Topical Semantics of Twitter Links

Jan Vosecky

About the paper

- **Topical Semantics of Twitter Links**
- WSDM'11
- Authors:
  - Michael J. Welch, Yahoo!
  - Uri Schonfeld, UCLA
  - Dan He, UCLA
  - Junghoo Cho, UCLA

Outline

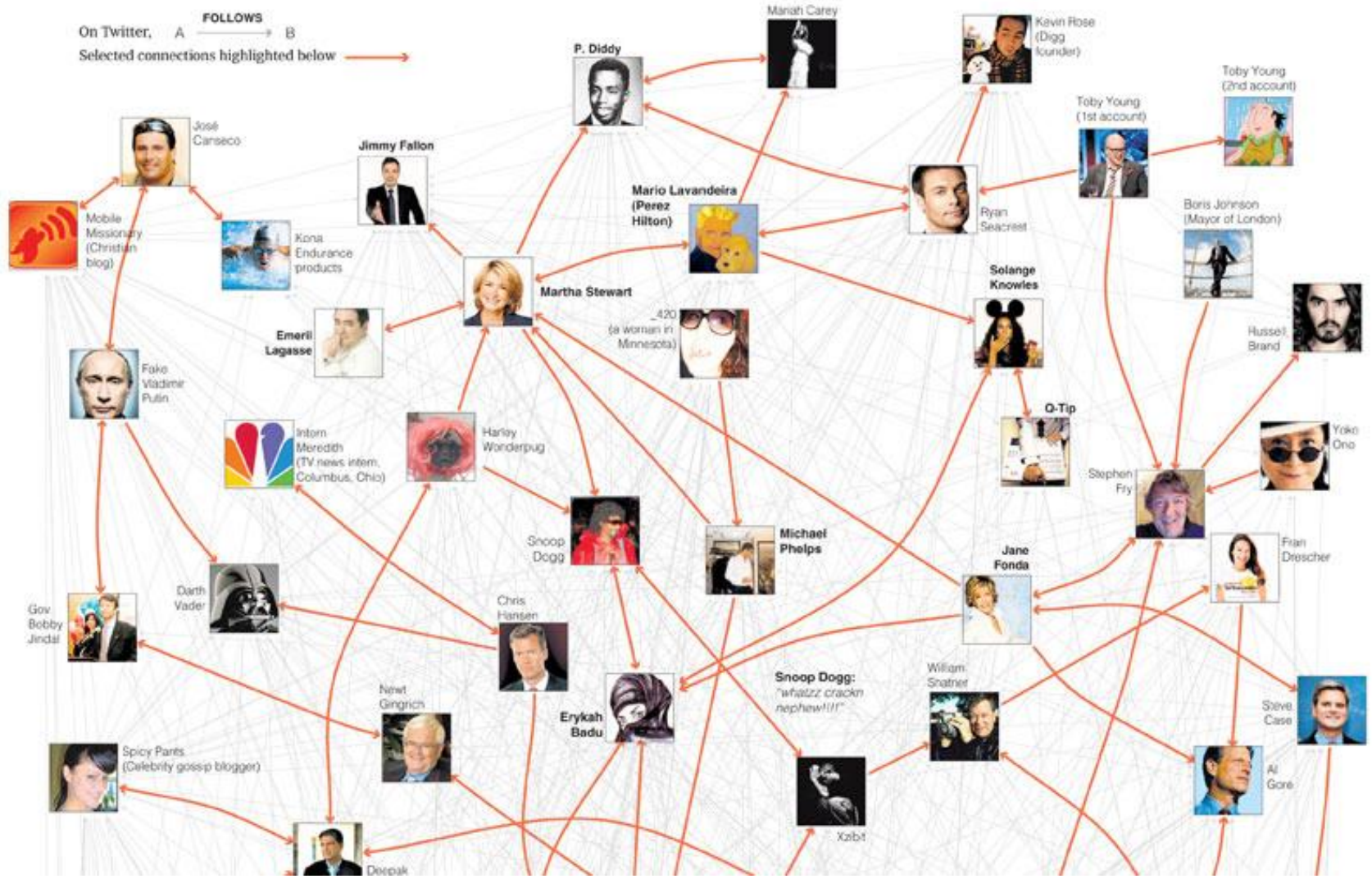
- Introduction, problem setting
- Modelling Twitter
  - Graph model
  - Graph analysis
- Link semantics
  - Implication for ranking
- Experiments, results
- Open questions

Introduction

Background: Twitter

- 10th highest internet traffic world-wide
  - Source of breaking news, announcements, comments and opinions
- Social network structure
  - Links
    - Follow-relationship
      - Following and reading content from another user
    - Re-tweet relationship
      - Re-posting content from another user
  - Semantics of the links? ('topics')
- User roles: reader / writer
- Ongoing efforts: finding influential users

