# Who says what to whom on Twitter
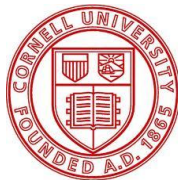
Shaomei Wu

sw475@cornell.edu

Information Science, Cornell University

Jake M. Hofman, Winter A. Mason, Duncan J. Watts

{hofman, winteram, djw}@yahoo-inc.com

Yahoo ! Research New York

WWW 2011, Hyderabad, India

# Motivation

- Lasswell's maxim (1948)
  - "Who says what to whom in what channel with what effect"
    - Hard to observe information flow in large population
    - Different channels have different attributes and effects

# Twitter: a new platform for studying the pattern of communications

- Advantages
  - Represents the full spectrum of communications
    - *Mass media*: CNN, NYTimes, organizations, governments
    - *"Masspersonal"*: celebrities, bloggers, journalists, experts
    - *Interpersonal*: friends and acquaintances
  - Enables easy tracking of information flow
    - URL shortening services (e.g. bit.ly, tinyurl)
- Limitations
  - Twitter is merely one communication channel
  - Hard to observe the "real" effect (e.g., behavior change, opinion forming)

# Data

- ## Twitter Firehose Corpus
  - 223 days (7/28/2009 – 3/8/2010)
  - 5B tweets, 260M (~5%) containing bit.ly URLs

- ## Follower graph (Kwak et al 2010)
  - Twitter as observed by 7/31/2009
  - 42M users, 1.5B following relationships

- Who is whom? (user classification)
- Who listens to whom?
- Who says what?

# Who is whom on Twitter

*mass media*
(Katz and Lazarsfeld 1955)
(Gitlin 1978)

*"masspersonal"*
(Walther et al 2010)

*interpersonal*

**Media**

**Organizations**

**Celebrities**

**Bloggers**

**Other**

CNN

yahoo

ladygaga

mashable

# Twitter Lists as Folksonomy of users

- Twitter Lists: Feature launched on 11/2/2009
- Use the name of a list as a tag of users it contains
- Very time-consuming to crawl all lists



Twitter List Examples

# Snowball sample of Twitter lists (I)

## Manually selected seeds

- **Media**: CNN, New York Times
- **Organizations**: Amnesty International, WWF, Yahoo! Inc, Whole Foods
- **Celebrities**: Barak Obama, Lady Gaga, Paris Hilton
- **Blogs**: BoingBoing, mashable, Chrisbrogan, Gizmodo, …
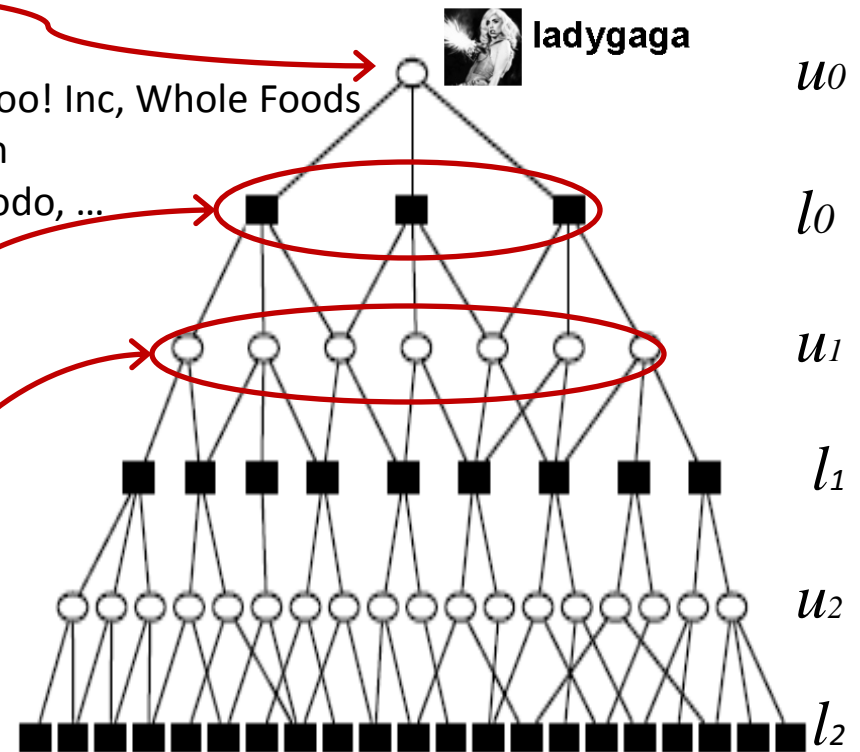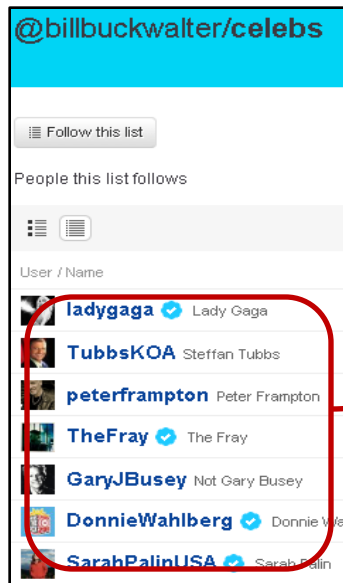
# Snowball sample of Twitter lists

