

What is Twitter, a Social Network or a News Media?

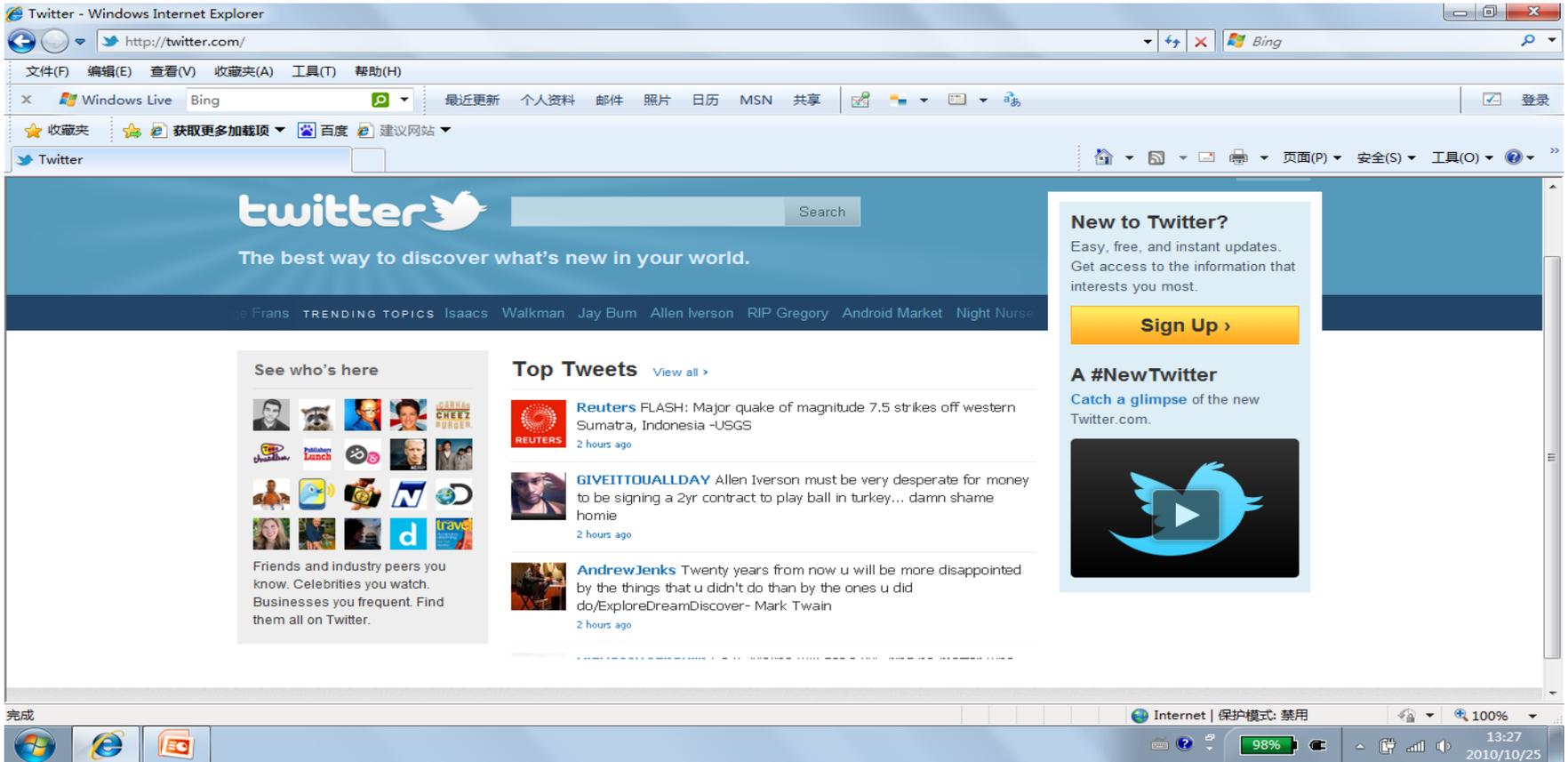
Paper Authors: H. Kwak, C. Lee, H.Park, S.Moon
– Department of Computer Science, Korea
Advanced Institute of Science and Technology

Presented by: Jidong Xiao

Overview

- This presentation contains
 - Background and Motivation
 - Data Collection
 - Data Analysis
 - Topological Analysis
 - Ranking Twitter Users
 - Trending the Trends
 - Impact of Retweet
 - Conclusion

Background



- Twitter: a website which offers a social networking and microblogging service, enabling its users to send and read other users' messages called tweets.