Background: Twitter



Topic-specific influence

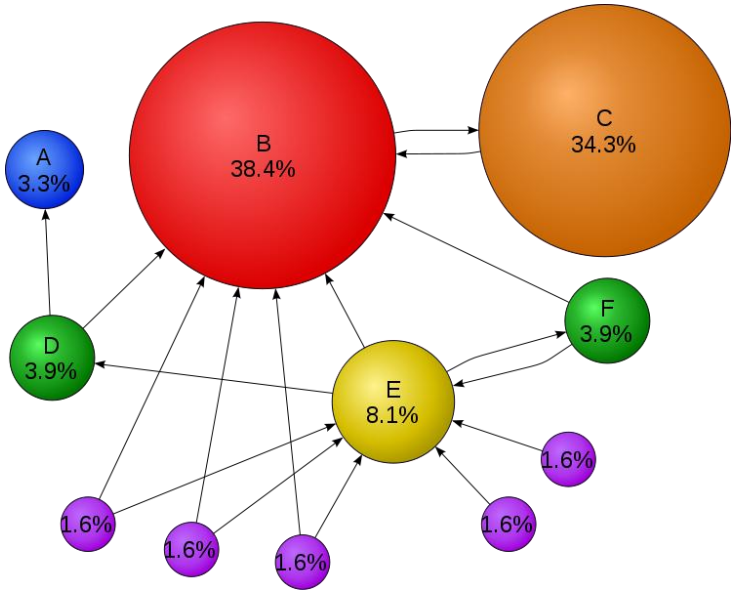
- Given a social network graph
  - Identify relevant and high-ranking users for a topic
    - Using e.g. PageRank
  - Evaluate topical relevance of high-ranked users
- Possible graphs in Twitter:
  - Follow-graph, retweet graph, etc.
- Questions:
  - Is topical relevance transitive?
  - Which relationship better preserves topical relevance?

Related work

- Structure and growth of the web
  - Web graph
    - Broder et al. (2000), Kumar et at. (1999)
    - Power-law distributions
    - Connected components
- Twitter graph analysis
  - Cha et al. Measuring User Influence in Twitter: The Million Follower Fallacy (ICWSM'10)
    - Follow, retweet and mention relationships
  - Weng et al. TwitterRank: Finding topic-sensitive influential twitterers (WSDM'10)
    - Analysis of follow relationships, posting frequency

Related work

- PageRank
  - PageRank (PR) of node u: $$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$



Related work

- Extensions of PageRank to Twitter
  - Utilize the global link structure
  - TunkRank, 2009 (http://tunkrank.com/)
    - Influence propagates over follow-links, no topic sensitivity
  - Weng, et al. TwitterRank: Finding topic-sensitive influential twitterers. WSDM '10
    - Follow-links as well as topical similarity derived from user's tweets

$$\overrightarrow{TR_t} = \gamma P_t \times \overrightarrow{TR_t} + (1-\gamma)E_t \qquad P_t(i,j) = \frac{|\mathcal{T}_j|}{\sum_{a:\ s_i\ follows\ s_a} |\mathcal{T}_a|} * sim_t(i,j)$$

- Pal and Counts, Identifying Topical Authorities in Microblogs. WSDM'11
  - Feature-based approach to rank users by authority
  - Influence does not propagate

Goal of the paper

- Recent efforts to rank users by quality and topical relevance
  - Mainly focus on the "follow" relationship
  - Topic-specific influential users
- Twitter's data offers additional implicit relationships
  - "retweets" and "mentions"
  - In this paper: investigate the semantics of the follow and retweet relationships
    - Rich graphical model
- Related questions
  - How does the Twitter graph compare with the Web graph?

Modelling Twitter

Modelling Twitter

- Full Twitter graph
  - Nodes: User, Post
  - Edges:
    - Publishes
    - Follows

    } *explicit*

    - Re-tweets
    - Mentions

    } *implicit*
  - Edge type is uniquely identified by the types of nodes it connects
    - → No special distinction of edge types needed

	User	Tweet
User	Follow	Publish
Tweet	Mention	Retweet

  - → Directed graph $G = (V, E)$ where $V = U + P$

Modelling Twitter

- Full Twitter graph
  - Matrix representation:
    - Similar to Web graph representation
    - *T:* |*U*| + |*P*| by |*U*|+|*P*| matrix, where |U| is the number of users and |P| is the number of posts
    - A non-zero value in *Tij* represents an edge between node i and node j

	U1	**U2**	**P1**	**P2**
U1	-	0	1	0
U2	1	-	0	1
P1	0	0	-	0
P2	0	0	1	-

Modelling Twitter

- Simplified graph
  - User-user only
  - Matrix representation:
    - *T:* |*U*| by |*U*| matrix, where |U| is the number of users
    - Each T_{ij} can have a value of:
      - *f*, indicating a follow-relationship
      - *r*, indicating a re-tweet relationship

