INFLUENCE AND ATTENTION ON TWITTER

Duncan Watts
Microsoft Research

Microsoft® Research

New Media…

- Historically communications research divided between
  - Mass media
  - Interpersonal communications
- In last few decades, traditional dichotomy has dissolved
  - Fragmentation of media
    - Cable, Web, satellite radio
  - Empowerment of individuals
    - Email lists, blogs, microblogs, social networking sites, You Tube
- Now have a near-continuous distribution of production
  - Emergence of “mass personal communication”
  - Search and recommendation engines → audience selection

…Old Questions

- In 1940's Harold Lasswell laid out the essential problem of social media:
  - "Who says what to whom, through which channel, and with what effect?"
  - Equally relevant today
- Although easy to ask, this question has proven difficult to answer
  - Measuring "who says what to whom" hard at scale
  - Difficulty compounded by multiplicity of channels
  - Measuring "effects" of all this (i.e. influence) even harder
- Fortunately, Web 2.0 revolution may finally bring the answer within reach

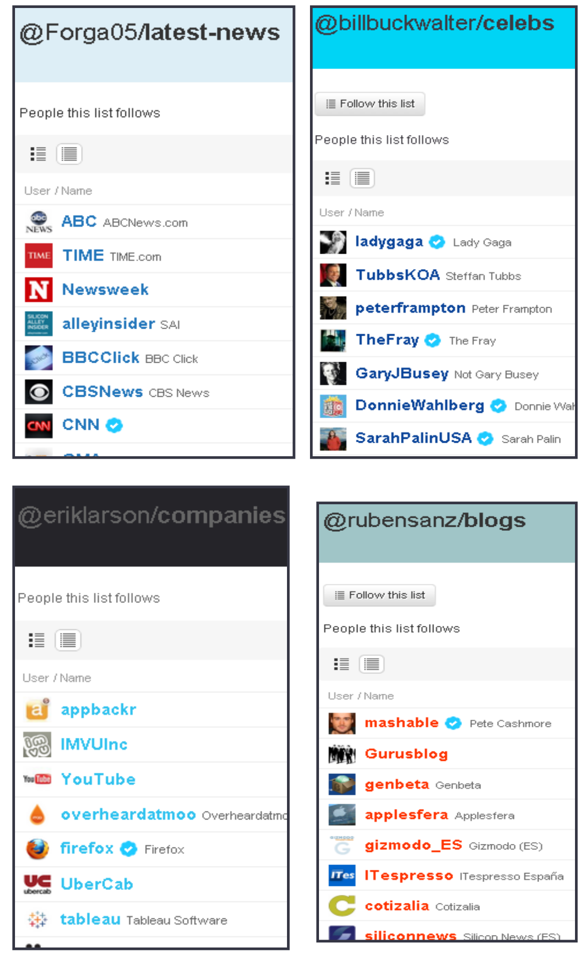
Twitter Well Suited To Lasswell's Maxim

- Full spectrum of production is present
  - Formal organizations (media, government, brands)
  - Celebrities (Ashton, Shaq, Oprah)
  - Public and Semi-Public Figures (bloggers, authors, journalists, public intellectuals)
  - Private Individuals
- Attention is well defined
  - The follower graph
- Information flow is explicit and observable
  - Especially when URLs are included
- Influence can be quantified
  - **Retweets**, click-throughs, conversions

Measuring Attention on Twitter

Wu, Hofman, Mason, Watts (2011)

- Follower graph (Kwak et al 2010)
  - Twitter as observed by 7/31/2009
  - 42M users, 1.5B edges
- Twitter Firehose
  - 223 day period (7/28/2009 – 3/8/2010)
  - 5B tweets, 260M containing bit.ly URLs
- Twitter Lists
  - Tens of millions of lists
  - Very time-consuming to crawl them all
  - Instead introduce two sampling methods



Twitter List Examples

Microsoft Research

Identifying Elite Users

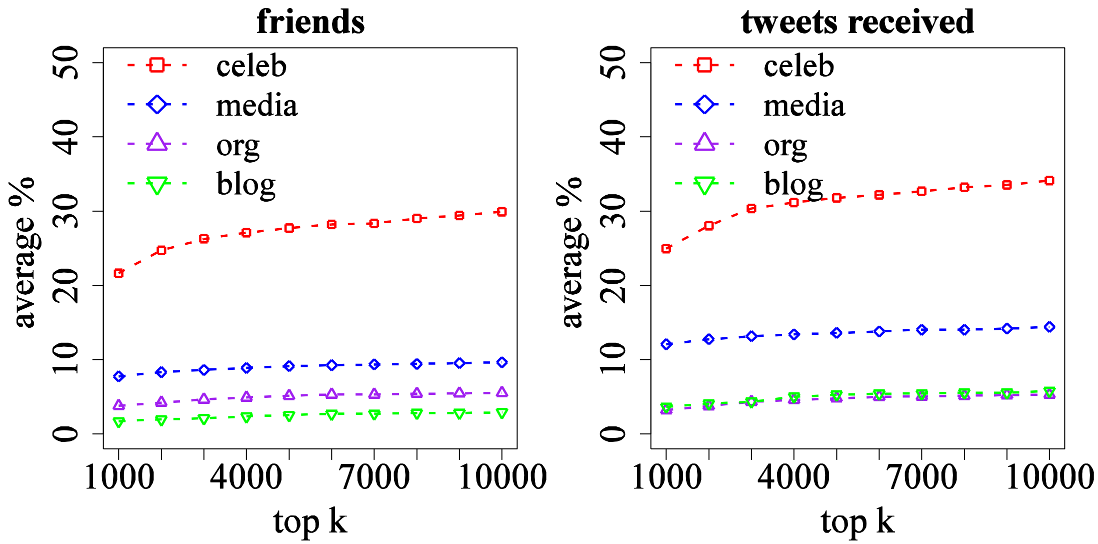
- Rank users by the frequency of being listed in each category

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

- Measure the flow of information from top *k* users in each category to the masses
  - randomly sample 100K ordinary (i.e. unclassified) users, calculate:
    - the average % of accounts they follow among the top k users in each category
    - The average % of tweets they receive from the top k users in each category