**Manually selected seeds**
- **Media**: CNN, New York Times
- **Celebrities**: Barak Obama, Lady Gaga, Paris Hilton
- **Organizations**: Amnesty International, WWF, Yahoo! Inc, Whole Foods
- **Blogs**: BoingBoing, mashable, Chrisbrogan, Gizmodo, …

**Keyword-pruned lists**
- **Media**: news, media, news-media
- **Celebrities**: star, stars, hollywood, celebs, celebrity, …
- **Organizations**: company, companies, organization, organizations, organisations, corporation, brands, products, ngo, charity, …
- **Blogs**: blog, blogs, blogger, bloggers

$u_0$

$l_0$

$u_1$

$l_1$

$u_2$

$l_2$
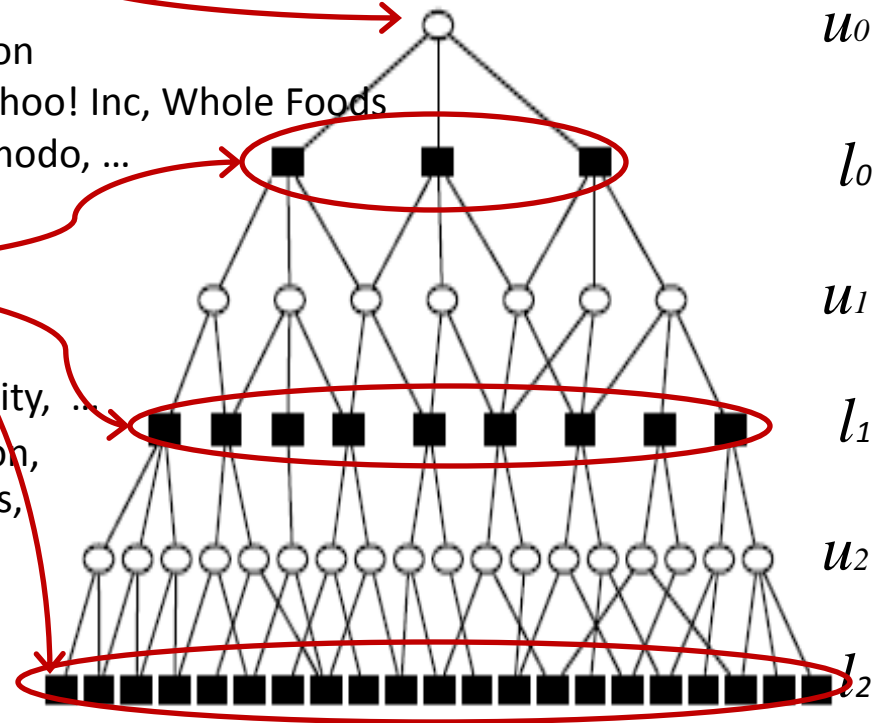
**Resolve ambiguity** (e.g. Oprah Winfrey in both "celebrity" and "media")
- Define *membership score*: $w_{ic} = n_{ic}/N_c$ ( $n_{ic}$ - # of lists in category c that contain user i, $N_c$ – total # of lists in category c )
- Assign user i to the category with highest membership score

9

# Activity sample Twitter lists

all users who tweeted **at least once every week**
    during entire observation period (750K users)

Keyword-pruned lists

- Total 5M lists, 113,685 after pruning

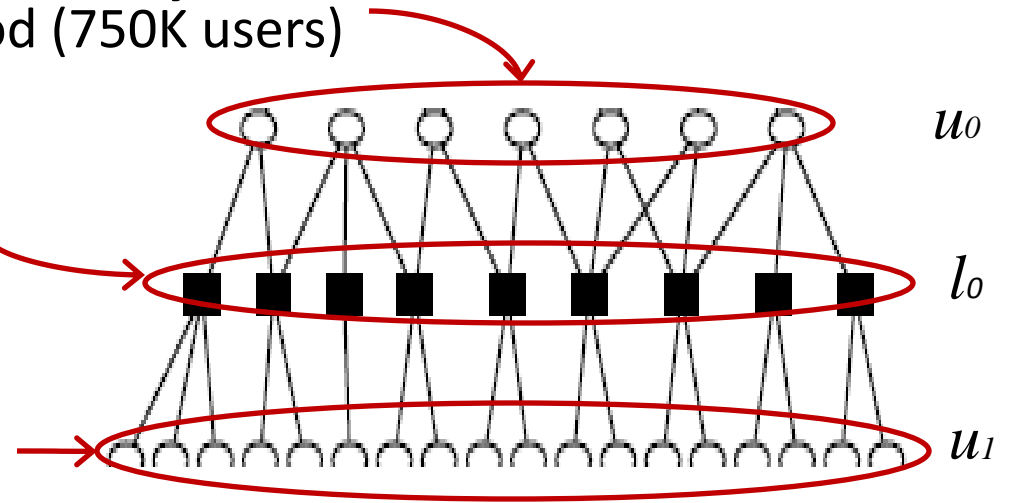**85%** also appear in snow-ball sample

$u_0$

$l_0$

$u_1$

Table 1: Distribution of users over categories

| category | Snowball Sample | | Activity Sample | |
|---|---|---|---|---|
| | # of users | % of users | # of users | % of users |
| celeb | 82,770 | 15.8% | 14,778 | 13.0% |
| media | 216,010 | 41.2% | 40,186 | 35.3% |
| org | 97,853 | 18.7% | 14,891 | 13.1% |
| blog | 127,483 | 24.3% | 43,830 | 38.6% |
| total | 524,116 | 100% | 113,685 | 100% |