	U1	**U2**
U1	-	-
U2	f,r	-

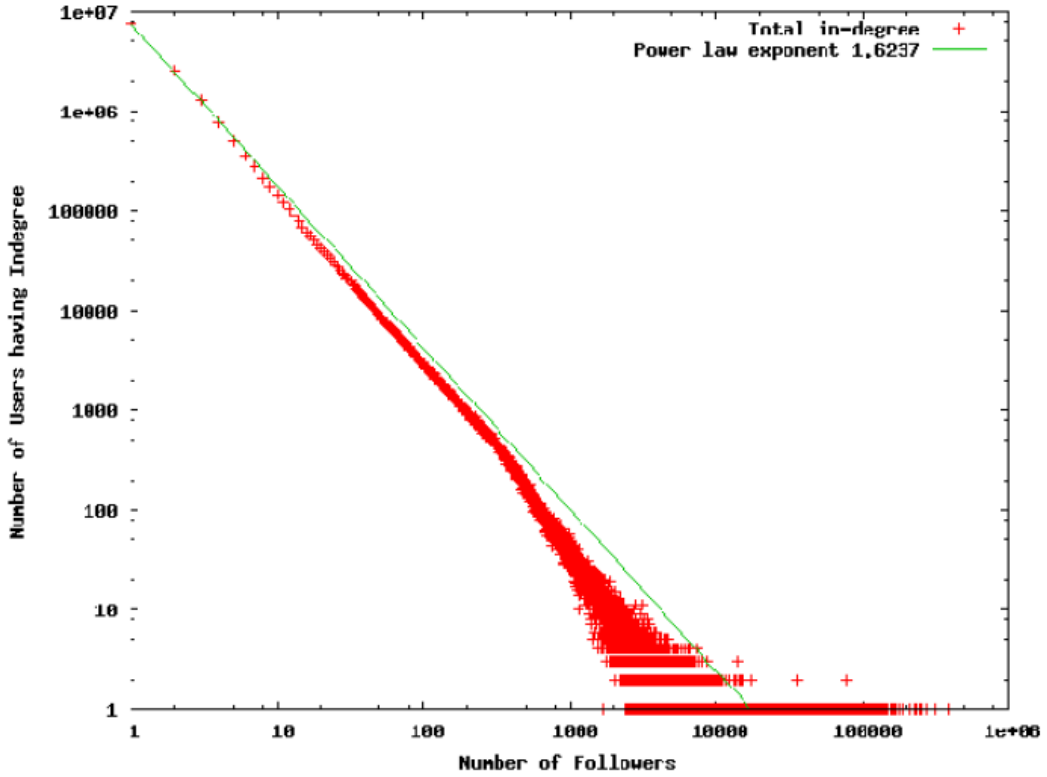
- Additional information – not included:
  - Time, hyperlinks, post content, location

Graph analysis

- Dataset
  - 1.1 million users
    - 273 million follow edges
    - 2.9 million re-tweet edges
  - October 2009 - January 2010

Graph analysis

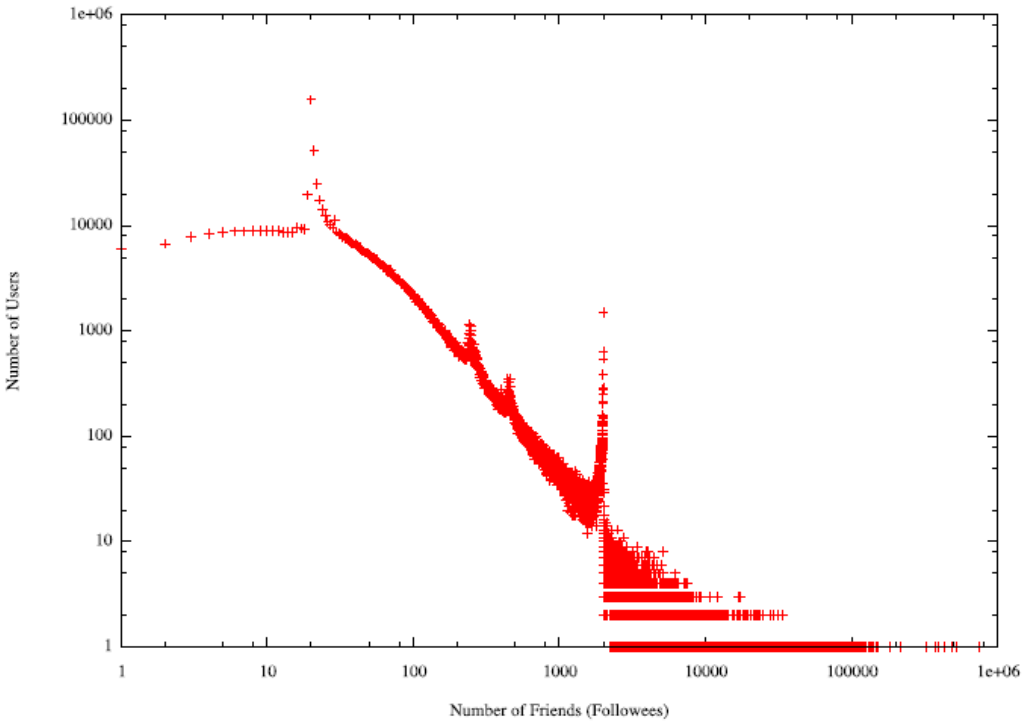
- Follow relationship
  - Inlink distribution (how users are followed as writers)
    - Power-law distribution



(a) Inlinks (followers)

Graph analysis

- Follow relationship
  - Outlink distribution (how many users people follow)