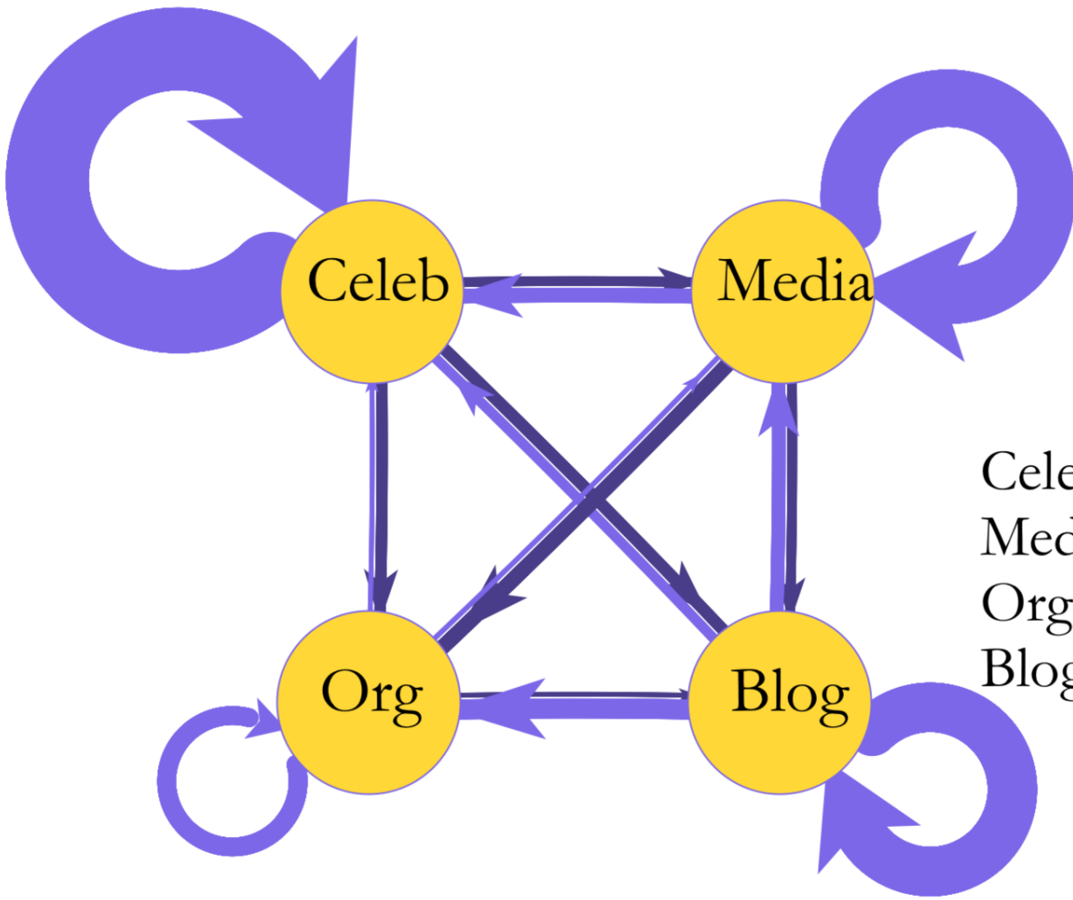
Identifying Elite Users

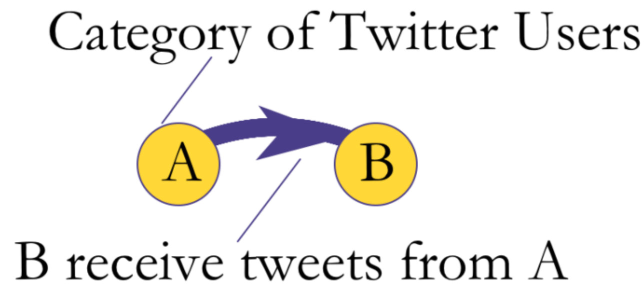


- High concentration of attention
  - Celebrities outrank all other categories
- Let k = 5000
  - Use only the top 5K users in snow-ball sample to represent each category
  - All rest fall into “ordinary” category
  - other values of k gives qualitatively indistinguishable results)
- Accounts for about 50% of all tweets received

