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

projects

please contact me about your HREC submission if you haven't done it yet

this lecture

a human-centric view of twitter

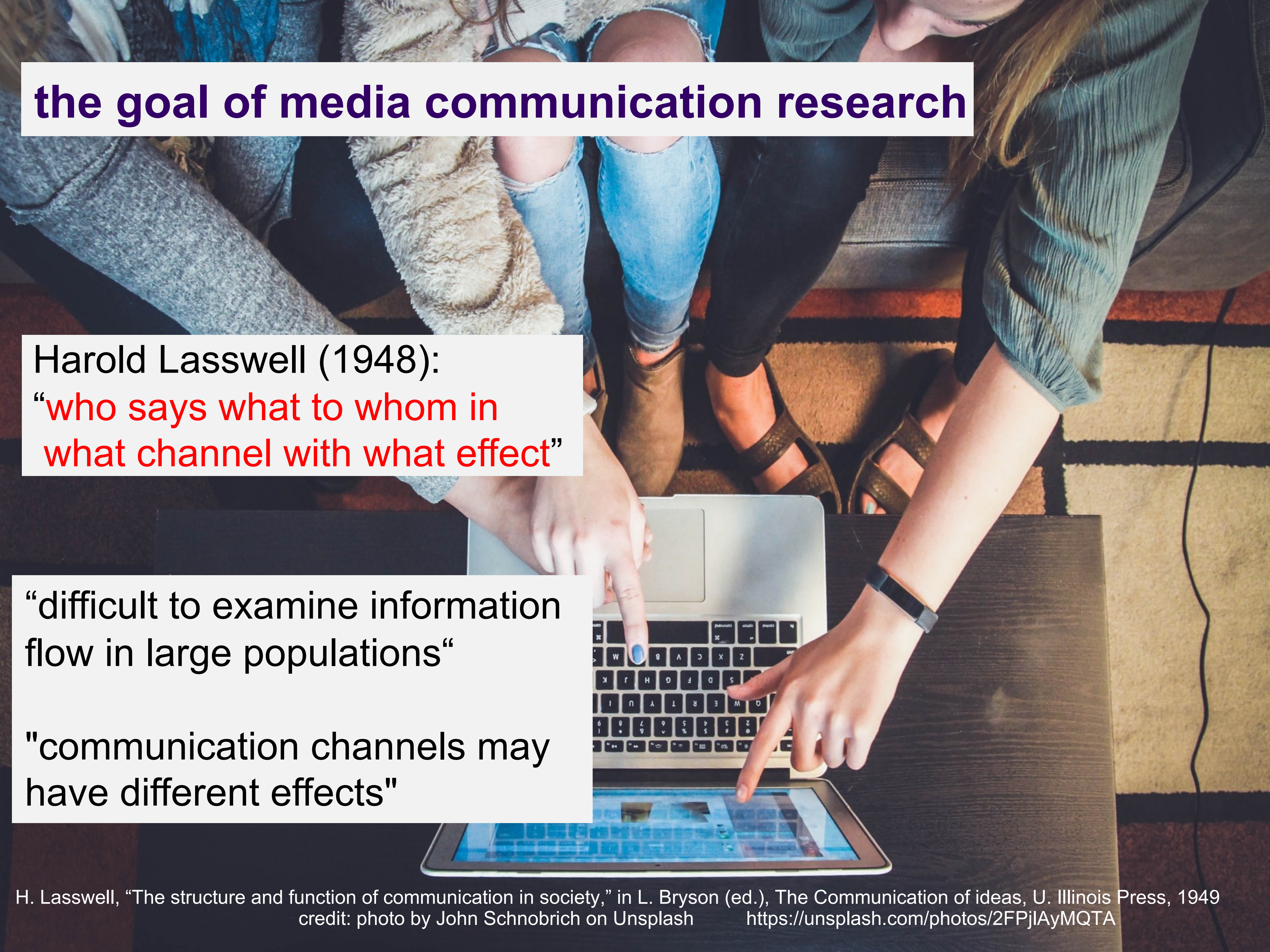
1. introduction
2. twitter users & uses
3. understanding large-scale human behavior
4. inferring real-world events & trends
5. spreading information in the real world

spreading information in the real world

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

1. who talks to whom on Twitter

S. Wu, J. M. Hofman, W. Mason, and D. Watts, “Who Says What to Whom on Twitter,” in Proc. WWW 2011.
Thanks to A. Olteanu for some of the slides.