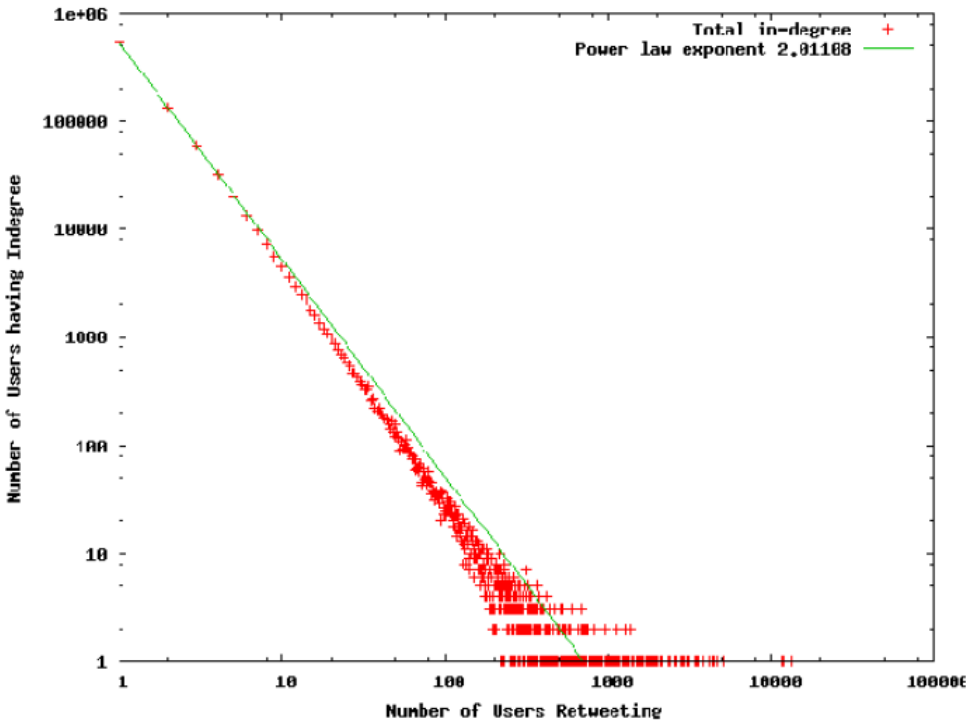


(b) Outlinks (friends)

- Spike around the 20-friend region
  - During signup, an initial set of 20 "recommended" users to follow
- Spike exactly on the 2000-friend mark
  - Restrictions on following more than 2000 users

Graph analysis

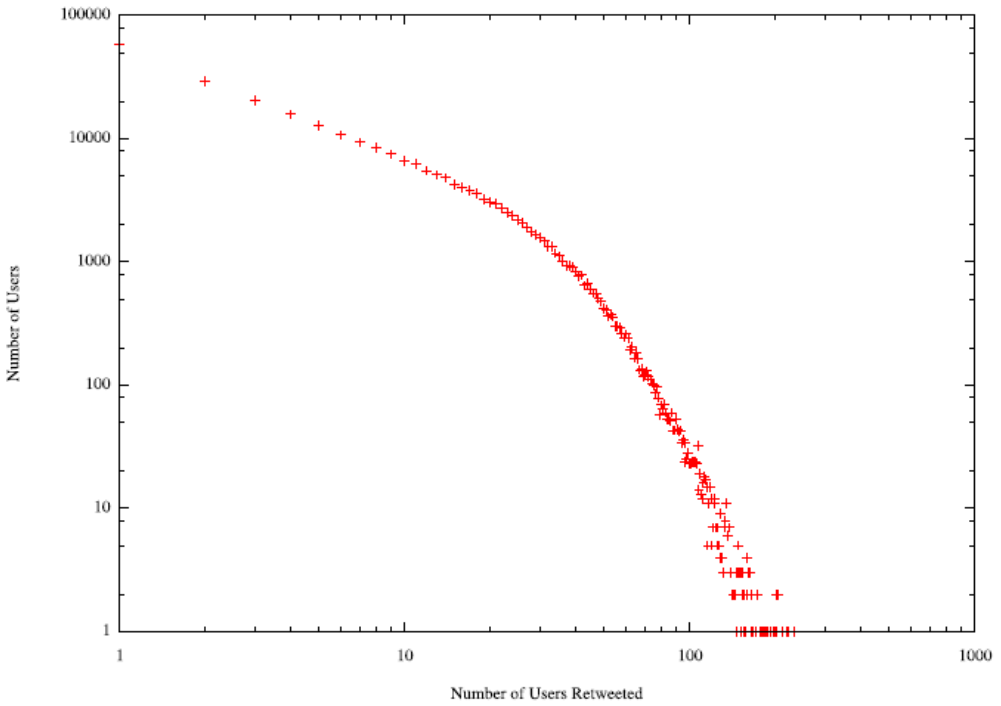
- Retweet relationship
  - Inlink distribution
    - number of unique users who retweeted at least one post of the user
    - Power-law distribution
    - distribution similar to hyperlinks on the Web



(a) Retweet Inlinks

Graph analysis

- Retweet relationship
  - Outlink distribution
    - number of unique users whose posts were retweeted by a given user
    - Does not follow a power-law distribution



(b) Retweet Outlinks

Graph analysis

- Tweet frequency
  - Over a period of 31 days
  - Large group of users who published only a single post
  - Large number of users wrote more than 100 posts