# Identify "Elite" Users

- Rank users by the frequency of being listed in each category

Table 3: Top 5 users in each 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 |

- Take the top *k* users in each category as "elite" users
- Leave all the rest as "ordinary" users

*How to set the cutoff value k?*

- For each value of $k$ measure the prominence of each category
  - randomly sample 100K ordinary (i.e. unclassified) users, calculate:
    - % of accounts they follow among the top k users
    - % of tweets they receive from the top k users
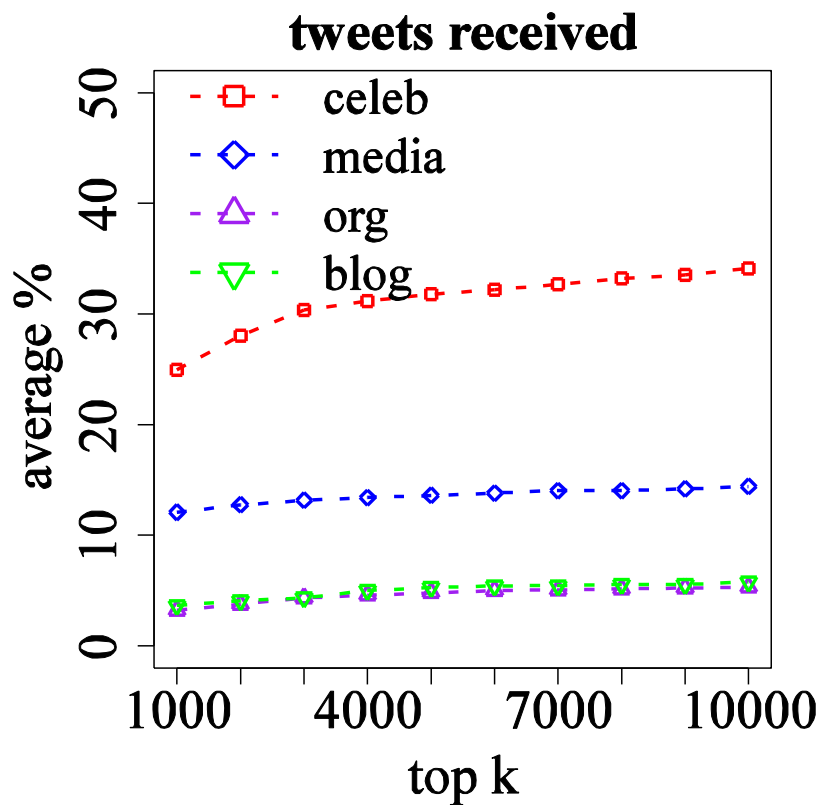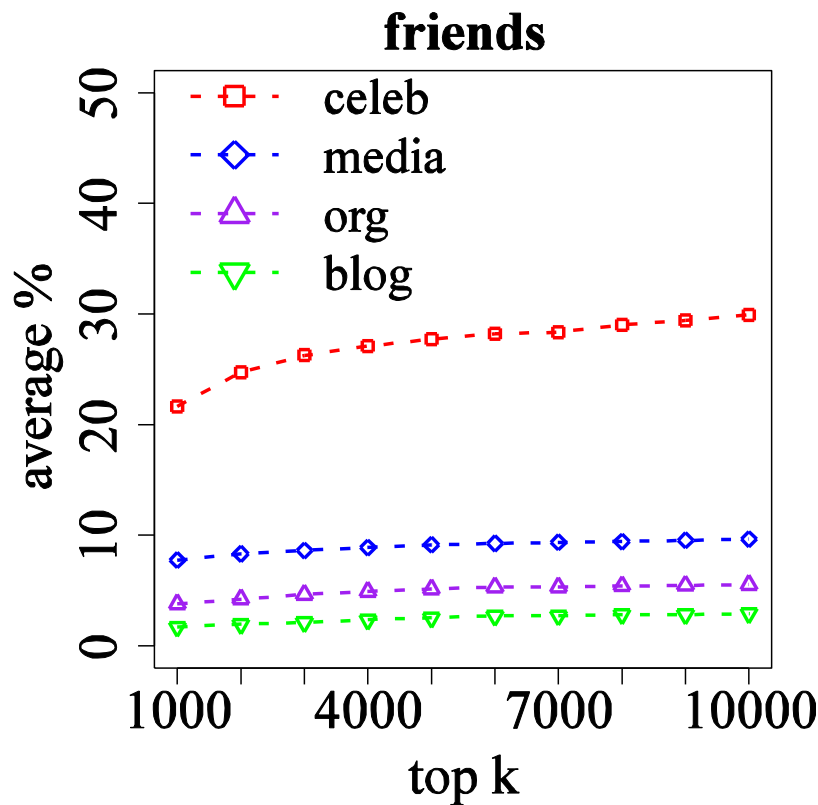
- For each value of $k$ measure the prominence of each category
  - randomly sample 100K ordinary (i.e. unclassified) users, calculate:
    - % of accounts they follow among the top k users
    - % of tweets they receive from the top k users