Microsoft Research

Attention Between Elites

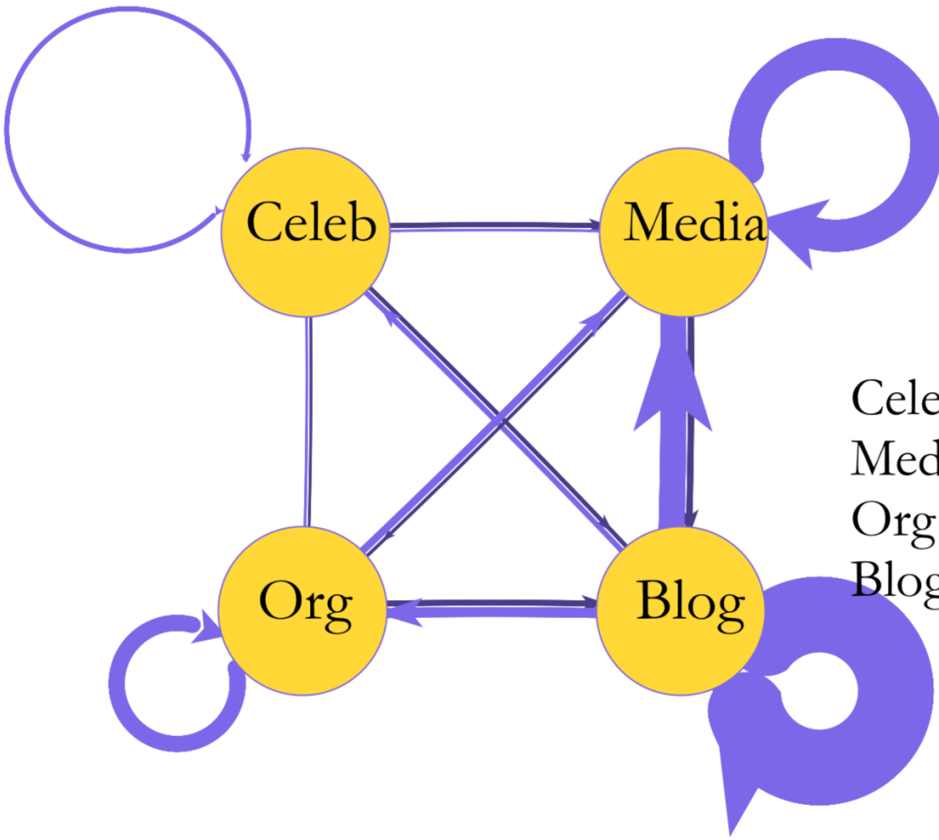


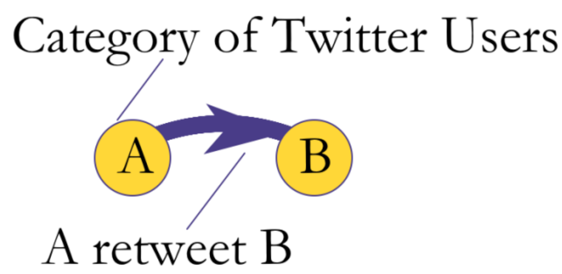


% of tweets received from

	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

Retweets

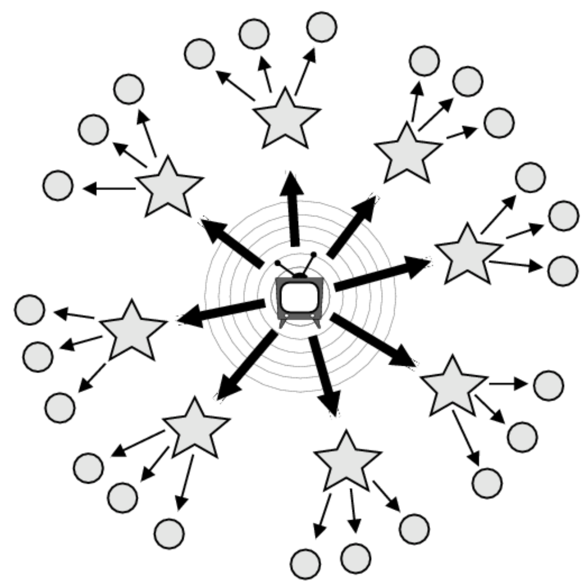




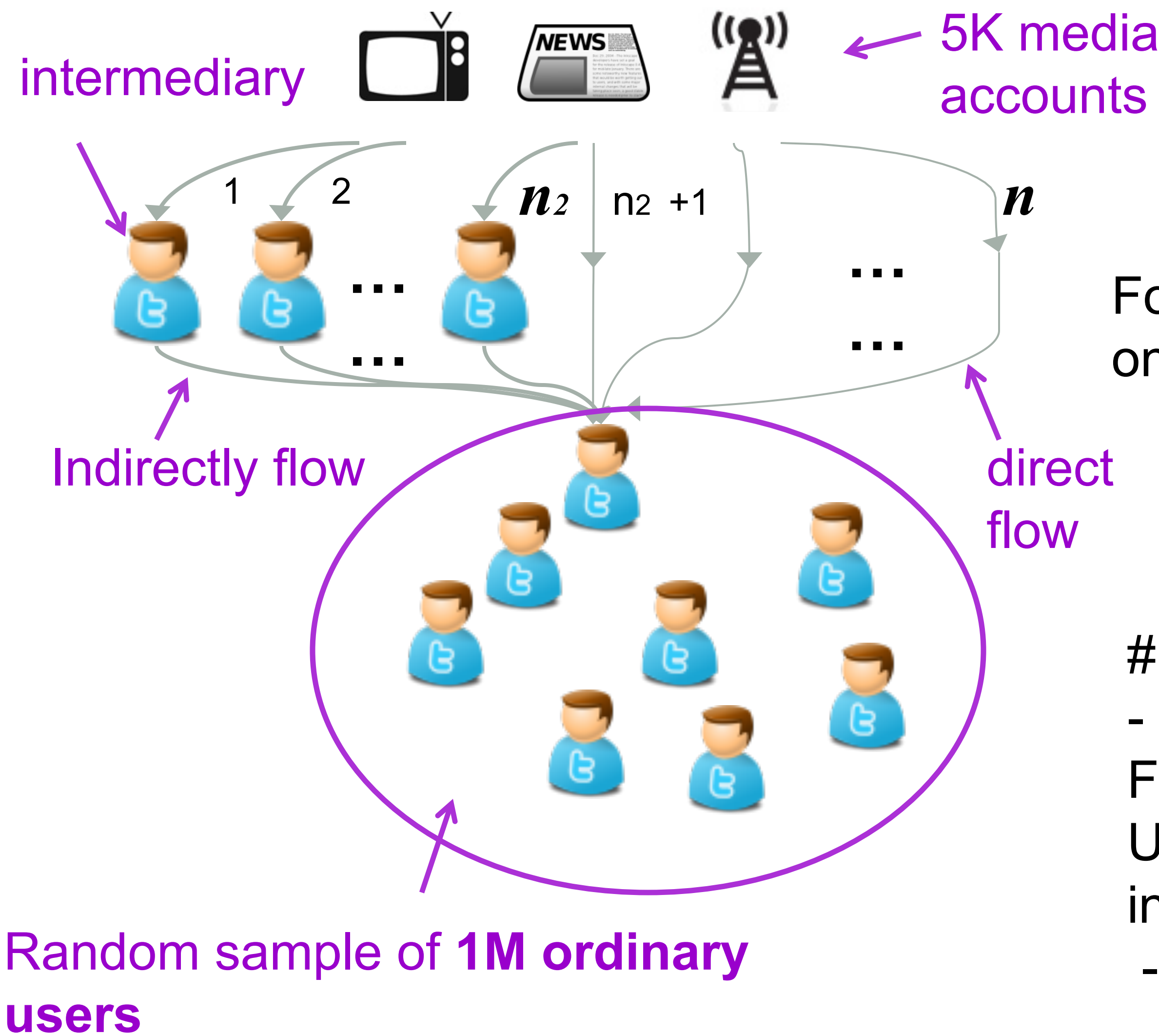
	# of retweets by			
	Celeb	Media	Org	Blog
Celeb	4,334	1,489	1,543	5,039
Media	4,624	40,263	7,628	32,027
Org	1,570	2,539	18,937	11,175
Blog	3,710	6,382	5,762	99,818

The Two-Step Flow of Information

- Research in 1950's emphasized importance of *personal* influence
  - Trusted ties more important than media influence in determining individual opinions
- Also found that not all people are equally influential
  - Opinion leaders act as intermediaries between mass media and the masses
    - More influential, and more exposed to the media
    - But dispersed throughout social strata
- Called this "the two-step flow" of information