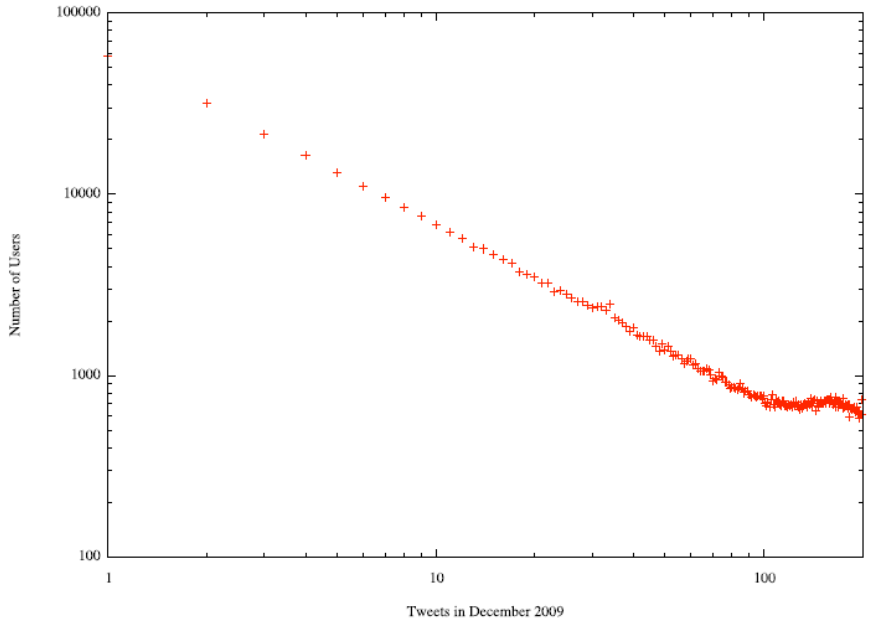


Figure 3: Tweet Frequency

Graph analysis

- Readers and Writers

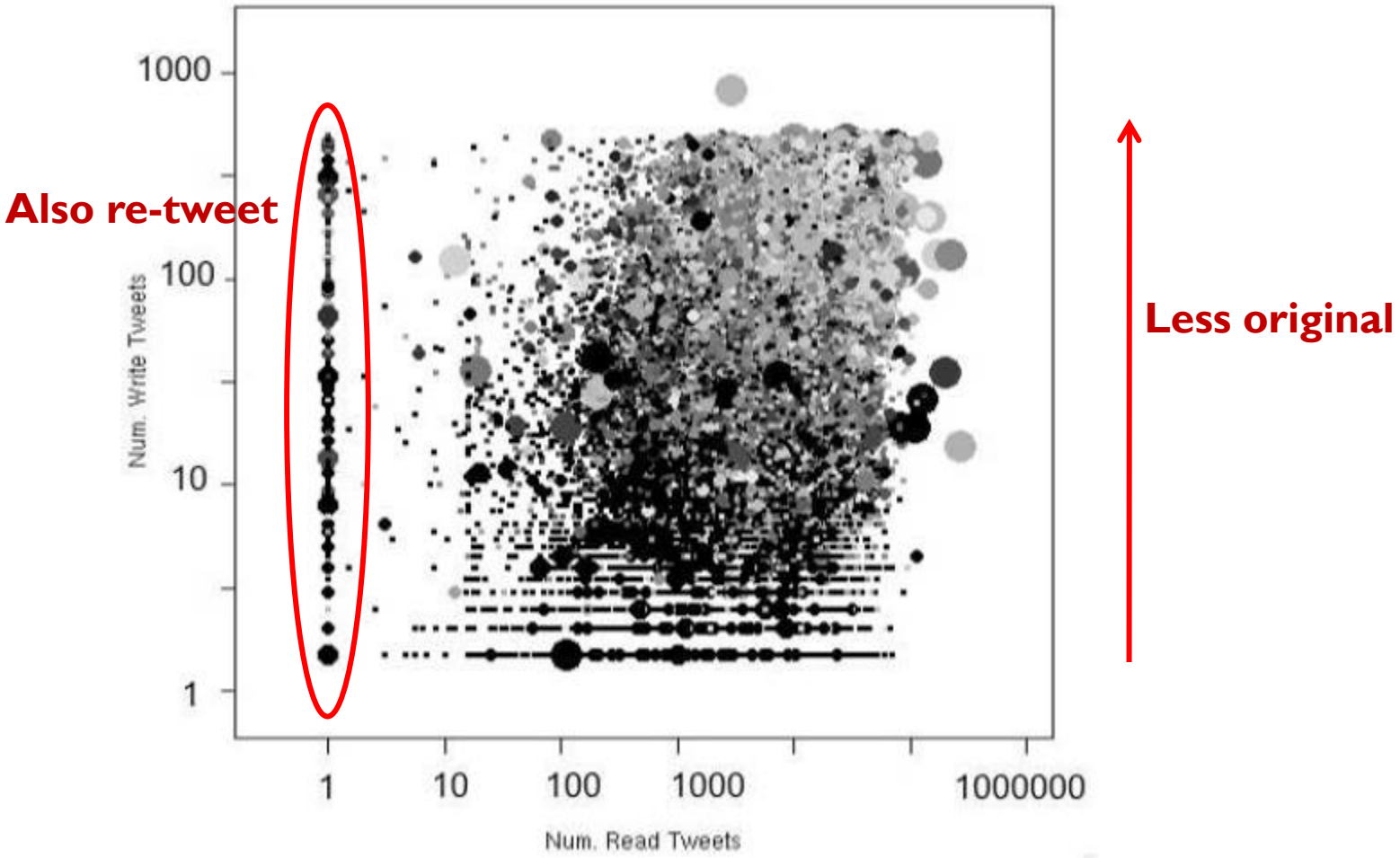
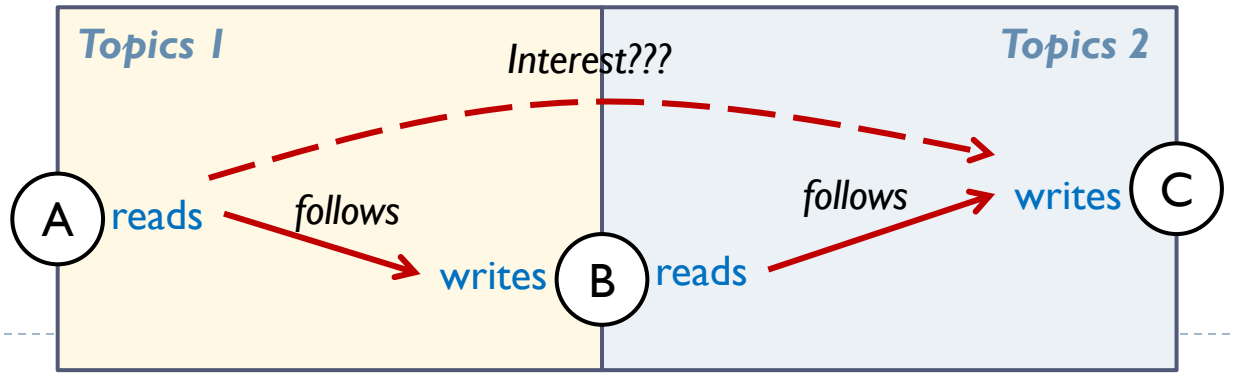