Background

- Tweet: text-based posts of up to 140 characters displayed on the user's profile page. Users may subscribe to other authors' tweets—this is known as following and subscribers are known as followers.
- Retweet: To post a copy of someone else's tweet on your own account. Similar to “forward” in email system.
- Trending Topics: the most common phrases appearing in messages.

Motivation

- Study the topological characteristics of Twitter and its power as a new medium of information sharing.
- How are people connected on Twitter? Who are the most influential people? What do people talk about? How does information diffuse via retweet?

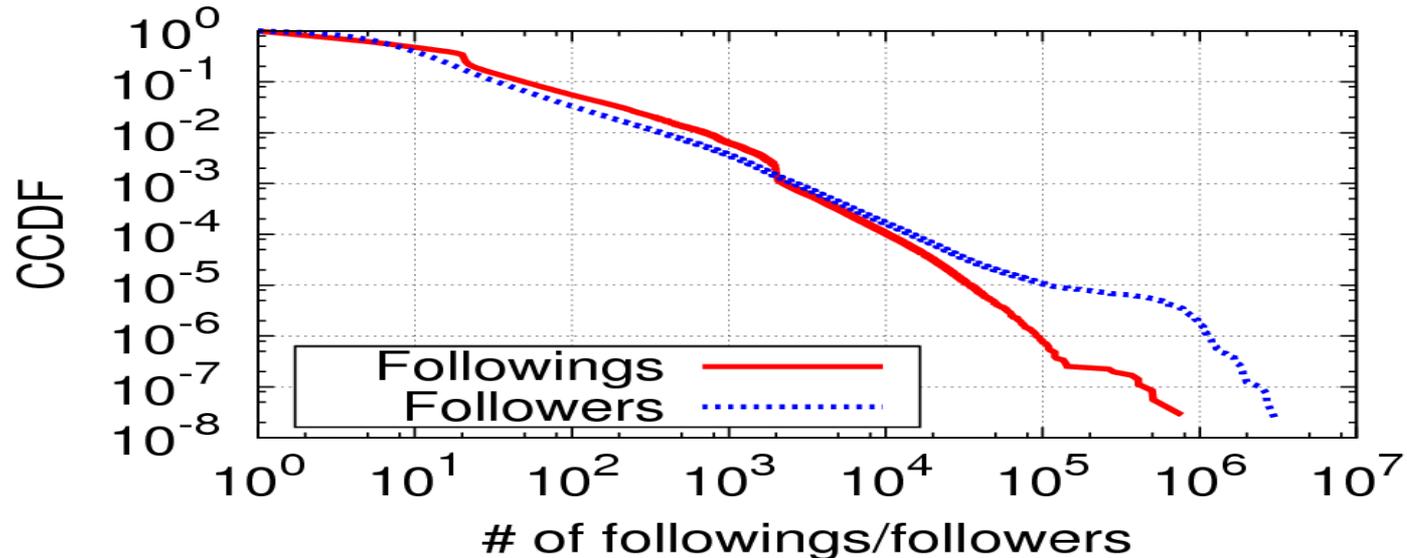
Data Collection

- Twitter offers an API that is easy to crawl and collect data.
- Starting from June, 2009, The authors crawled and collected profiles of all users on Twitter. The final number of user profiles they collected is 41.7 million.
- In addition, they collected profiles of users who mentioned trending topics from June 3rd to September 24th, 2009, and they collected 4,262 unique trending topics and their tweets. The final number of collected tweets is 106 millions.