Quantify 2-step flow on Twitter



Focus now exclusively on Media-originating URLs

\# of sampled users with n > 0

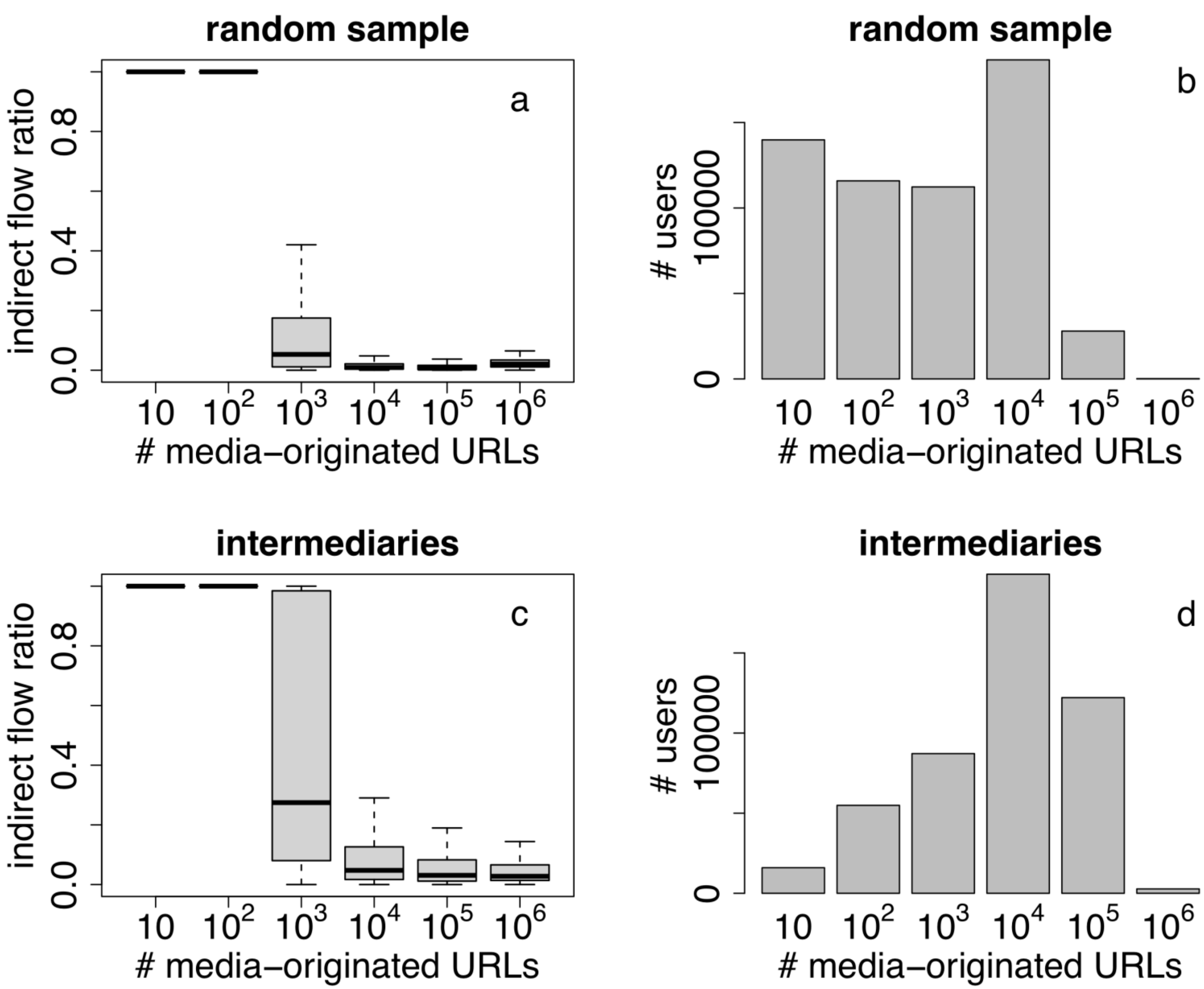
- **600K (60%)**

Fraction of media-originated URLs received through intermediaries

- **avg(n_2/ n) = 0.46**

Microsoft Research

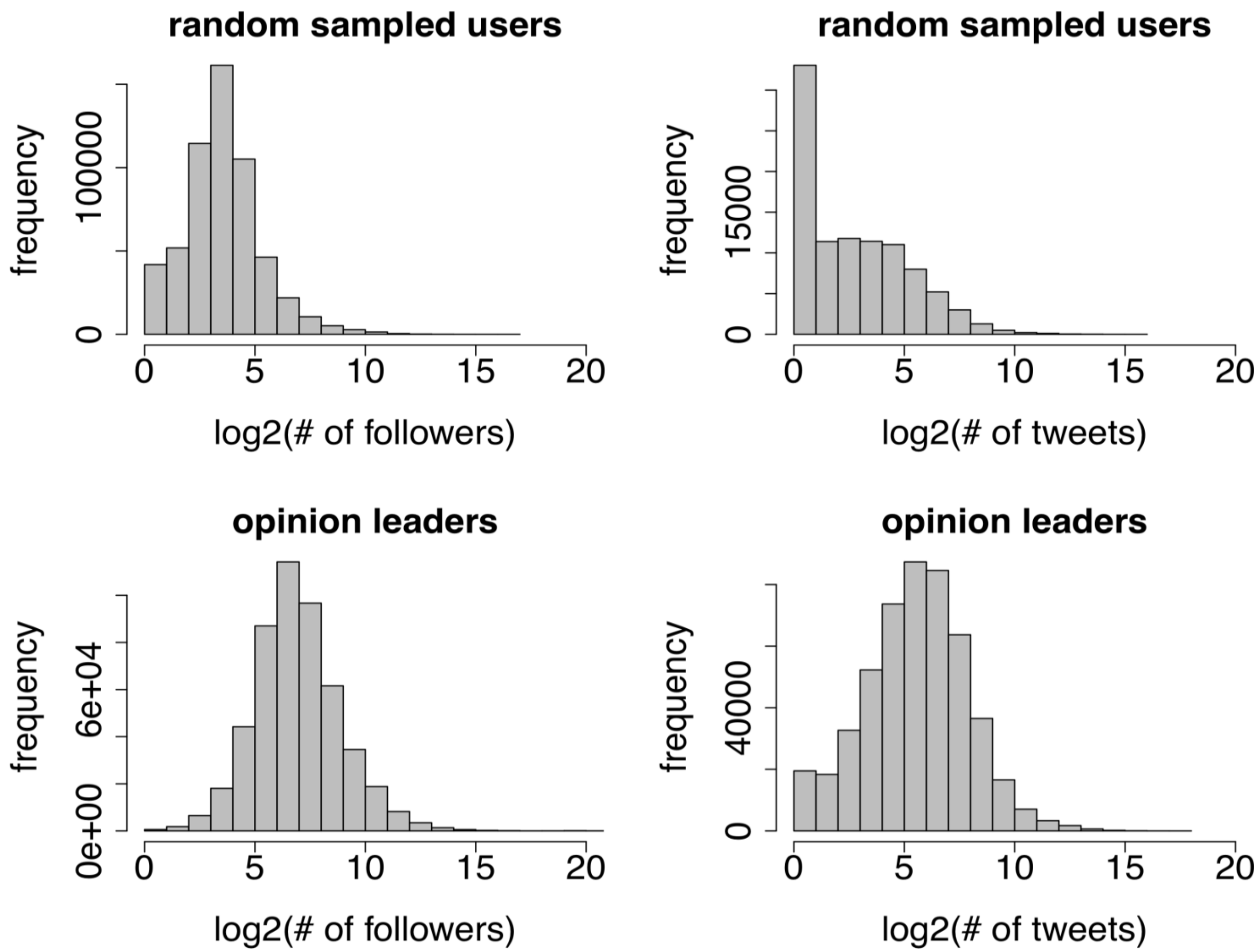
Who Are The Opinion Leaders?



- Not surprisingly, they intermediate more than random users
- Also consume more Media URLs

Microsoft Research

They also tweet more, have more followers



Microsoft Research

Conclusions

- Attention has fragmented, but remains remarkably concentrate on tiny fraction of population
- Surprising support for the Two-step flow
  - Intermediaries have more followers, tweet more, and consume more media
  - Just like the original theory claimed
- Lifespan of content on Twitter reflects the nature of the content, not the influence of the source
  - Twitter really a subset of a larger media ecosystem, from which it draws and redraws content

From Attention to Influence

- Opinion leaders are interesting in part because they appear to generate a “multiplier effect”
  - Influence one opinion leader and they will influence X others
- Two-step flow has become conflated with diffusion research to produce notion of “Influencers”
  - “Law of the Few” (Gladwell, 2000)
  - “One in ten Americans tells the other nine how to vote, where to eat, and what to buy.” (Keller and Berry, 2003)
  - “Influencers have become the ‘holy grail’ for today’s marketers.” (Rand, 2004)

BUT GRAILS ARE HARD TO FIND…

Microsoft Research

Can One Predict Influencers?