Figure 4: Reading, Writing, PageRank, and Originality

Link Semantics

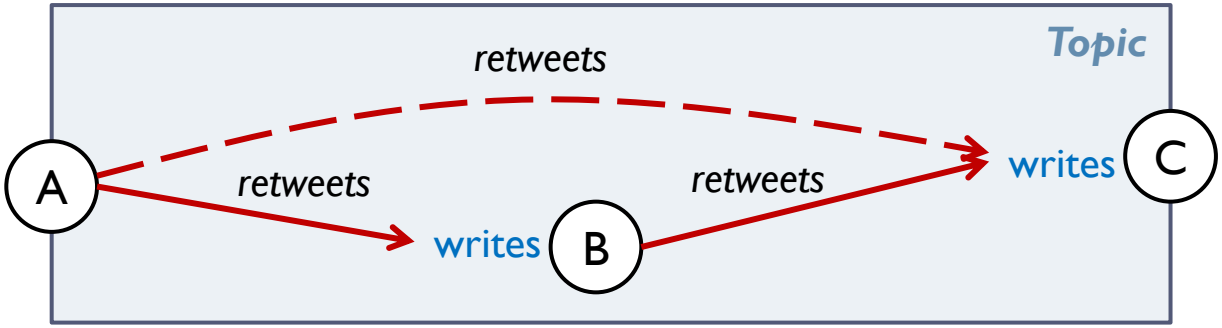
Link Semantics

- What do links in Twitter mean?
- On the web: link from page A to page B
  - Endorsement of quality of B
  - Relevance of B to A
- In Twitter: user A follows user B
  - Endorsement of quality of/interest in user B
  - Also: A as a reader is interested in B as a writer
    - Is this relationship transitive? Is topic preserved?



Link Semantics

- User A re-tweets user B
  - Endorsement of quality of/interest in user B
  - A is interested in writing about what B wrote
    - A as a writer is interested in B as a writer
    - Better transitivity, better preservation of topic