Data Analysis

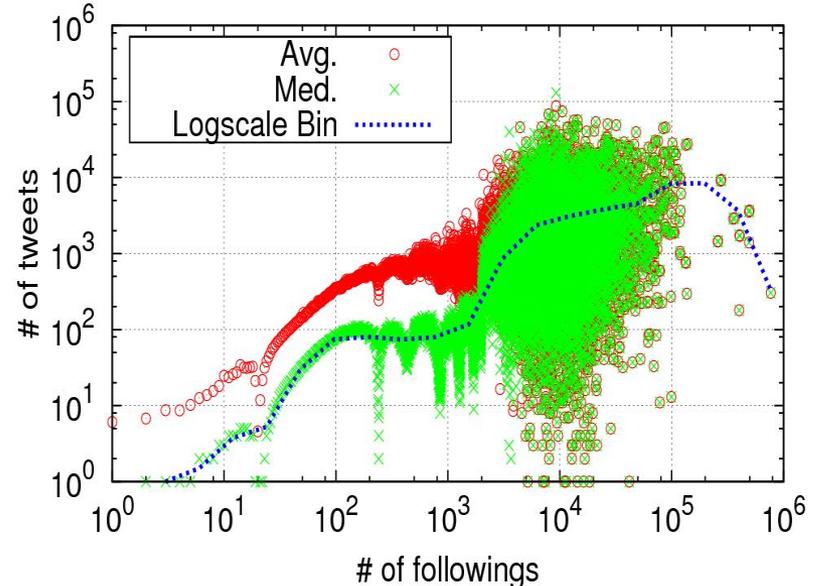
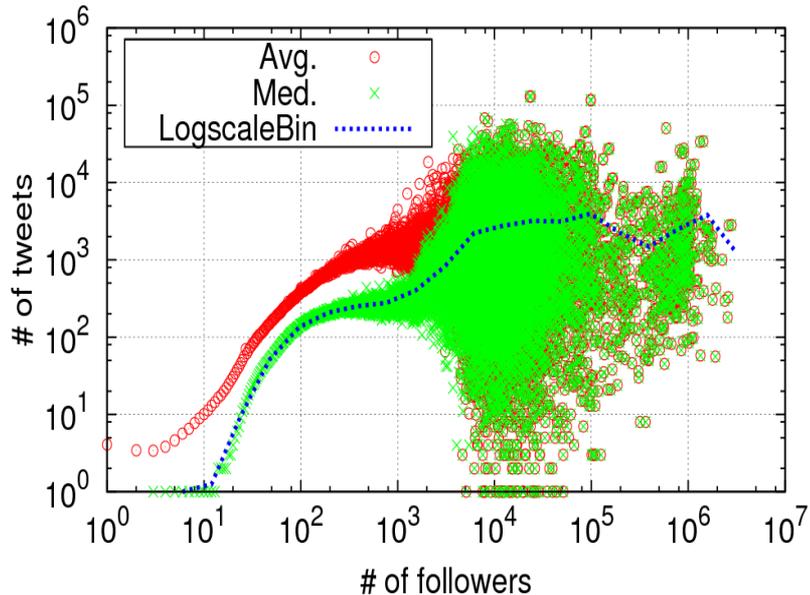
- **Topological Analysis**
- Ranking Twitter Users
- Trending the Trends
- Impact of Retweet

Topological Analysis - Basic Analysis



- The distribution of the number of followings and that of followers.
- A very small number of users follow more than 10,000.
- Only 40 users with more than a million followers and all of them are either celebrities or mass media.
- The dashed line in the figure up to $x=10^5$ fits to a power law distribution with the exponent of 2.276. The data points beyond $x=10^5$ represent users who have many more followers than the power-law distribution predicts.

Followers vs. Tweets, Followings vs. Tweets



- The dashed line shows a positive trend.
- Both figures demonstrate the median number of tweets increase up to $x=100$ and remains relatively flat up till $x=1000$, then beyond $x=5000$, the number of tweets increase by an order of magnitude or more.

