announcements

- reading #2 will be presented today

M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, PNAS, 2013
This lecture

A human-centric view of Twitter

1. Introduction
2. Twitter users & uses
3. Twitter-specific phenomena
4. Twitter, large-scale human behavior & real-world events
1: 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></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>
"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
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

Bashõ (17th century)

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
Charge to the account of

THE FRANKLIN INSTITUTE

Western Union

Charge to the account of

THE FRANKLIN INSTITUTE

Send the following telegram, subject to the terms on back hereof, which are hereby agreed to

WANT A REPLY?
"Answer by WESTERN UNION" or similar phrases may be included without charge:

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
2: understanding twitter users & uses
what is twitter made of?

follow (2006)
users subscribe to other users’ tweets

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

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

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

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/ericrice
source: The Evolution of Languages on Twitter, Brian Lehman, 10.03.2014, http://blog.gnip.com/twitter-language-visualization/
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”

**tool for communication**
- everyday small talk
- (citizen) journalism
- political grassroots activism
- emergencies and disasters
- community participation
- fake news

**tool for self-promotion**
- “influentials”
  - celebrities, stars
  - politicians
- enables organization/management of fans/audiences/voters

“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

\[ x = 20: \text{twitter recommended new users to follow 20 users to start with} \]

\[ x = 2000: \text{before 2009 there was limit on the number of people one could follow (rule removed later)} \]

CCDF: Complementary Cumulative Distribution Function
\[ \text{CCDF}(x) = 1 - F(x) = P(X>x) \]

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

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.

3: studying twitter-specific phenomena
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"


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</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>
</tbody>
</table>
| Personal           | Others (the rest of us) | }

---

*CNN, The New York Times, WWF, Whole Foods*
goals of the study

who?  
user categorization

who listens to whom?  
information flow & consumption

who says what?  
information production
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
How to obtain a set of Twitter users?
Use Lists (feature that groups 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)
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>Snowball Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of users</td>
</tr>
<tr>
<td>celeb</td>
<td>82,770</td>
</tr>
<tr>
<td>media</td>
<td>216,010</td>
</tr>
<tr>
<td>org</td>
<td>97,853</td>
</tr>
<tr>
<td>blog</td>
<td>127,483</td>
</tr>
<tr>
<td>total</td>
<td>524,116</td>
</tr>
</tbody>
</table>

Statistics of the snowball sample

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>Starbucks</td>
<td>kibeloco</td>
</tr>
<tr>
<td>taylorswift13</td>
<td>BreakingNews</td>
<td>Starbucks</td>
<td>naosalvo</td>
</tr>
<tr>
<td>Oprah</td>
<td>TIME</td>
<td>twitter</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

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

celebrities dominate (users get 25% of their tweets from the top 1000 celebrities, and 20% of the users they follow are top-1000 celebrities)
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 attract ~50% of all attention… information flows have not become egalitarian."

(5000 per category, add values for k=5000 in Fig. in last slide)

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

tweets (with URL) received

re-tweets
two-step flow of information

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

![Diagram of two-step flow of information]

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)

ordinary users: two patterns
* users receiving up to 100 media URLs, receive them essentially all through intermediaries
* others receive them most of them via the media

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

Twitter as an information network

Audience attention is fragmented, not longer dominated by classical mass media

Attention is concentrated: 20K elite users get half of the attention (early Twitter days)

Homophily is present: celebrities follow celebrities & media follows media

Two-step flow of information: about half of media URLs pass indirectly through intermediaries
questions?
daniel.gatica-perez@epfl.ch