the goal of media communication research

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"

H. Lasswell, “The structure and function of communication in society,” in L. Bryson (ed.), The Communication of ideas, U. Illinois Press, 1949
credit: photo by John Schnobrich on Unsplash https://unsplash.com/photos/2FPjlAyMQTA

three models of communication

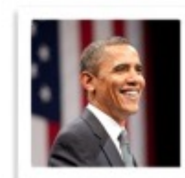
mass communication:
"one-way message transmission from one source to a large, relatively undifferentiated and anonymous audience"

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"

J. B. Walther, C. T. Carr, S. S. W. Choi, D. C. DeAndrea, J. Kim, S. T. Tong, and B. Van Der Heide. Interaction of interpersonal, peer, and media influence sources online. In Z. Papacharissi, (ed.) A Networked Self: Identity, Community, and Culture on Social Network Sites, Routledge, 2010.

who is on twitter?

Communication Type	User Category Examples	User Examples
Mass media	Media, Organizations	
Mass-personal	Celebrities, Bloggers	 
Personal	Others (the rest of us)	

questions

who talks on Twitter?

user categories

who listens to whom?

information flow & consumption

credit: photo by Marten Bjork on Unsplash:
https://unsplash.com/photos/FVtG38Cjc_k

quick detour: what is the "full" dataset of users?

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

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

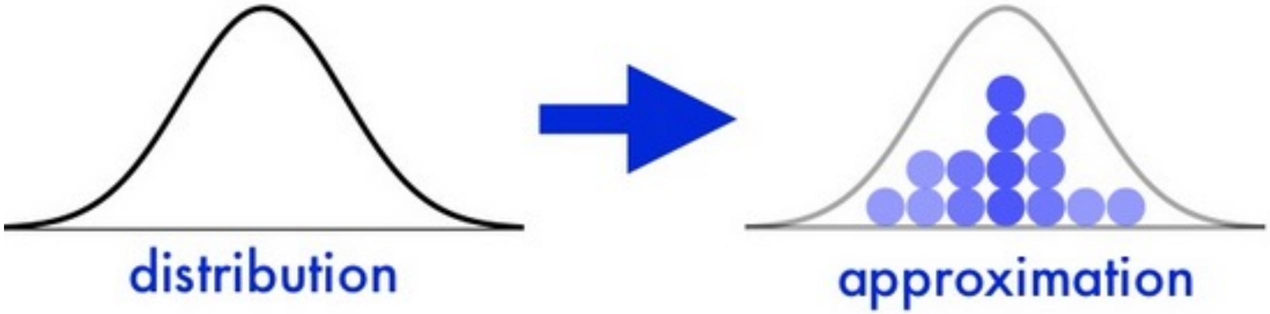
estimates for each case might exist
(with varying levels of uncertainty)

more often than not, we work with
partial data a.k.a. <u>samples</u>

sampling

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

Monte Carlo: sampling p(X) provides 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 for research (up to 2021)

Developer | Use cases | Products | Docs | More | Labs

STANDARD

Our free, standard APIs are great for getting started, testing an integration, or validating a concept.

Featuring:

Foundational, free APIs
Basic query complexity
Forum access

PREMIUM

Our premium APIs offer scalable access to Twitter data for those looking to grow, experiment, and innovate.

Featuring:

Scalable access to increased data
Free sandbox and flexible month-to-month contracts
Forum access

ENTERPRISE

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.

Featuring:

Enterprise-grade APIs
Tailored packages and annual contracts
Dedicated account managers and technical support

fully random sampling: impossible unless you were Twitter or paid for data: Twitter API - Enterprise category

convenience sample: Twitter API - Standard category

https://developer.twitter.com/en/pricing (link no longer active)

Twitter data samples for academic research (since Nov 2021)

Enhance your academic research with global, real-time and historical data

Get more precise, complete, and unbiased data from the public conversation for free. This specialized access includes access to all Twitter API v2 endpoints, a higher monthly **Tweet cap**, and enhanced features designed to support research.

Who it's for

Academic researchers with specific research objectives are encouraged to apply. This includes graduate students working on a thesis, PhD candidates working on a dissertation, or research scholars affiliated with or employed by an academic institution.

Use cases

Find global data for your thesis or dissertation, gather historical or real-time data for your research lab, and study the public conversation with the API v2.