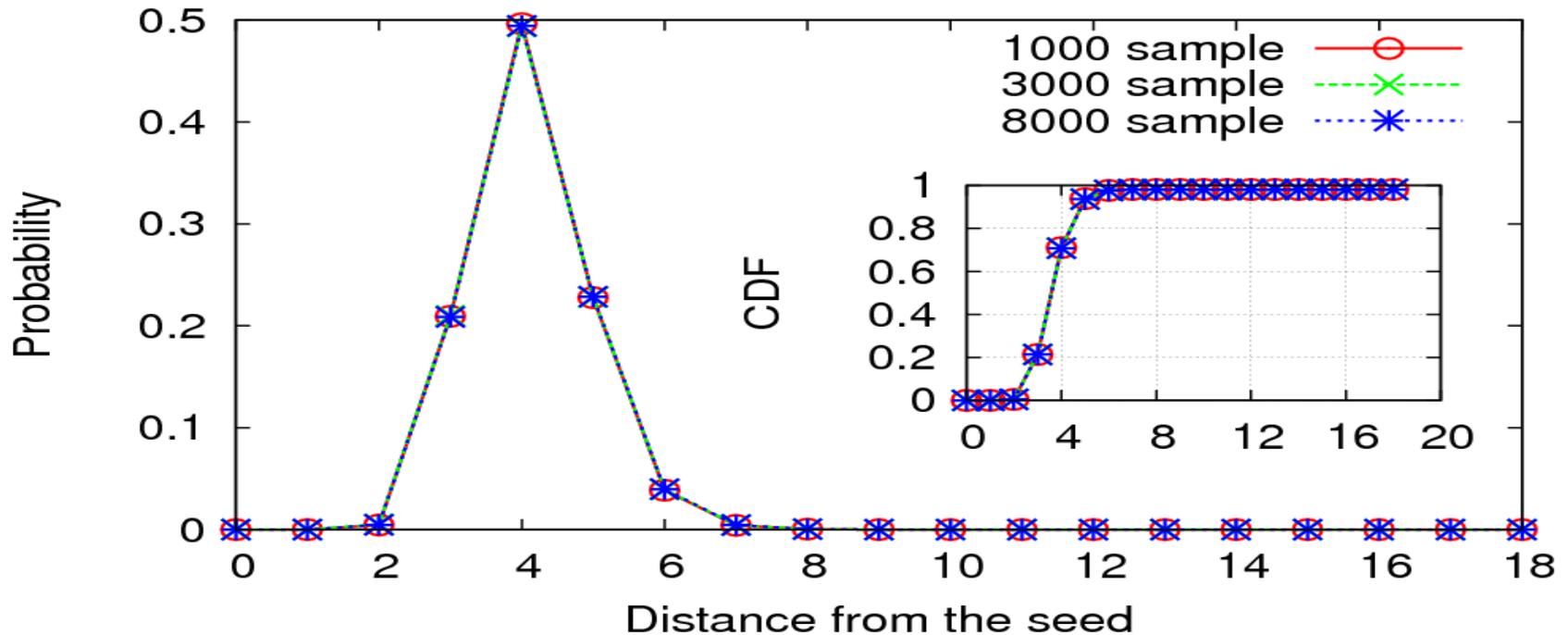
Reciprocity

- Previous studies have reported high reciprocity on other social network services: 68% on Flickr and 84% on Yahoo!360.
- Twitter shows a low level of reciprocity; 77.9% of user pairs with any link between them are connected one-way, and only 22.1% have reciprocal relationship between them.
- Moreover, 67.6% of users are not followed by any of their followings in Twitter.