*High concentration of attention on a small set of "elite" users:*
- *~30% tweets from celebs*
- *~15% from media*
- *~5% from orgs and blogs*

**tweets received**
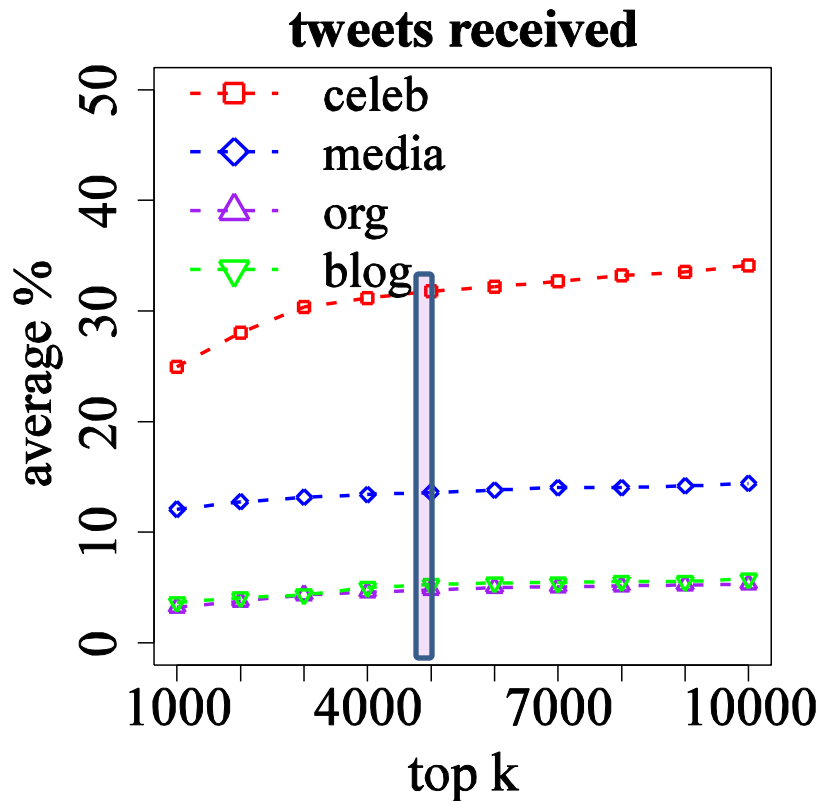
- Who is whom? (user classification)
- Who listens to whom?
- Who says what?

# "Who listens to whom"

- Concentrated attention
  - 20K (0.05%) elite users account for 50% of all attention within Twitter

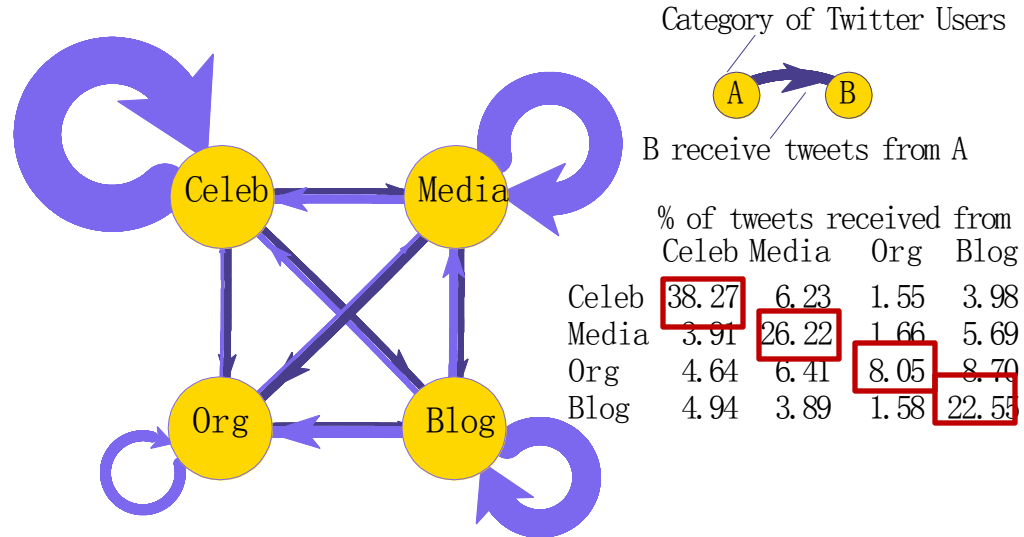  $\Rightarrow$ How do elite users listen to each other?