Non-commercial use

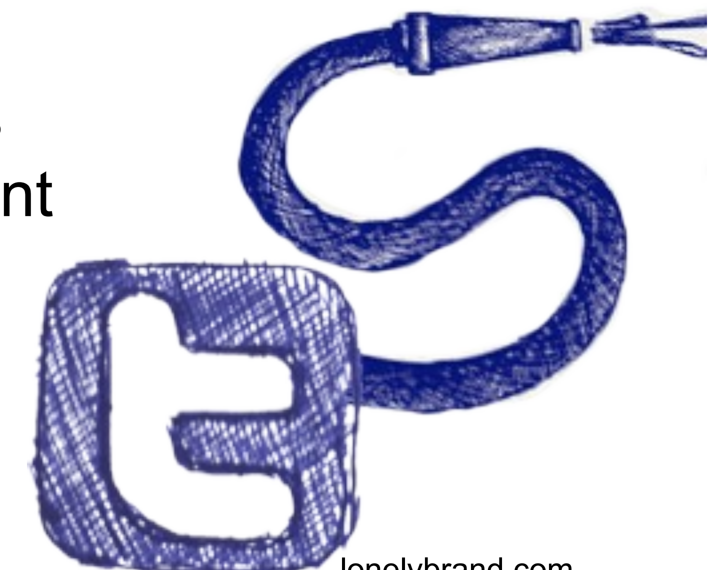
Reserving this access for only non-commercial use makes it possible to provide long-term support for researchers who rely on the Twitter API to do their work.

https://developer.twitter.com/en/products/twitter-api/academic-research

back to the main topic:
datasets

1. follower graph [Kwak et al, WWW 2010]
   collected in Jul 2009, 42M users, 1.5B edges

2. twitter firehose (full stream)
   223 days (Jul 2009 – Mar 2010)
   5B tweets
   260M tweets with bit.ly URL links
      URLs are easier to track content
      & give access to rich content