Ranking: follow-based vs. retweet-based

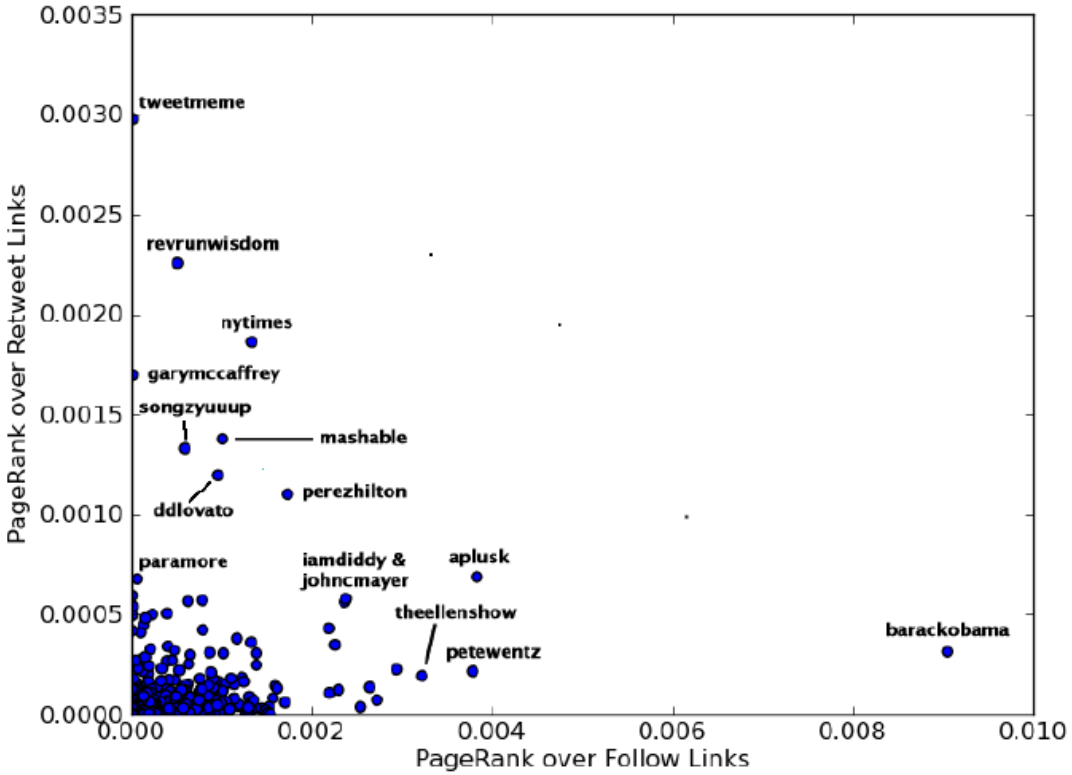
- PageRank computed over
  - Follow-graph
  - Retweet-graph

username	Follow-based	Retweet-based
barackobama	1	32
aplusk	2	9
petewentz	3	54
theellenshow	4	57
the_real_shaq	5	51
mrskutcher	7	87
johncmayer	9	12
iamdiddy	10	15

Figure 7: Top 10 Follow-based and Top 100 Retweet-based

username	Retweet-based	Follow-based
nytimes	3	30
mashable	5	60
ddlovato	7	72
perezhilton	8	15
aplusk	9	2

Figure 8: Top 10 Retweet-based and Top 100 Follow-based

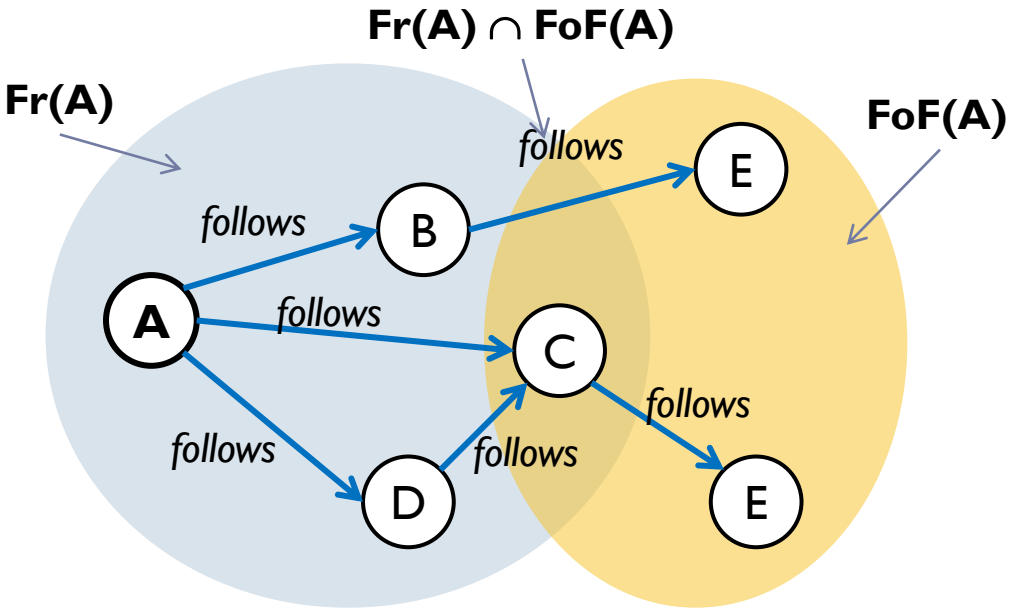


Ranking: follow-based vs. retweet-based

- Empirical analysis of the two rankings:
  - Follow links capture the quality of a user being popular or well known
  - Re-tweet links capture the quality of being influential or producing newsworthy/topically relevant posts

Link "Virality"