Degree of Separation

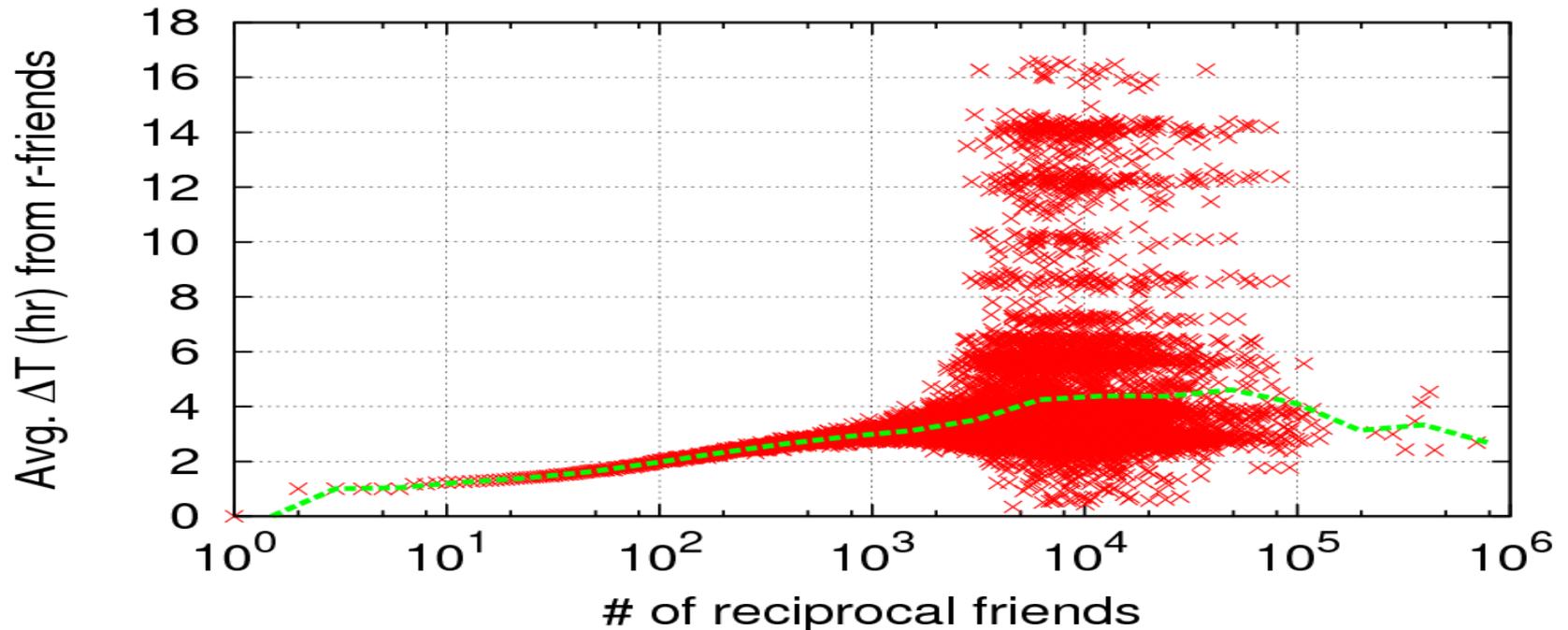
- Previous study shows any two people could be connected on average within six hops from each other.
- Recent report on MSN messenger network shows that the median and the 90% degrees of separation are 6. and 7.8, respectively.
- As only 22.1% of user pairs are reciprocal in twitter, the authors expect the average path length between two users in Twitter to be longer than other known networks.
- However, their evaluation results in next slide goes in the opposite direction of their expectation.