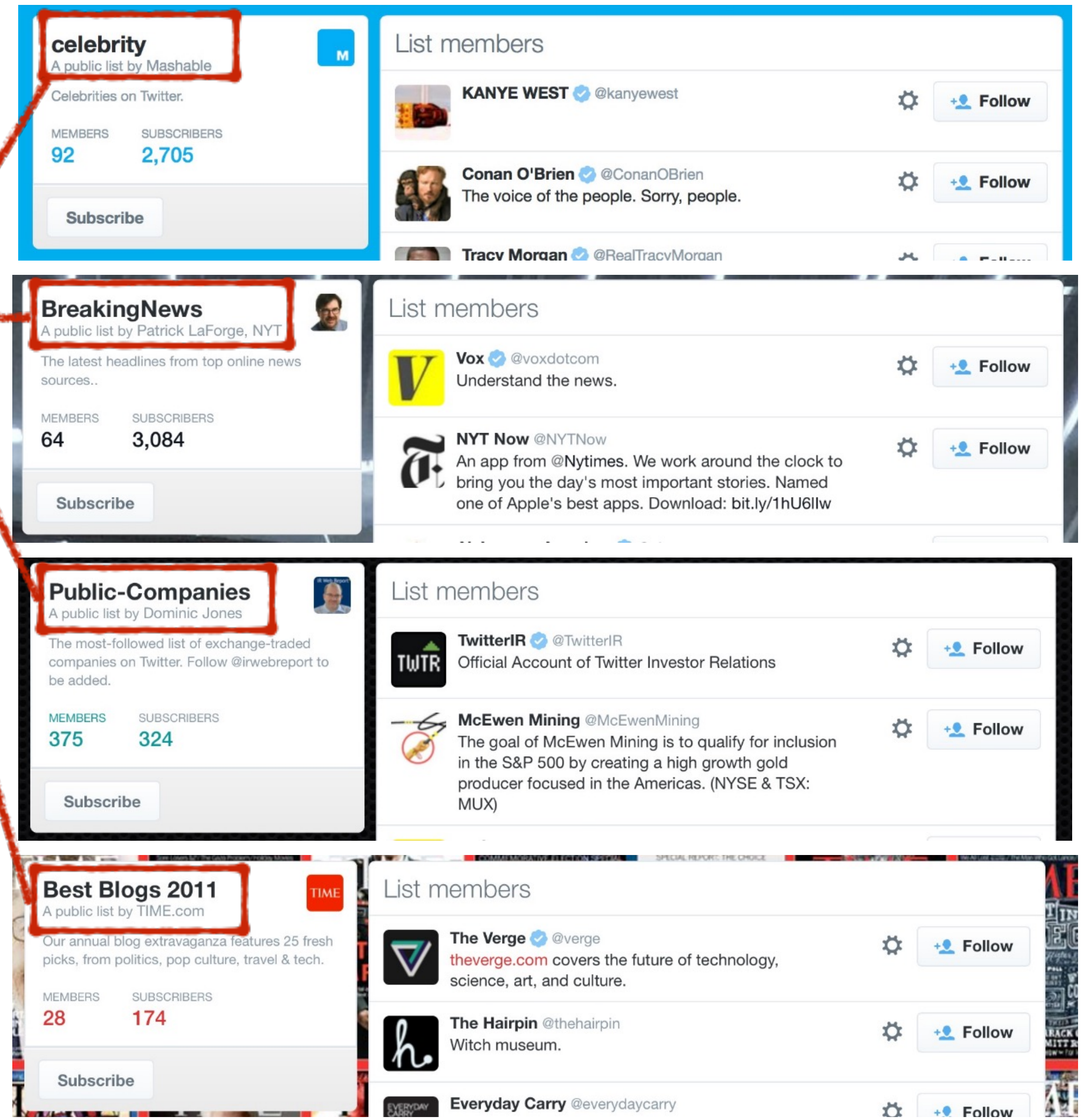


lonelybrand.com

lists: feature to groups users

lists allow to organize users into sets

list names are meaningful labels to describe listed users
→ user categorization



https://help.twitter.com/en/using-twitter/twitter-lists

snowball user sample: using lists of popular users

Media Celebrities Organizations Blogs

u_0: manual **seed users** (4 categories)

check all lists that seed users belong to & manually select keywords

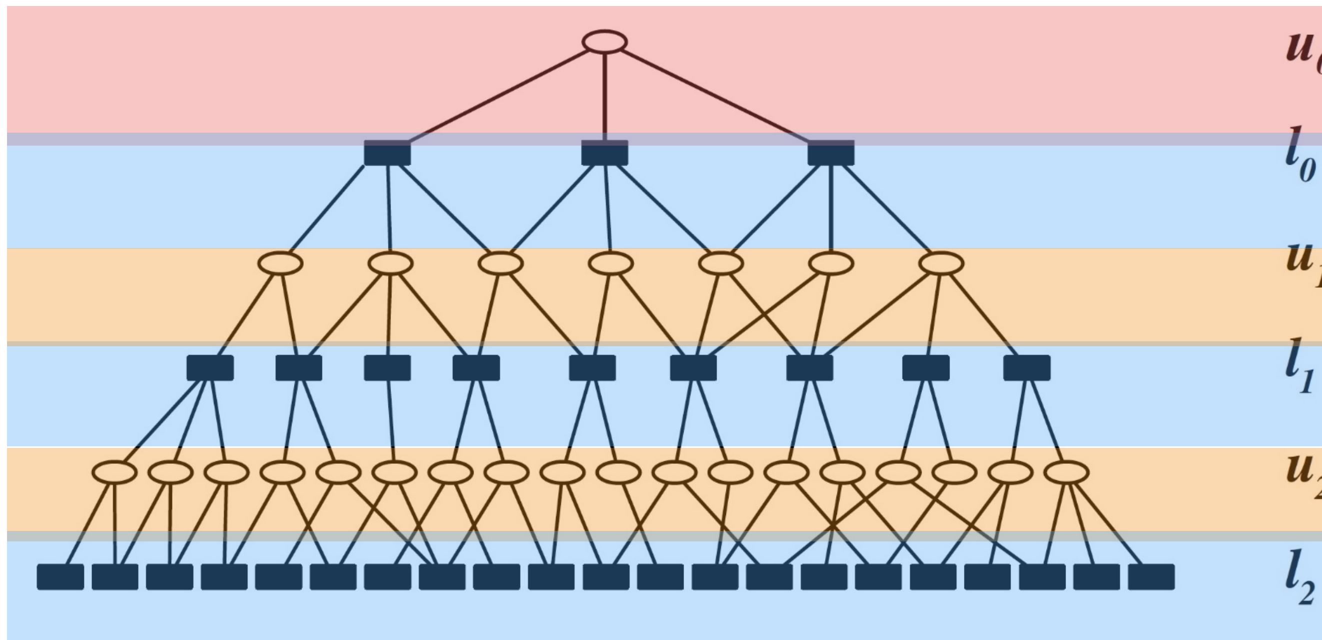
Media (news, news-media), Celebrities (stars, celebs)
Organizations (company, ngo, brand), Blogs (blog, blogger)