- Fragmented audience
  - Ordinary users receive information from thousands of distinct sources
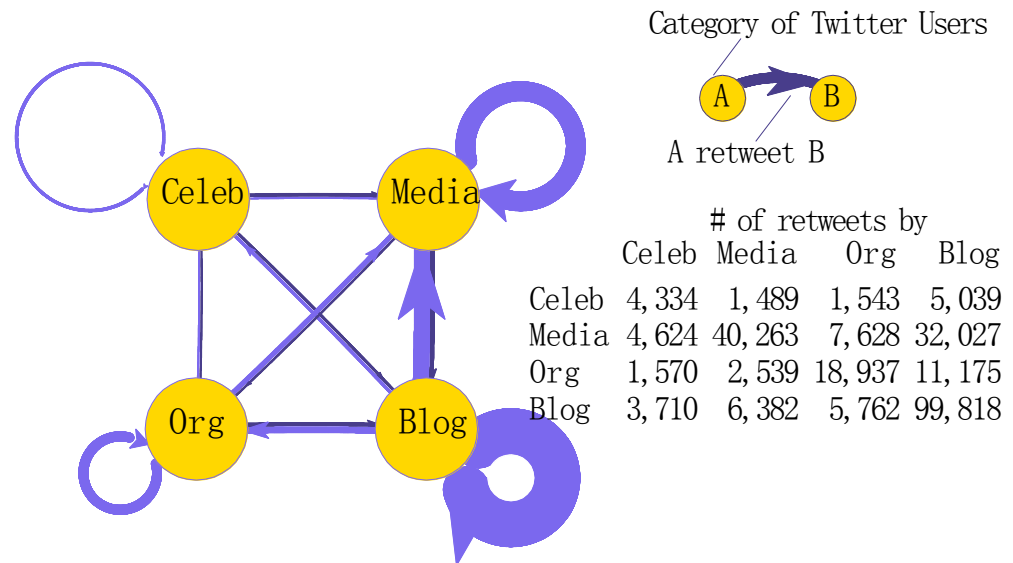  - Only 15% of information ordinary users receive directly from the media

  $\Rightarrow$ How does information flow from the media to the masses?