- After the fact, can always tell a story about why X succeeded
  - Can identify some group of individuals who were involved early on
  - They will seem to have been influential
- But to *make use* of influencers, need to identify them in advance
- Very little evidence that marketers (or anyone else) can do this consistently

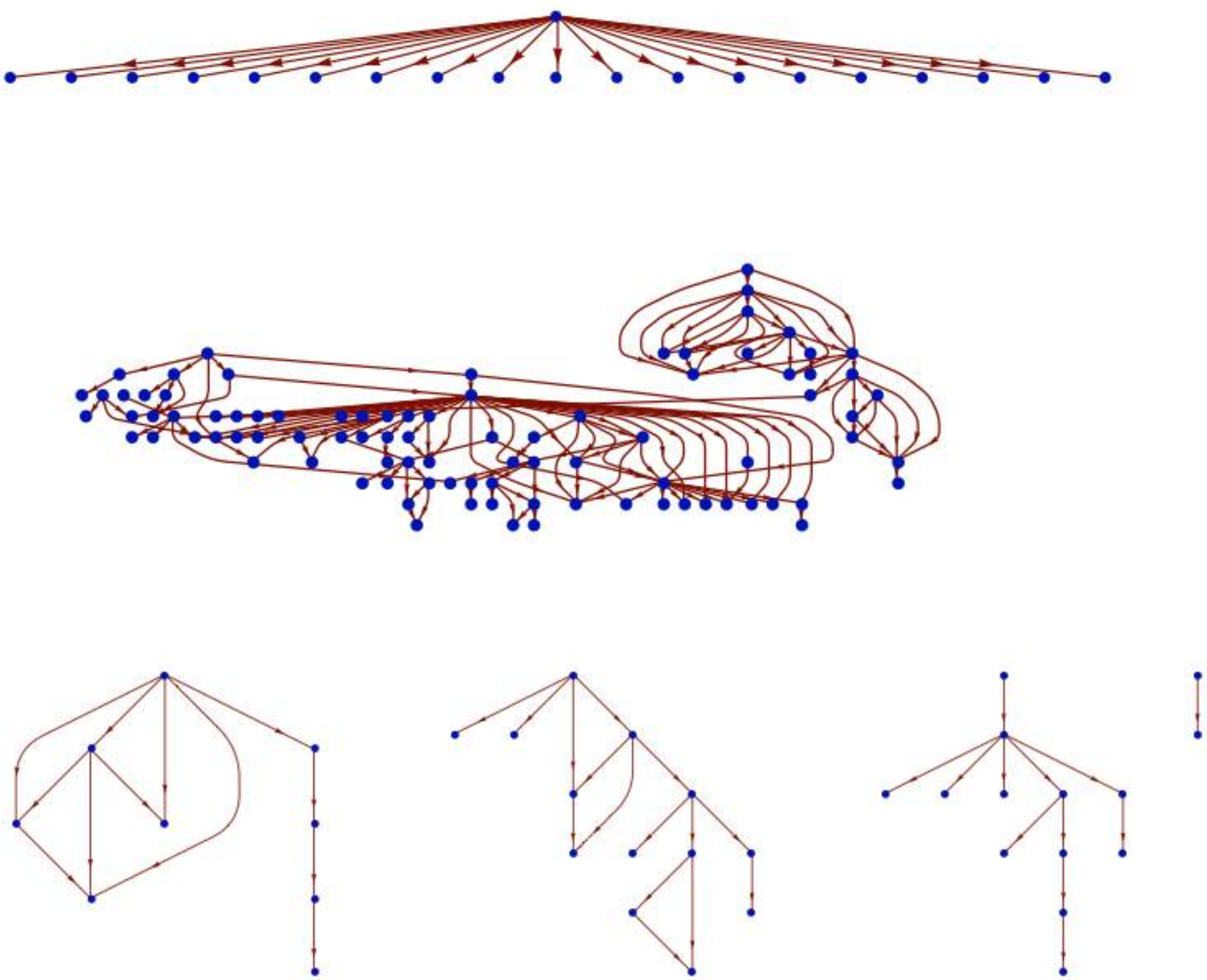
Microsoft Research

Influence on Twitter

Bakshy, Hofman, Mason, Watts (2011)

- An individual “seed” user tweets a URL (here we consider only bit.ly)
- For every follower who subsequently posts same URL (whether explicit “retweet” or not), seed accrues 1 pt
- Repeat for followers-of-followers, etc. to obtain total influence score for that “cascade”
  - Where multiple predecessors exist, credit first poster
  - Can also split credit or credit last poster (no big changes)
- Average individual influence score over all cascades
  - Highly conservative measure of influence, as it requires not only seeing but acting on a tweet
  - Click-through would be good, but not available to us

