search and clustering

Retweet this to your followers?

Join The Flock @JoinTheFlock · 4h
Twitter is hiring! Vertical Marketing Manager - Automotive
jobvite.com/m?39Mfgwe #job

Retweet RT (2007, official 2009)
repost tweets towards one’s followers
enables trends by retweeting

https://about.twitter.com/press/brand-assets
https://about.twitter.com/milestones
J. van Dijck The culture of connectivity, Oxford University Press, 2013
hashtags

- link **strangers** into larger conversations

- facilitate **impromptu** interactions

- not directed communication but a **stream**

- enable the **emergence** of trending topics
Chris Messina
@chrismessina

how do you feel about using # (pound) for groups. As in #barcamp [msg]?

Eric Rice
@ericrice

ReTweet: jmalthus @spin Yes! Web2.0 is about social media, and guess what people like to be social about? Themselves. Social Narcissism

https://twitter.com/chrismessina/status/223115412

https://twitter.com/ericrice
Is Twitter Getting Rid of @-Replies and Hashtags?

Twitter's head of news, Vivian Schiller, momentarily sent Twitter into a self-reflective mini-frenzy when she revealed that the service might be eliminating some of its core features, namely @-replies and hashtags.
geolocalized tweets in Switzerland
who uses twitter?
users and usage

2006: older professional users in business and news

2009: shift to younger adults, then mainstream

from “social network” to “information network”

<table>
<thead>
<tr>
<th>tool for communication</th>
<th>tool for self-promotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>- everyday small talk</td>
<td>- “influentials”</td>
</tr>
<tr>
<td>- (citizen) journalism</td>
<td>- celebrities, stars</td>
</tr>
<tr>
<td>- political grassroots activism</td>
<td>- politicians</td>
</tr>
<tr>
<td>- emergencies and disasters</td>
<td>- enables organization/management</td>
</tr>
<tr>
<td>- community participation</td>
<td>of fans/audiences/voters</td>
</tr>
<tr>
<td>- fake news</td>
<td></td>
</tr>
</tbody>
</table>

“the impulse to make life a publicly annotated experience has blurred the distinction between advertising and self-expression, marketing and identity” (Hagan 2011)

J. van Dijck The culture of connectivity, Oxford University Press, 2013
Twitter basic descriptive statistics (2009)
41.7M user profiles

CCDF: Complementary Cumulative Distribution Function
$CCDF(x) = 1 - F(x) = P(X > x)$

following vs. friending

“connection with very low expectation” (Murthy, 2013)

low reciprocity, highly asymmetric links (Kwak, 2010)

“77.9% of user pairs with any link between them are connected one-way, and only 22.1% have reciprocal relationship between them.”

“67.6% of users are not followed by any of their followings. We conjecture that for these users Twitter is rather a source of information than a social networking site”

weak and “strong” ties in twitter

**weak tie** (following): easy to follow a large number of people

**“strong” tie** (“friend”): direct communication ‘@’ (at least twice) in the observation period

309k users
211k users posted at least twice
206 days (on average)

Figure 4: Number of friends as a function of the number of followees. The total number of friends saturates while the number of followees keeps growing due to the minimal effort required to add a followee.

topic 2: studying twitter-specific phenomena
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"


photo credit: United Workers (cc) https://www.flickr.com/photos/unitedworkers/14138566864
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 media</td>
<td>Media, Organizations</td>
<td>CNN, The New York Times</td>
</tr>
<tr>
<td>Mass-personal</td>
<td>Celebrities, Bloggers</td>
<td>WWF, Whole Foods</td>
</tr>
<tr>
<td>Personal</td>
<td>Others (the rest of us)</td>
<td></td>
</tr>
</tbody>
</table>
goals of the study

who?
user classification

who listens to whom?
information flow & consumption

who says what?
information production

credit: mark hillary (cc): http://www.flickr.com/photos/markhillary/5412943529/
data

1. follower graph [Kwak et al, WWW 2010] collected July 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
twitter lists:
feature that groups sets of users into categories

useful to organize users into sets

user annotation & categorization

list names are meaningful labels to describe the listed users
user crawling (1): snowball sampling of twitter lists of elite users: media, celebrities, organizations, blogs

- Manual seed users (4 categories)
- Check lists, then manually select keywords
- Crawl all lists where seeds appear
- Prune lists to keep only those that contain keywords
- Crawl all users in pruned lists

Media (news, news-media), Celebrities (stars, celebs), Organizations (company, ngo, brand), Blogs (blog, blogger)
user crawling (2): activity-based sampling