# How elite categories listen to each other

**tweets received**



Category of Twitter Users

A → B

B receive tweets from A

% of tweets received from

|       | 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 |

**RT behavior**



Category of Twitter Users

A → B

A retweet B

# of retweets by

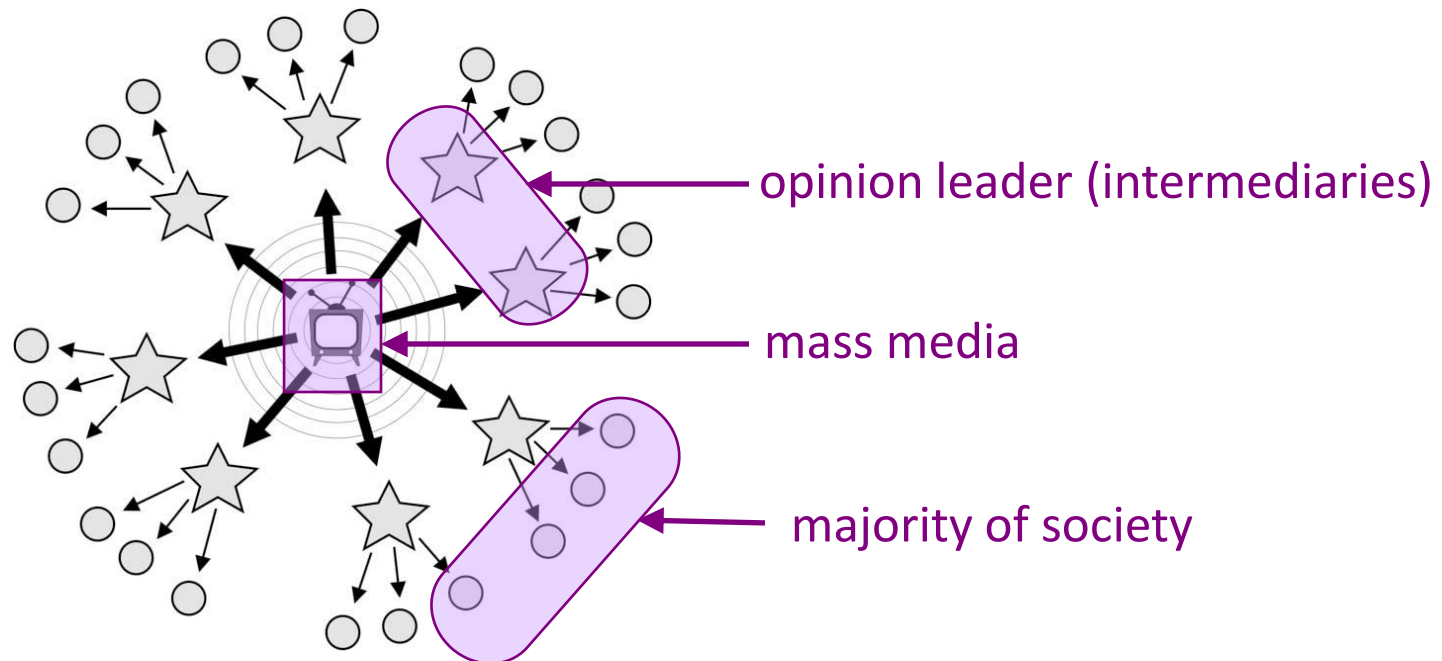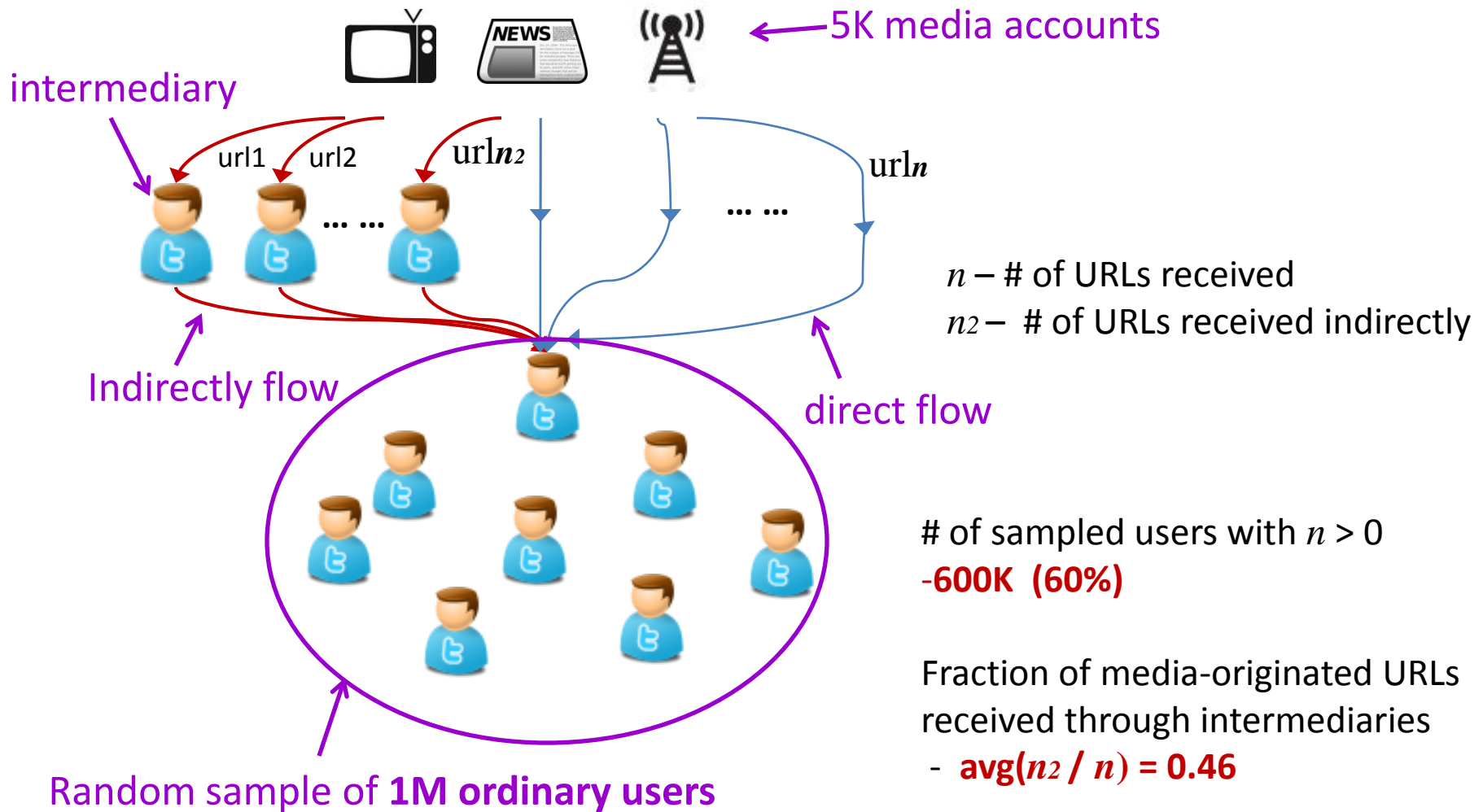|       | Celeb | Media  | Org    | Blog   |
|-------|-------|--------|--------|--------|
| Celeb | 4,334 | 1,489  | 1,543  | 5,039  |
| Media | 4,624 | 40,263 | 7,628  | 32,027 |
| Org   | 1,570 | 2,539  | 18,937 | 11,175 |
| Blog  | 3,710 | 6,382  | 5,762  | 99,818 |