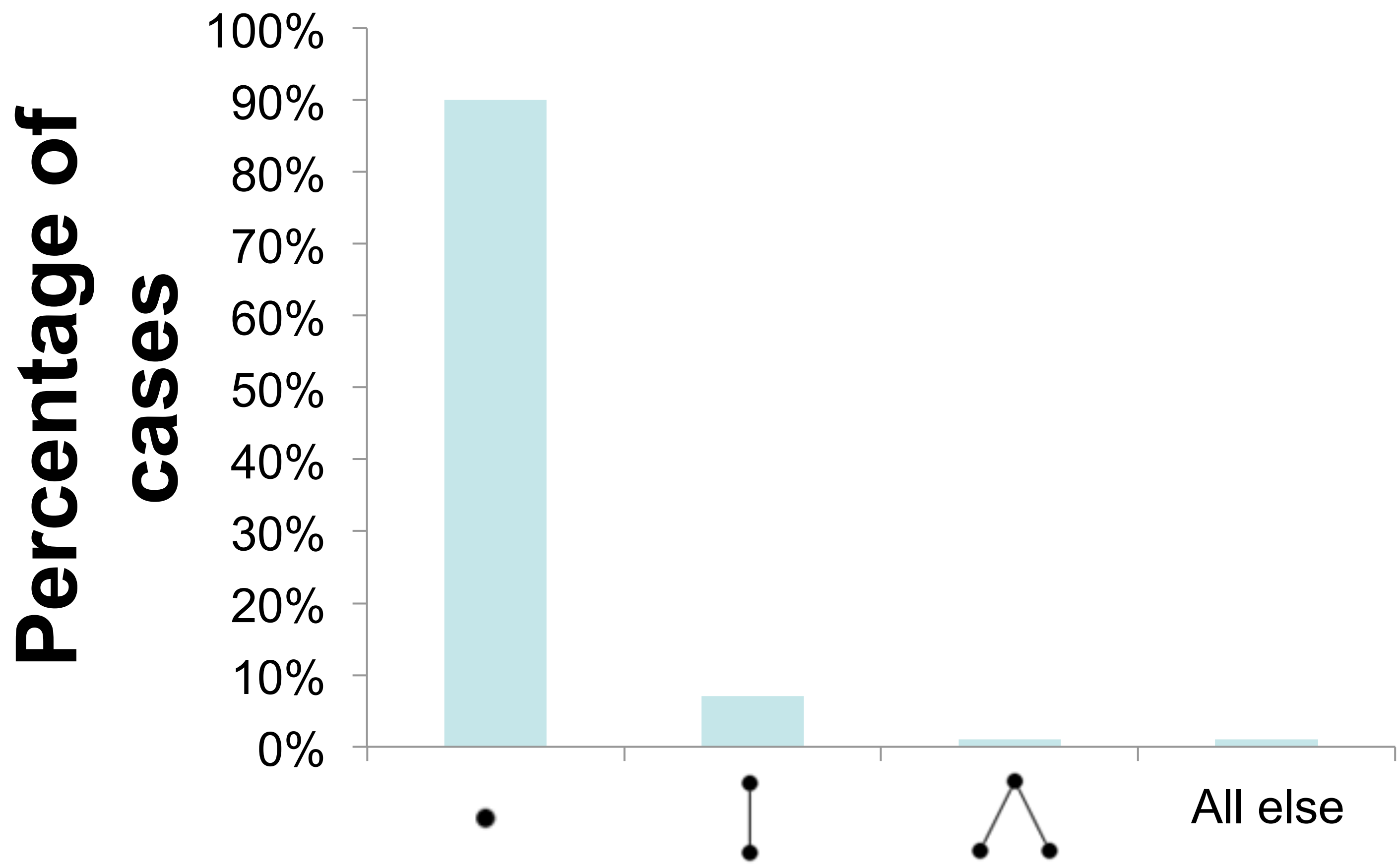
Cascades on Twitter



- 1.6M distinct "seeds"
- Each seed posts average of 46.3 bit.ly URL's
- Hence 74M cascades total
- Average cascade size 1.14
  - Median cascade size 1
- Average influence score is 0.14