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



the concept of elite users: top 5000 users (ranked by how frequently they are listed in each category)

statistics of the snowball sample

category	Snowball Sample	
	# of users	% of users
celeb	82,770	15.8%
media	216,010	41.2%
org	97,853	18.7%
blog	127,483	24.3%
total	524,116	100%

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

Celebrity	*Media*	*Org*	*Blog*
aplusk	cnnbrk	google	mashable
ladygaga	nytimes	Starbucks	problogger
TheEllenShow	asahi	twitter	kibeloco
taylorswift13	BreakingNews	joinred	naosalvo
Oprah	TIME	ollehkt	dooce

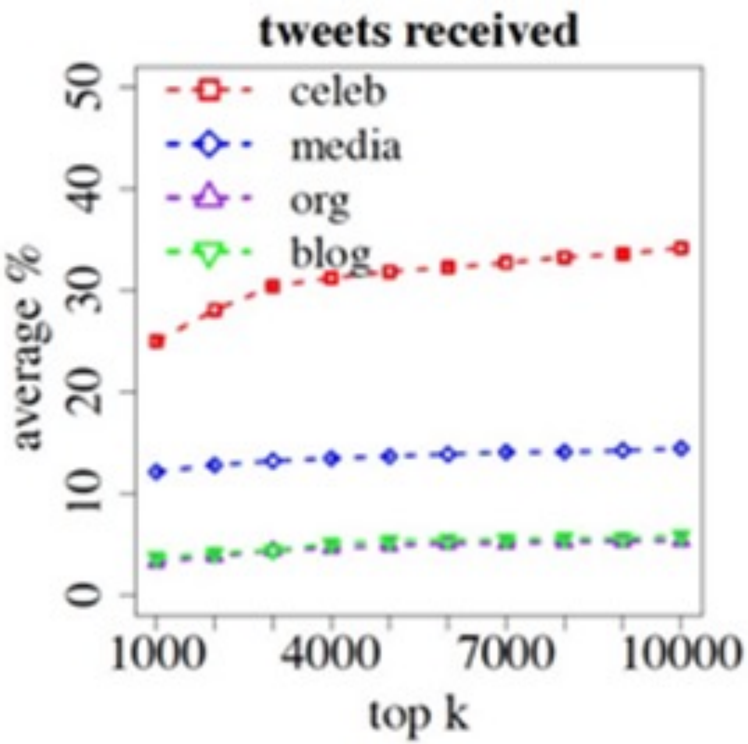
counts of URLs initiated by each category composed of 5000 elite users

category	# of URLs	# of URLs per-capita
celeb	139,058	27.81
media	5,119,739	1023.94
org	523,698	104.74
blog	1,360,131	272.03
ordinary	244,228,364	6.10