Degree of Separation



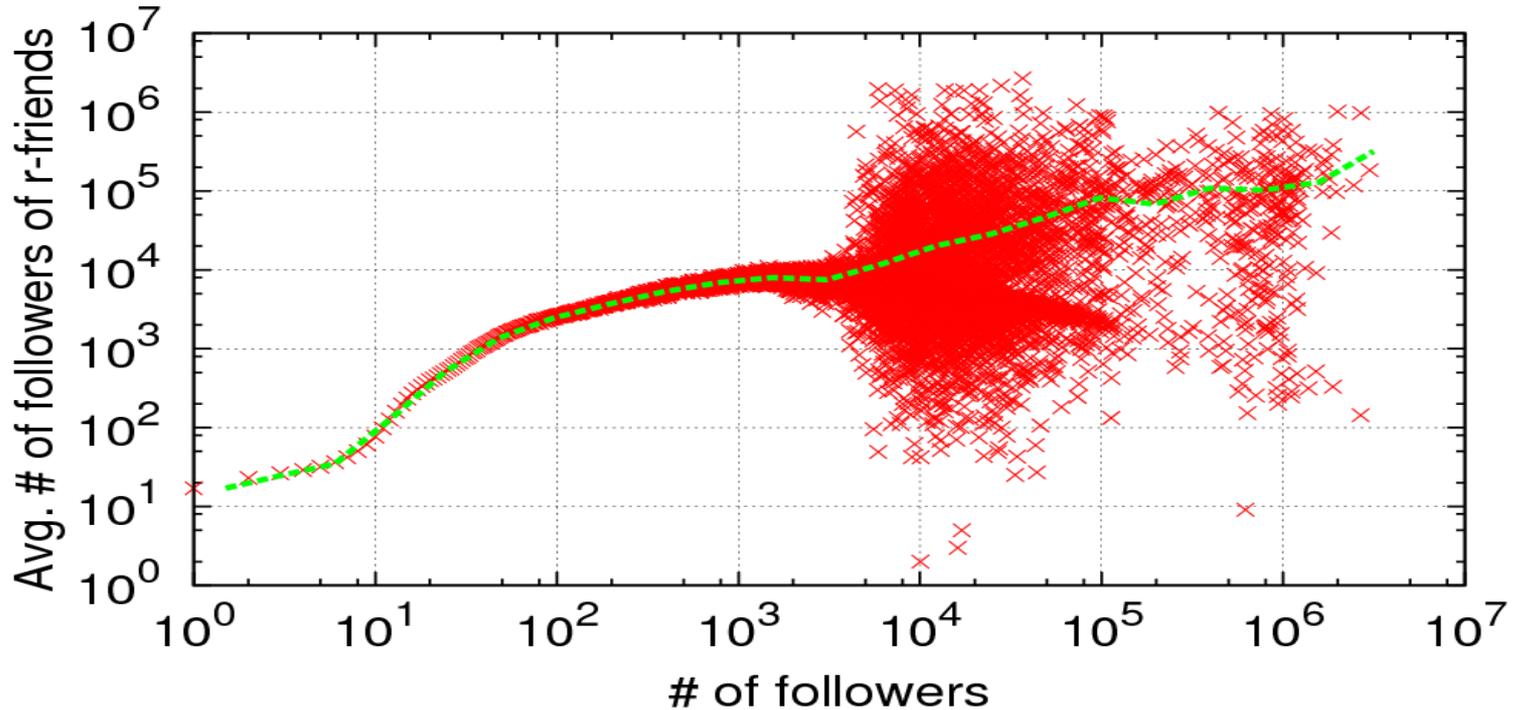
- The median of the distribution is 4. The average path length is 4.12. The effective diameter (i.e., the 90th percentile distance) is 4.8. For 70.5% of node pairs, the path length is 4 or shorter, and for 97.6% it is 6 or shorter.
- The authors say this interesting phenomenon may indicate that people follow others not only for social networking, but for information.

Homophily



- Users choose one of the 24 time zones around the world. And the authors calculate the average time differences between a user and reciprocal friends.
- The above Figure shows for those with 2,000 reciprocal friends or fewer, the median time difference stays below 3 hours. The authors conclude that these users and their reciprocal friends are likely to be geographically close.

Homophily (cont'd)



- Does a user of certain popularity follow other users of similar popularity and they reciprocate?
- The above Figure shows a positive correlation slightly below $x=1,000$ and dispersion beyond that point. Which means users with followers 1,000 or less are likely to have similar popularity with their reciprocate friends.