Microsoft Research

Most Tweets Don't Spread

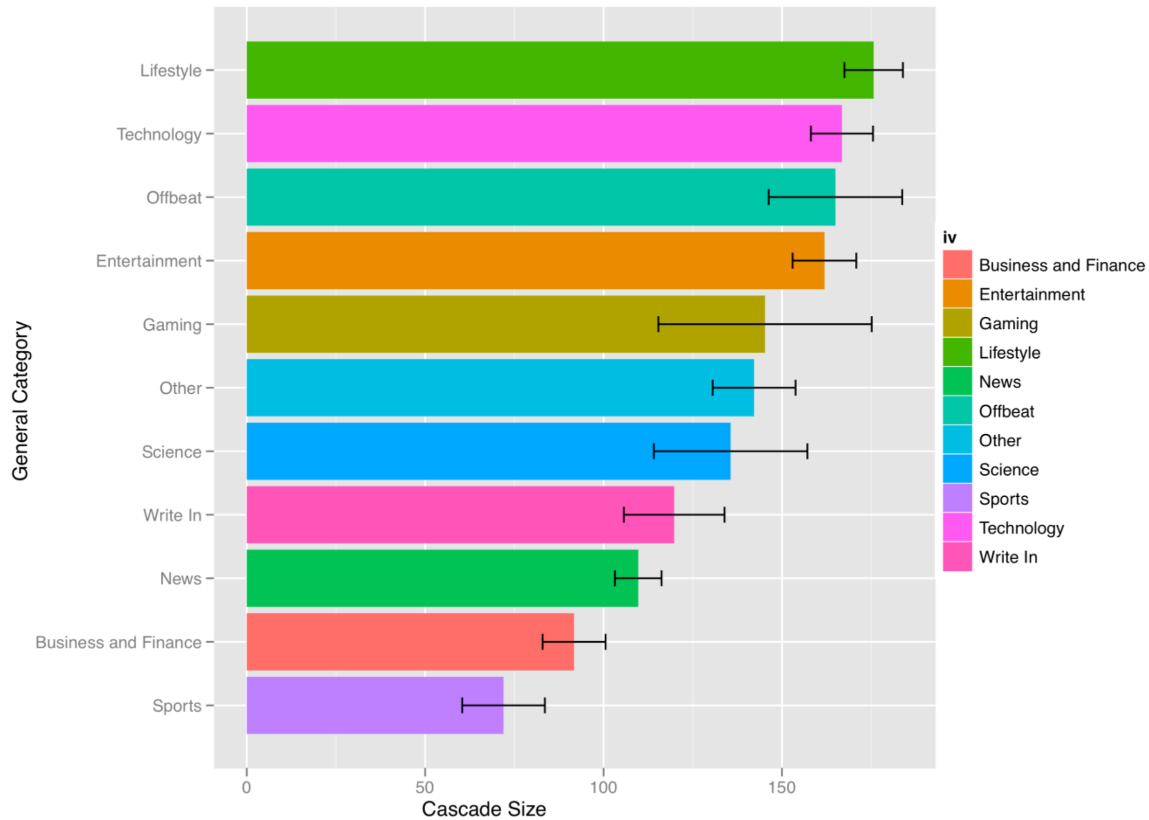


~ 90% of <u>adoptions</u> are direct from the source

~ 99% of <u>adoptions</u> are within 1 hop from the source

Microsoft Research

Content and Cascade Size

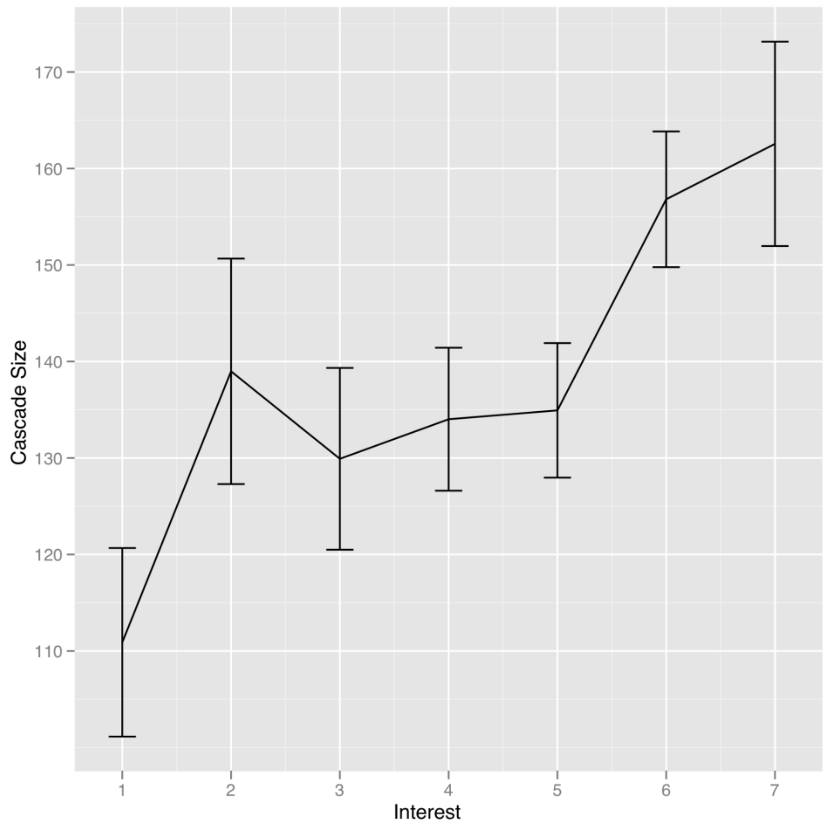


URLs in the "Lifestyle" category spread farthest

Very local and very global topics (Sports & News) spread the least

Microsoft Research

Interest and Cascade Size



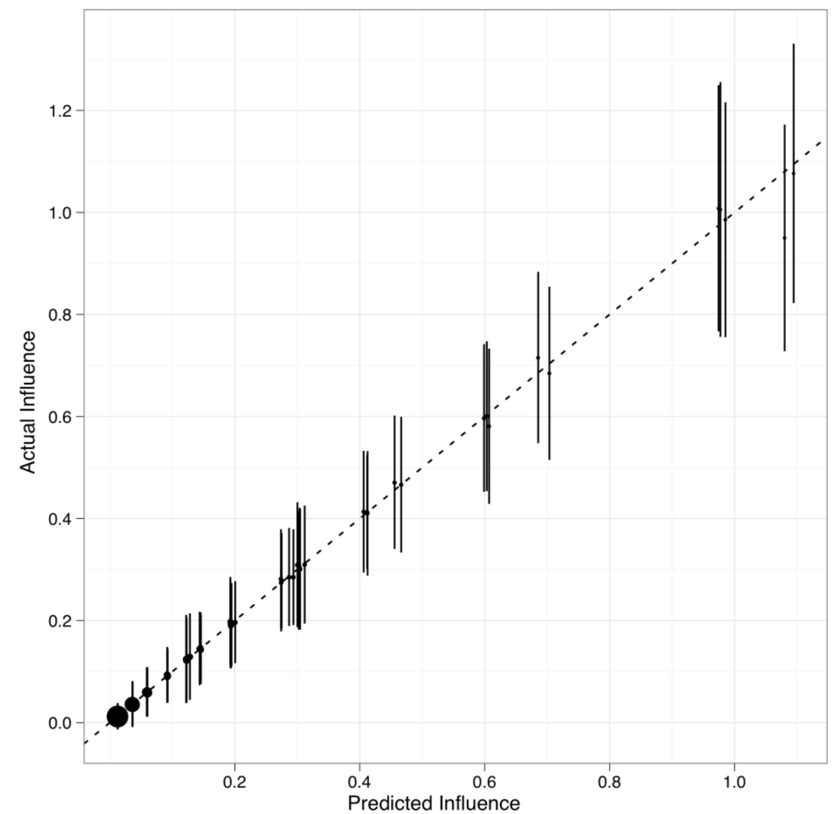
Unsurprisingly, on average more interesting URLs spread farther

Predicting Influence

- Objective is to predict influence score for future cascades as function of
  - # Followers, # Friends, # Reciprocated Ties
  - # Tweets, Time of joining
  - Past influence score
- Fit data using regression tree
  - Recursively partitions feature space
  - Piecewise constant function fit to mean of training data in each partition
  - Nonlinear, non-parametric
    - Better calibrated than ordinary linear regression
  - Use five-fold cross-validation
    - For each fold, estimate model on training data, then evaluate on test data
    - Every user gets included in one test set

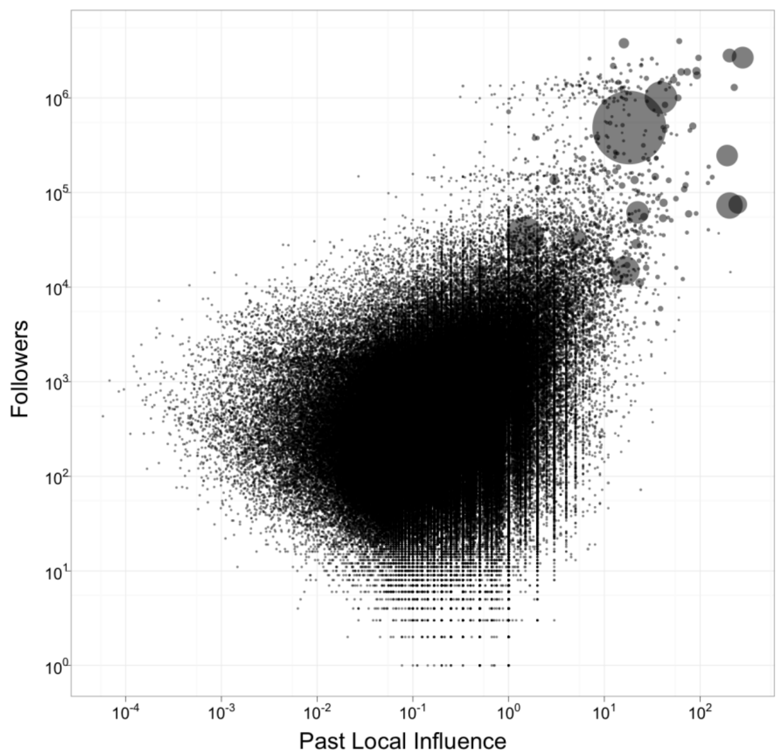
Results

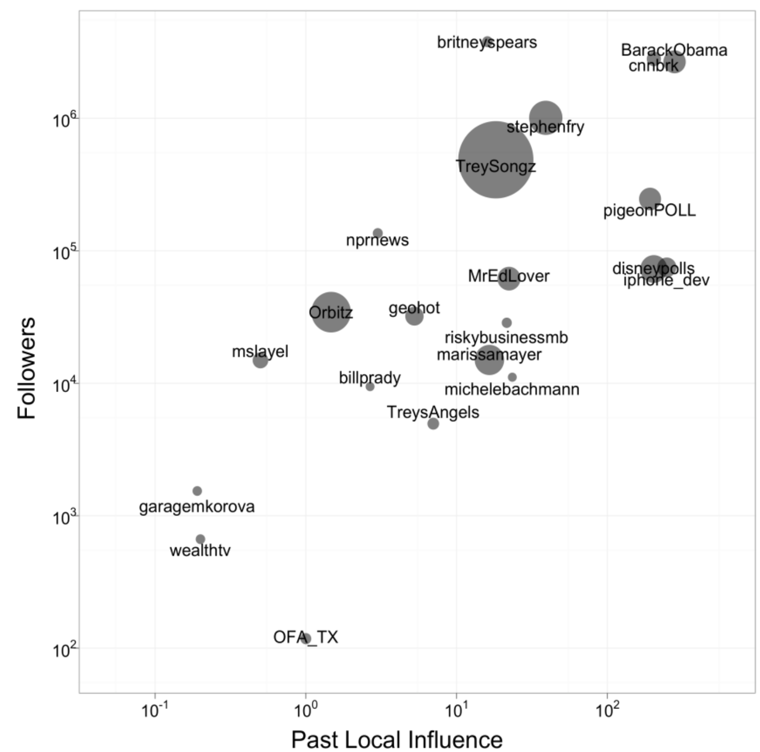
- Only two features matter
  - Past local influence
  - # Followers
- Surprisingly, neither # tweets nor # following matter
- Also surprisingly, content doesn't help
- Model is well calibrated
  - average predicted close to average actual within partitions
- But fit is poor (R^2 = 0.34)
  - Reflects individual scatter