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

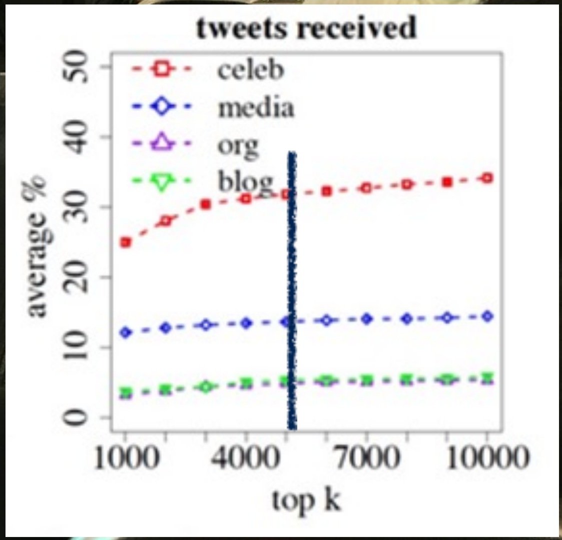
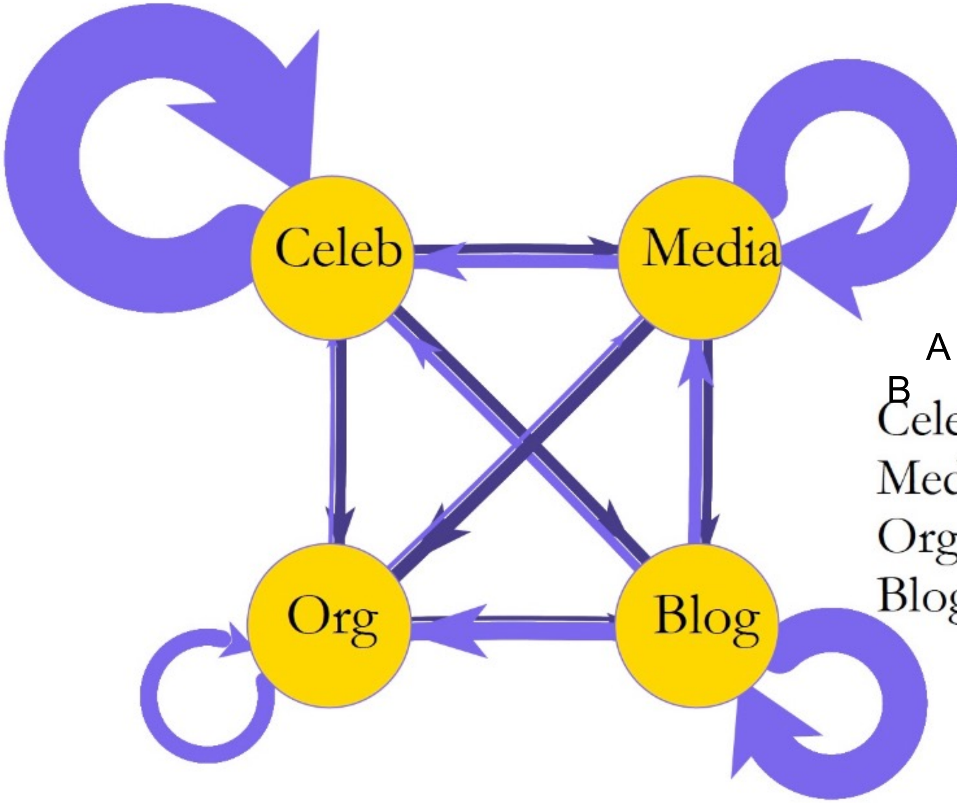
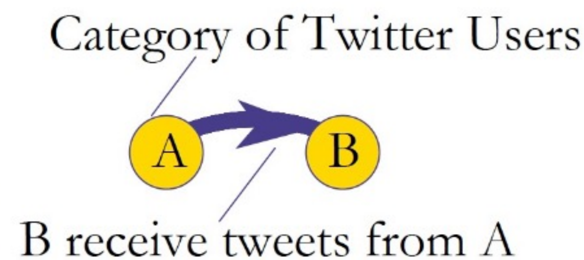


photo credit: John Schnobrich on Unsplash: https://unsplash.com/photos/FIPc9_VocJ4

who listens to whom among the 4 categories?