Data Analysis

- Topological Analysis
- **Ranking Twitter Users**
- Trending the Trends
- Impact of Retweet

Ranking Twitter Users

- By the Number of Followers
- By PageRank (The algorithm which google use to rank web pages in their search results)
- By the Retweets

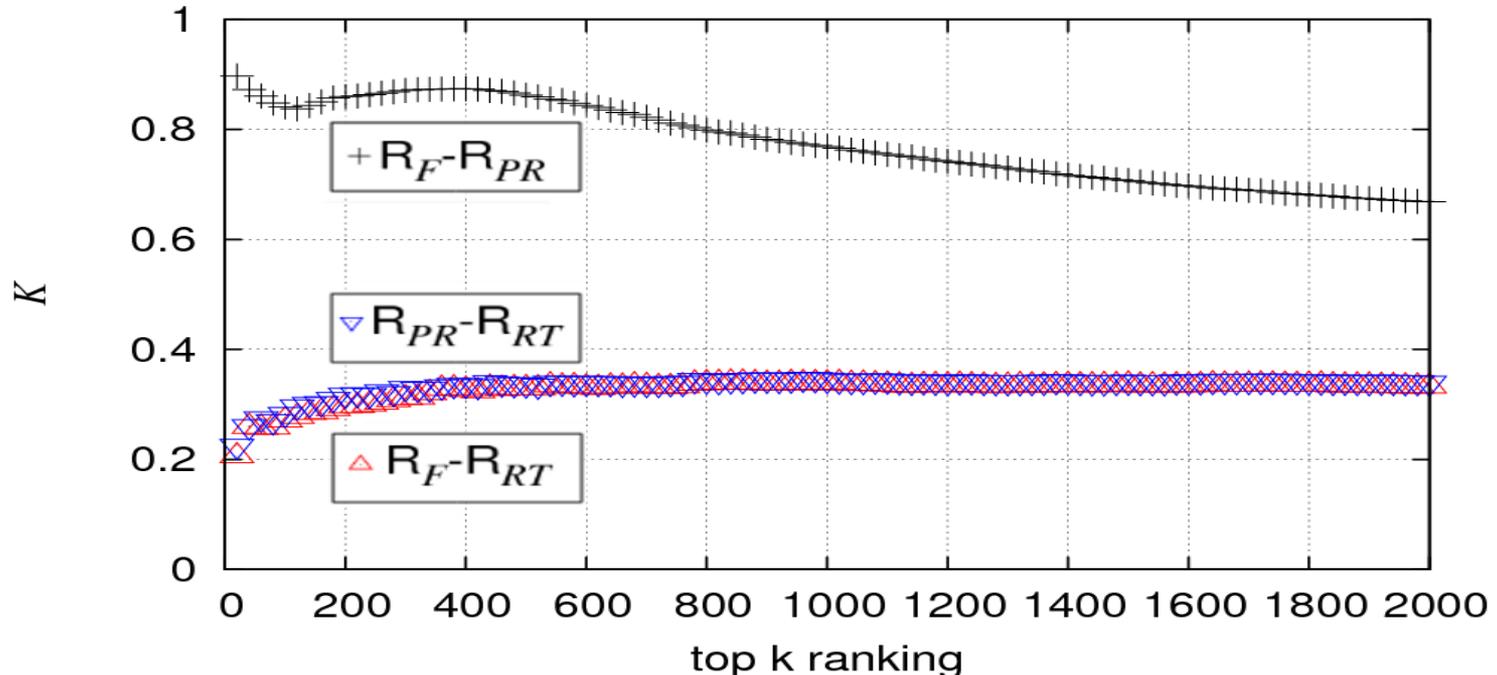
Ranking Twitter Users

Rank	Ranking by # of followers			Ranking by PageRank in the following/follower network			Ranking by # of retweet in the diffusion network		
	ID	Name	Remark	ID	Name	Remark	ID	Name	Remark
1	aplusk	ashton kutcher	actor	aplusk	ashton kutcher	actor	mashable	Pete Cashmore	news on social media
2	britneyspears	Britney Spears	musician	BarackObama	Barack Obama	president of U.S.	BreakingNews	BNO News	news
3	TheEllenShow	Ellen DeGeneres	show host	cnnbrk	CNN Breaking News	news	tweetmeme	TweetMeme	news on Twitter
4	cnnbrk	CNN Breaking News	news	TheEllenShow	Ellen DeGeneres	show host	oxfordgirl	oxfordgirl	journalist
5	Oprah	Oprah Winfrey	show host	britneyspears	Britney Spears	musician	cnnbrk	CNN Breaking News	news
6	twitter	Twitter	subject of this paper	Oprah	Oprah Winfrey	show host	TechCrunch	Michael Arrington	news on technology
7	BarackObama	Barack Obama	president of U.S.	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	myfabolouslife	Fabulous	musician
8	RyanSeacrest	Ryan Seacrest	show host	johncmayer	John Mayer	musician	nytimes	The New York Times	news
9	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	twitter	Twitter	subject of this paper	lilduval	lil duval	comedian
10	KimKardashian	Kim Kardashian	model	RyanSeacrest	Ryan Seacrest	show host	IranRiggedElect	Iran	about Iran
11	johncmayer	John Mayer	musician	lancearmstrong	Lance Armstrong	sports star	espn	ESPN Sports News	news
12	mrskutcher	Demi Moore	actress	jimmyfallon	Jimmy Fallon	actor	persiankiwi	persiankiwi	about Iran
13	iamdiddy	iamdiddy	musician	iamdiddy	iamdiddy	musician	aplusk	ashton kutcher	actor
14	jimmyfallon	Jimmy Fallon	actor	mrskutcher	Demi Moore	actress	StopAhmadi	Raymond Jahan	about Iran
15	lancearmstrong	Lance Armstrong	sports star	PerezHilton	Perez Hilton	power blogger	Alyssa_Milano	Alyssa Milano	actress
16	alгоре	Al Gore	politician	nytimes	The New York Times	news	huffingtonpost	HuffingtonPost.com	news
17	miley Cyrus	Miley Cyrus	actress / musician	miley Cyrus	Miley Cyrus	actress / musician	iamdiddy	iamdiddy	musician