Crawl all lists for all users that tweeted at least once a week

keyword-pruned lists

get users

<table>
<thead>
<tr>
<th>category</th>
<th>Snowball Sample</th>
<th></th>
<th>Activity Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of users</td>
<td>% of users</td>
<td># of users</td>
</tr>
<tr>
<td>celeb</td>
<td>82,770</td>
<td>15.8%</td>
<td>14,778</td>
</tr>
<tr>
<td>media</td>
<td>216,010</td>
<td>41.2%</td>
<td>40,186</td>
</tr>
<tr>
<td>org</td>
<td>97,853</td>
<td>18.7%</td>
<td>14,891</td>
</tr>
<tr>
<td>blog</td>
<td>127,483</td>
<td>24.3%</td>
<td>43,830</td>
</tr>
<tr>
<td>total</td>
<td>524,116</td>
<td>100%</td>
<td>113,685</td>
</tr>
</tbody>
</table>
elite users: top 5000 users
(ranked by how frequently they are listed in each 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>joinred</td>
<td>problogger</td>
</tr>
<tr>
<td>Oprah</td>
<td>TIME</td>
<td>ollehkt</td>
<td>naosalvo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>dooce</td>
</tr>
</tbody>
</table>

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

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

measure proportion of elite accounts that ordinary users follow or receive tweets from for 4 categories of interest

snowball vs. activity: similar distributions

celebrities dominate (users get 25% of their tweets from the top 1000 celebrities, and 20% of the users they follow are top-1000 celebrities)

- how many others a user follows
- tweets produced by the accounts a user follows

average fraction of friends or tweets for a random user that are accounted for by top K elite users.

(a) Snowball sample

(b) Activity sample
who listens to whom?

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

"Audiences have become increasingly fragmented."

“Only about 15% of tweets received by ordinary users are received directly from the media"

"20K elite users (<0.05% of all users) attract ~50% of all attention… information flows have not become egalitarian."

5000 per category, add values for k=5000 in Figure in last slide

credit (cc): institute of network cultures
who listens to whom among the 4 categories?

Category of Twitter Users

A → B receive tweets from A

<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>

tweets (with URL) received

re-tweets

Category of Twitter Users

A retweet 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>4,334</td>
<td>1,489</td>
<td>1,543</td>
<td>5,039</td>
</tr>
<tr>
<td>Media</td>
<td>4,624</td>
<td>40,263</td>
<td>7,628</td>
<td>32,027</td>
</tr>
<tr>
<td>Org</td>
<td>1,570</td>
<td>2,539</td>
<td>18,937</td>
<td>11,175</td>
</tr>
<tr>
<td>Blog</td>
<td>3,710</td>
<td>6,382</td>
<td>5,762</td>
<td>99,818</td>
</tr>
</tbody>
</table>
two-step flow of information

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

information passes through intermediaries via
(1) retweets
(2) tweets of URLs

For 1M random ordinary users, 46% of received URLs generated by top 5000 media users were received via intermediaries (2-step flow)
two-step flow of information (2)

indirect flow ratio: % of media URLs received via intermediaries

two patterns for ordinary users
* users receiving up to 100 media URLs, receive them essentially all through intermediaries
* others receive them virtually all via the media
two-step flow (3): what about the intermediaries?

Intermediaries (pass along media content to at least one other user):  
* 99% are ordinary users, not elite  
* they are exposed to more media content than random ordinary users (9165 vs. 1377 URLs)  
* they are more active (543 vs. 34 followers; 180 vs. 7 tweets)

--> prominent intermediaries are a small fraction
URLs lifespan by category

URL lifespan
- time lag between first & last appearance of a URL on Twitter

- Media generated a large proportion of short lived URLs

- URLs from bloggers are better represented among long-lived ones
in summary

1. “audience attention has fragmented among a wider pool of content producers than classical mass media”

2. “yet attention remains highly concentrated – 20K elite users attract ~50% of all attention for the sample”

3. “attention & homophily – media follows media; celebrities follow celebrities”

4. “support for two-step flow of information model – almost half of the media URLs pass indirectly via intermediaries”

5. “different content has different lifespans – media-originated URLs are short-lived; blogger-originated URLs are more long-lived.”
announcements: paper #3
to be presented 19.04.2018

S. Vosoughi, D. Roy, S. Aral,
The spread of true and false news online,
Science, 359, pp. 1146–1151, March 2018
http://science.sciencemag.org/content/359/6380/1146

we need a presenter, a discussant, and a scribe

let me know by today 5pm if you want to volunteer (first come first served). if nobody volunteers, I will pick 3 people randomly
reminder of logistics

role 1: presenter

read paper & prepare slides: 15 minute presentation

points: (1) what problem is addressed? (2) what are the contributions? (3) what is the technical approach? (4) what are the main findings?

role 2: discussant

read paper & prepare questions to guide discussion

examples: (1) how to improve / extend the work? (2) what implications do the ideas have in society or in computing? (3) are there any controversial issues? (4) any limitations?

role 3: scribe

take notes of the group discussion for your paper

write up summary of discussion (one page max)
questions?

gatica@idiap.ch
daniel.gatica-perez@epfl.ch
@dgaticaperez