Microsoft Research

Who are the Influencers?





Circles represent individual seeds (sized by influence)

Microsoft Research

Necessary but not sufficient

- Seeds of large cascades share certain features (e.g., high degree, past influence)
- However, many small cascades share those features, making "success" hard to predict at individual level
- Common problem for rare events
  - School shootings, Plane crashes, etc.
  - Tempting to infer causality from "events," but causality disappears once non-events accounted for
- Lesson for marketers:
  - Individual level predictions are unreliable, even given "perfect" information
- Fortunately, can target *many* seeds, thereby harnessing average effects

Microsoft Research

Should Kim Kardashian Be Paid $10,000 per Tweet?

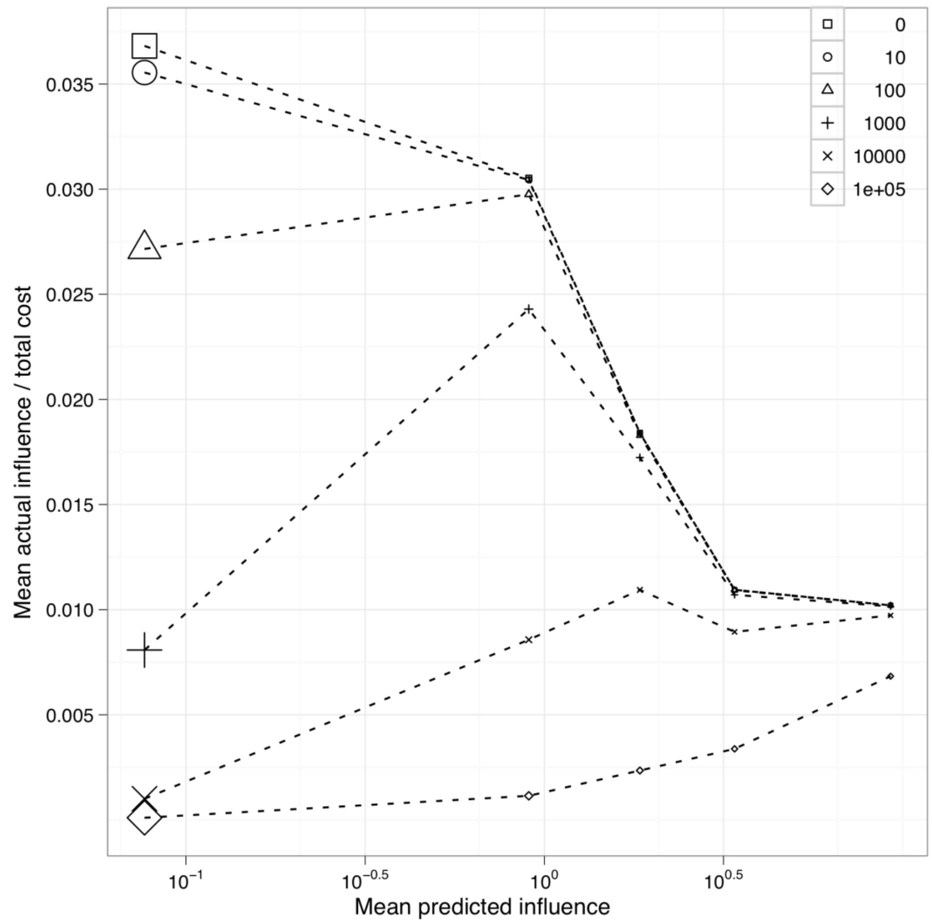
- On average, some types of influencers are more influential than others
  - Many of them are highly visible celebrities, etc. with millions of followers
  - But these individuals may also be very expensive (i.e. Kim Kardashian)
- Assume the following cost function
  - $c_i = c_a + f_i * c_f$, where c_a = acquisition cost; c_f = per-follower cost
  - Also $c_a = a * c_f$, where a expresses cost of acquiring individual users relative to sponsoring individual tweets
- Should you target:
  - A small # of highly influential seeds?
  - A large # of ordinary seeds with few followers?
  - Somewhere in between?

Microsoft Research

"Ordinary Influencers" Dominate

- Assume c_f = \$0.01
  - Equivalent to paying \$10K per tweet for user with 1M followers
- When c_a = \$1,000, (a = 100,000) highly influential users are most cost effective
- But for lower ratios, most efficient choice can be individuals who influence at most one other

Influence per Follower



Microsoft Research

Conclusions

- Attention on Twitter is surprisingly concentrated
  - 50% of attention is directed to one of ~ 0.1% of users
- Nevertheless, influence is hard to predict
  - Most cascades are tiny
  - Large cascades are more likely to start with highly visible users
  - But efficiency is often maximized by targeting "ordinary" influencers (who influence just one other on average)
- By targeting many seeds, can improve predictive power dramatically
  - Consistent with "big seed" model, not "epidemics"
  - No free lunch, but a cheap snack isn't bad

Microsoft Research

References

Shaomei Wu, Jake Hofman, Winter A. Mason, and Duncan J. Watts. "Who says what to whom on Twitter" *Proceedings of the 20th international conference on World Wide Web*, Hyderabad, India (2011)

Eytan Bakshy, Jake Hofman, Winter Mason, and Duncan J. Watts. "Everyone's an influencer: Quantifying Influence on Twitter" *Proceedings of the 4th International Conference on Web Search and Data Mining*, Hong Kong (2011)

<u>Background:</u>

D. J. Watts and P. S. Dodds. "Networks, influence, and public opinion formation." *Journal of Consumer Research*, 34(4), 441-458 (2007).

D. J. Watts. Challenging the "Influentials Hypothesis." *Measuring Word of Mouth, Vol. 3. Word of Mouth Marketing Association* (2007).

D. J. Watts. "The Accidental Influentials." *Harvard Business Review*, p. 22-23 (February, 2007)

D. J. Watts and J. Peretti. Viral marketing in the real world. *Harvard Business Review* (May, 2007)

Microsoft Research