# How does Information flow from media to the masses

- Two-step flow theory (Katz and Lazarsfeld 1955)
  - Media exerts indirect influence on the masses via an intermediate layer of opinion leaders

opinion leader (intermediaries)

mass media

majority of society

# Quantify 2-step flow on Twitter



5K media accounts

intermediary

url1  url2  url$n_2$

url$n$

… …

… …

Indirectly flow

direct flow

$n$ – # of URLs received

$n_2$ – # of URLs received indirectly

# of sampled users with $n > 0$
- **600K (60%)**

Fraction of media-originated URLs received through intermediaries
- **avg($n_2$ / $n$) = 0.46**
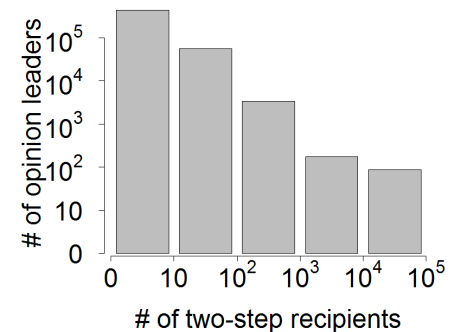
Random sample of **1M ordinary users**

# Who are the intermediaries?

- A large population (490K users) act as intermediaries for 600K users
  - Most (99%) are ordinary
  - Also receive information via two-step flows
  - More exposed to the media
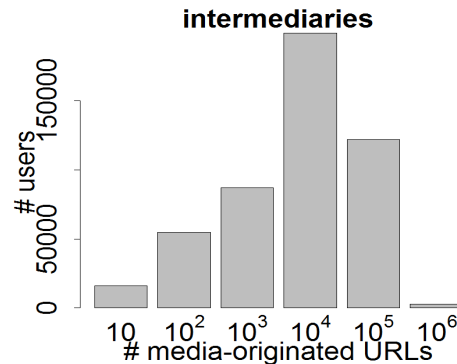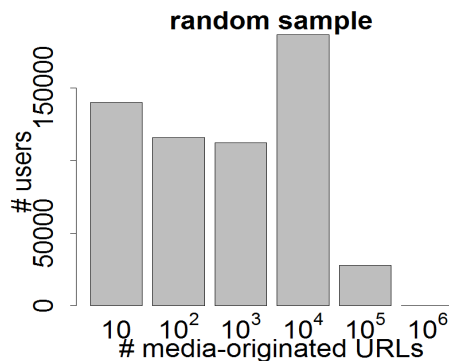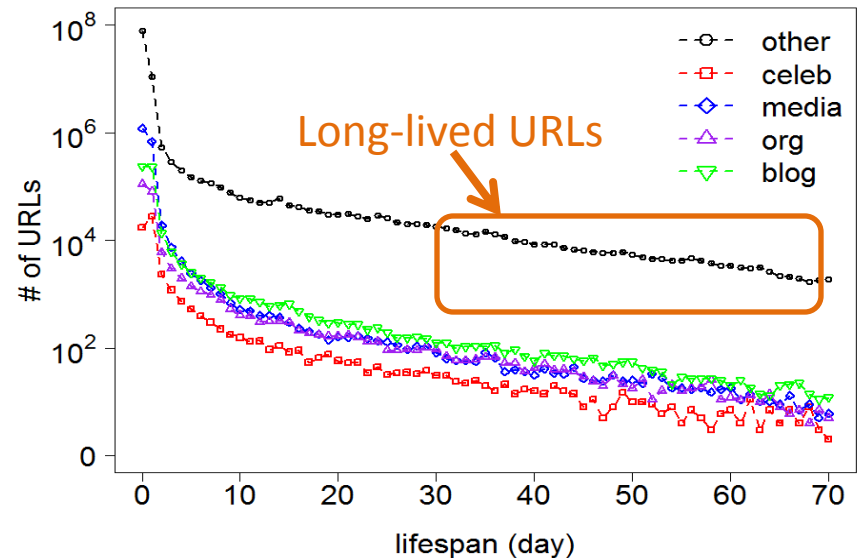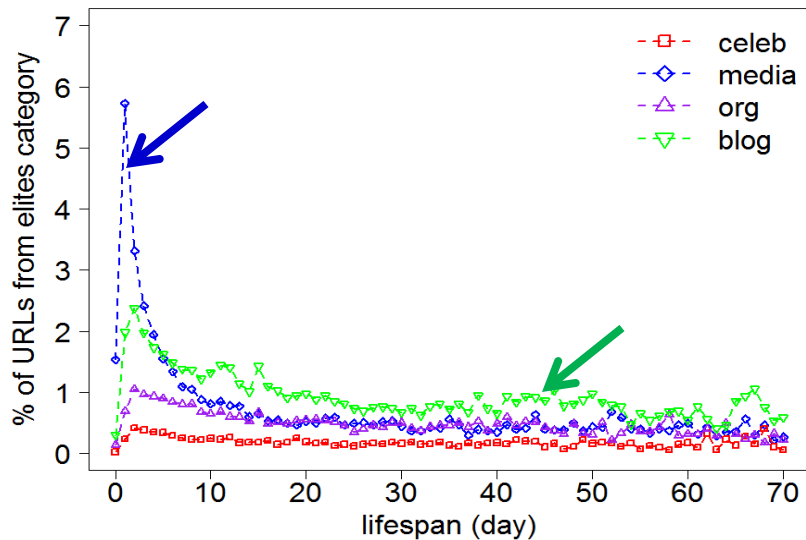


- Opinion leadership is NOT a binary attribute

*Findings consistent with the two-step theory.*

- Who is whom? (user classification)
- Who listens to whom?
- Who says what?