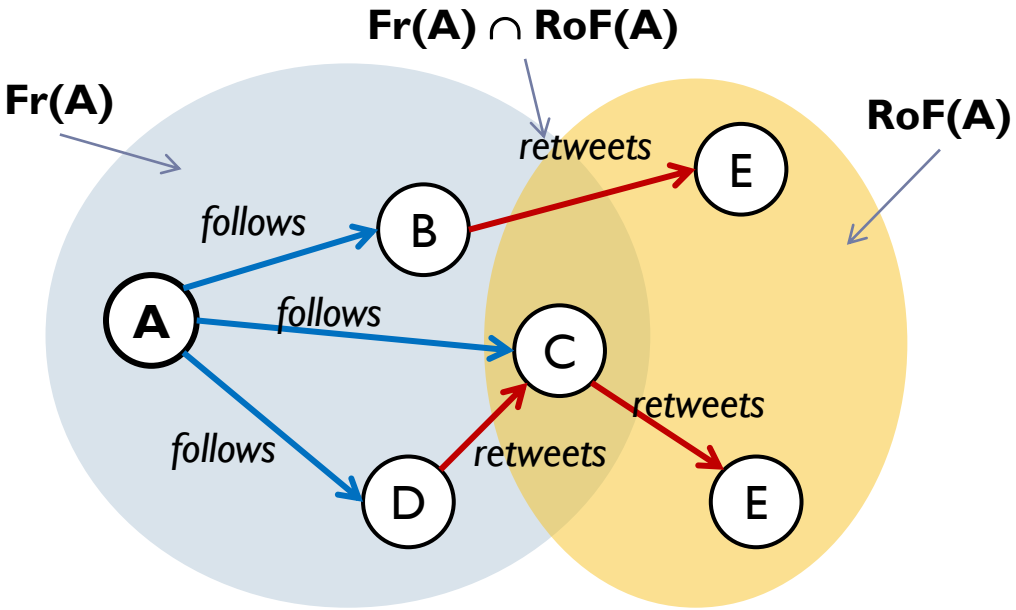
- Follow virality: $f_v(u) = \frac{FoF(u) \cap Fr(u)}{FoF(u)}$
  - Fr(u): users followed by u
  - FoF(u): 'friends of friends', users followed by Fr(u)



- *Probability that a follower of user u_a is following user u_b, given that u_a follows u_b*

Link "Virality"

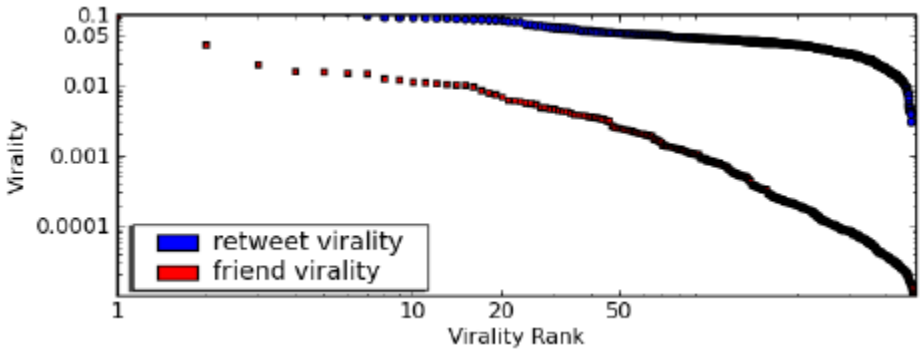
- Re-tweet virality: $r_v(u) = \frac{RoF(u) \cap Fr(u)}{RoF(u)}$
  - Fr(u): users followed by u
  - RoF(u): users retweeted by Fr(u)



- *Probability that a follower of user u_a is following user u_b, given that u_a retweeted a post from u_b*

Link "Virality"

- Retweet virality vs. Follow virality



- Possible conclusion:
  - Users are more likely to follow people they see retweeted than those who are merely "Friends of Friends".

Experiments and Results

Experiments

- Dataset
  - 1.1 million users
    - 273 million follow edges
    - 2.9 million re-tweet edges
  - October 2009 - January 2010

Experiments

- Use topic sensitive PageRank
  - Rank users relevant for a particular topic
  - Study difference in topical relevance carried by follow and retweet links
- Steps
  1. List of seed users for a given topic
     - 9 topical lists from listorious.com (avg. 155 users each)
  2. Compute PageRank scores
     - Follow graph, retweet graph
  3. Evaluate high-ranking users for topical relevance
     - 30 highest-ranking non-seed users
     - User survey (binary judgement of relevance)

Experiments

- Precision and Relevance of Top-ranked Users

Link	Precision	Relevance
Follow	0.451	0.548
Retweet	**0.601**	0.704

Table 2: Precision and Relevance by Link Type

$$\text{Precision}(U) = \frac{1}{k} \sum_{k} \frac{|R_k(U) \cap U|}{|U|}$$

$$\text{Relevance}(U) = \frac{|\bigcup_k R_k(U)|}{|U|}$$

- Precision improved by over 30% by using retweet links

Topical relevance vs. popularity

- Observations
  - Retweet links → more topically relevant users
    - But have fewer followers than those discovered by follow links
      - Relevant follow-based users: avg. number of followers 257, 088
      - Relevant retweet-based users: avg. number of followers 75, 851
    - Number of followers a user has is not directly related to their relevance for a particular topic

Conclusions

- Link semantics
  - Follow links, even from a set of topically similar users, quickly diffuse into a broad range of topics
  - Retweet links, meanwhile, remain more concentrated on the original topic
- Importance for topic-sensitive ranking:
  - Propagating a user's topical relevance over links is not trivial
  - Different link types produce significantly different results

Summary

Summary

- Graph model of Twitter
- Link types and their properties
- Significance of link types for topic preservation
  - Propose retweet links as an alternative source of information

- Open questions:
  - How to model other types of links?
    - @-links (tweet → user)
    - URLs (tweet → website)
    - #tags (tweet → tag)
  - What are their semantics? How can we use them?
  - General framework for topic propagation in the graph?