tweets (with URL) received

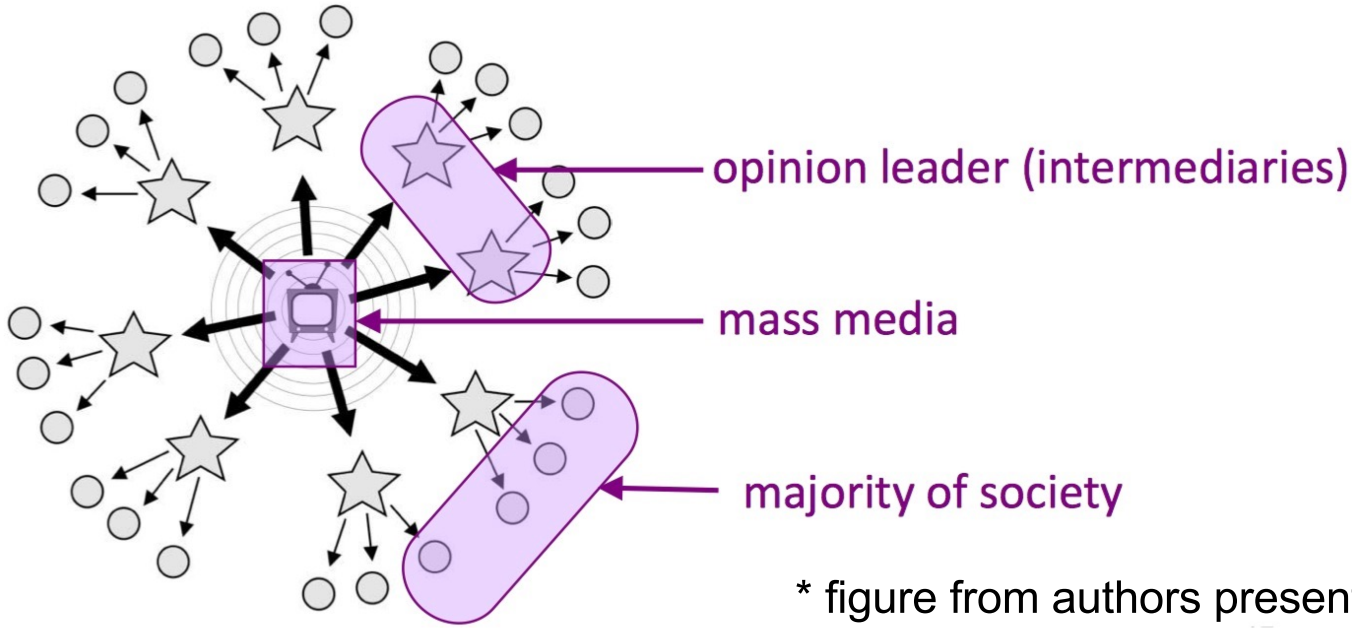


% of tweets received from

B \ A	Celeb	Media	Org	Blog
Celeb	38.27	6.23	1.55	3.98
Media	3.91	26.22	1.66	5.69
Org	4.64	6.41	8.05	8.70
Blog	4.94	3.89	1.58	22.55

two-step flow of information

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



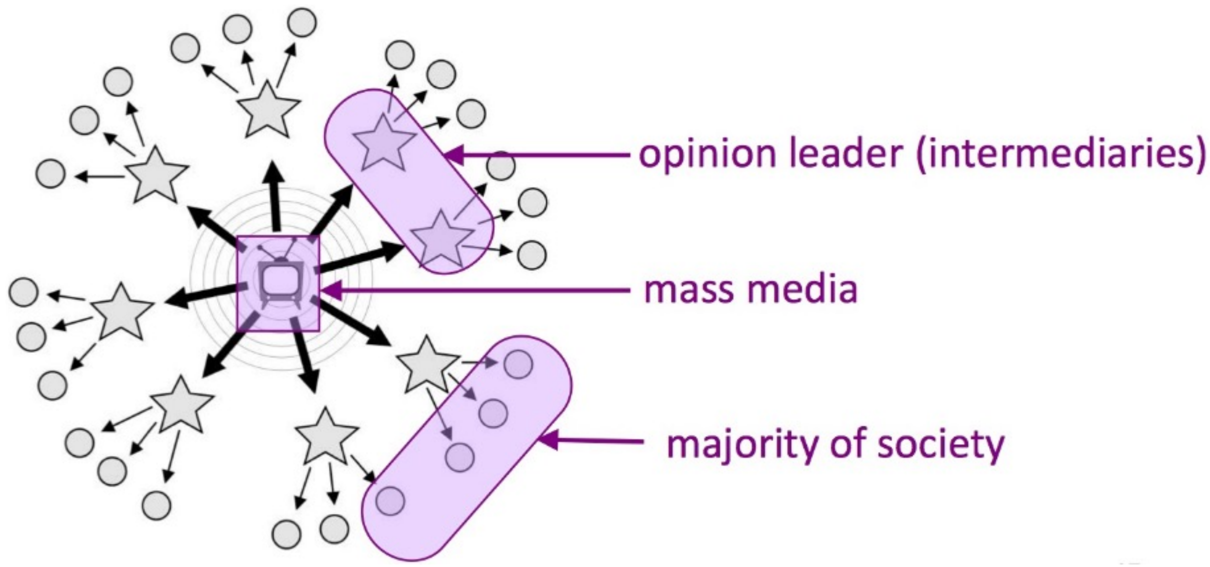
* figure from authors presentation

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

sampling

critical issue for computational social science

who talks to whom on Twitter

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

questions?

daniel.gatica-perez@epfl.ch