# Lifespan of URLs

- URL lifespan
  - the time lag between the first and last appearance of a given URL on Twitter
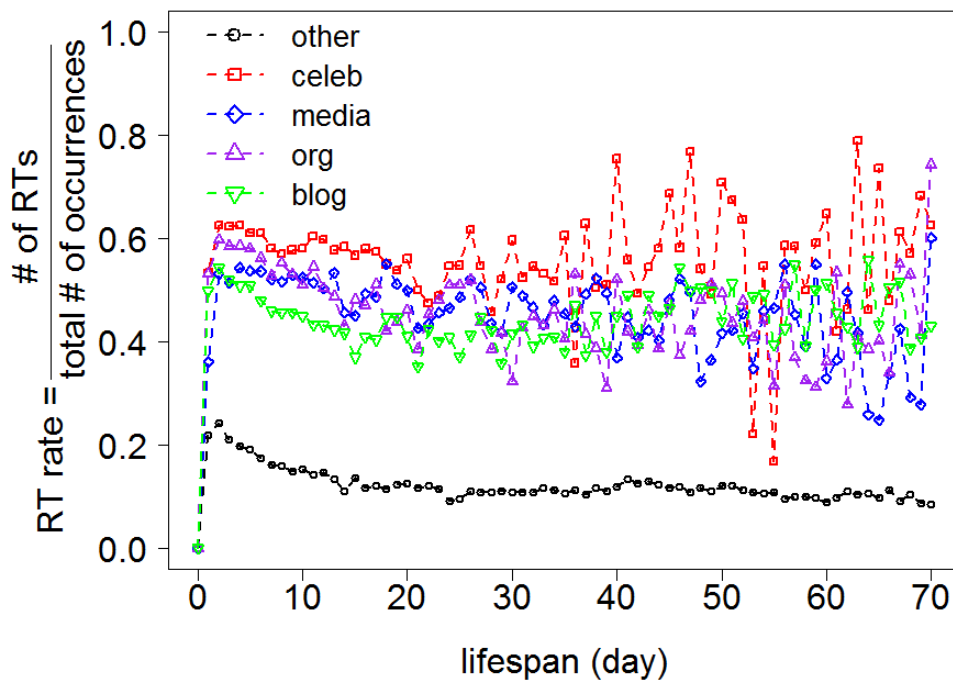
- Lifespan of URLs introduced by different categories



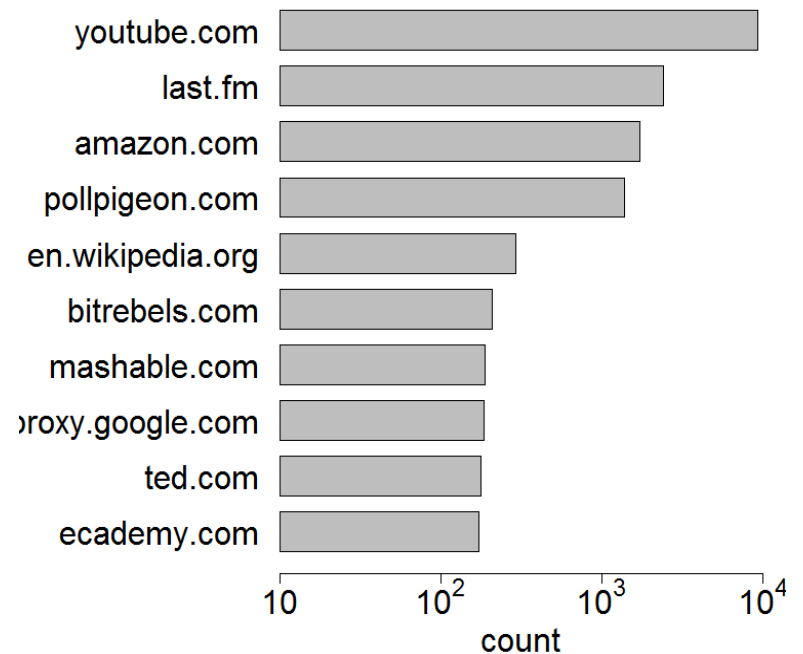* lifespan = 0  means the URL only appeared once

# The mechanism for persistence

## Content vs. Network structure

Average RT rate as a function of lifespan



Top 10 domains for URLs that lived more than 200 days

# Conclusion

- Introduce a method for classifying users into "elite" and "ordinary" categories, using Twitter Lists
- Investigate the flow of information among categories
  - High concentration of attention on a minority of elites
  - A large population of intermediaries passing information from mass media to the masses
- Study the types of contents
  - Different types of content exhibit different characteristic lifespans
  - The persistence of information as a result of content, not structure