18	nytimes	The New York Times	news	stephenfry	Stephen Fry	actor	iranbaan	Fershteh Ghazi	about Iran
19	coldplay	Coldplay	musician	TheOnion	The Onion	news	nprnews	NPR News	news
20	TheOnion	The Onion	news	KimKardashian	Kim Kardashian	model	PerezHilton	Perez Hilton	power blogger

Ranking Twitter Users

- The top 20 users ranked by PageRank are almost the same to the top 20 users ranked by the number of followers. Only two differences.
- Only 4 out of 20 users are common in all three lists.
- There are mainstream news media and independent news media rise in ranking by the retweets, which speaks that followers of these media think that tweets of these media are worth propagating.

Comparison among Rankings



- The authors present a quantitative comparison between the three ranking. In their algorithm, $K=1$ means complete agreement while $K=0$ means complete disagreement. The result in this figure shows, Rank by the number of followers and Rank by PageRank are similar, while Rank by retweets is different, which indicates a gap between the number of followers and the popularity of one's tweets.

Data Analysis

- Topological Analysis
- Ranking Twitter Users
- **Trending the Trends**
- Impact of Retweet

Comparison with Trends in Other Media

- The authors have collected top 40 search keywords per day from Google Trend during the same period as their Twitter data collection, also collected CNN Headline News of the same period.
- Only 126 (3.6%) out of 3,479 unique trending topics from Twitters exist in 4,597 unique hot keywords from Google.
- On average 95% of topics each day are new in Google while only 72% of topics are new in Twitter.
- More than half the time CNN was ahead in reporting, but some news broke out on Twitter before CNN and they are of live broadcasting nature (e.g. Sports matches and accidents).

Singleton, Reply, Mention, and Retweet

- Singleton: if a tweet has no reply or retweet, then we call it singleton.
- Mention: People often write a tweet addressing a specific user, we call such a tweet a mention.
- Among all tweets mentioning those trending topics, singleton are most common, followed by replies and retweets. Mentions are least common in tweets.

User Participation in Trending Topics

- Out of 41 million Twitter users, a large number of users (8,262,545) participated in trending topics and about 15% of those users participated in more than 10 topics during four months.

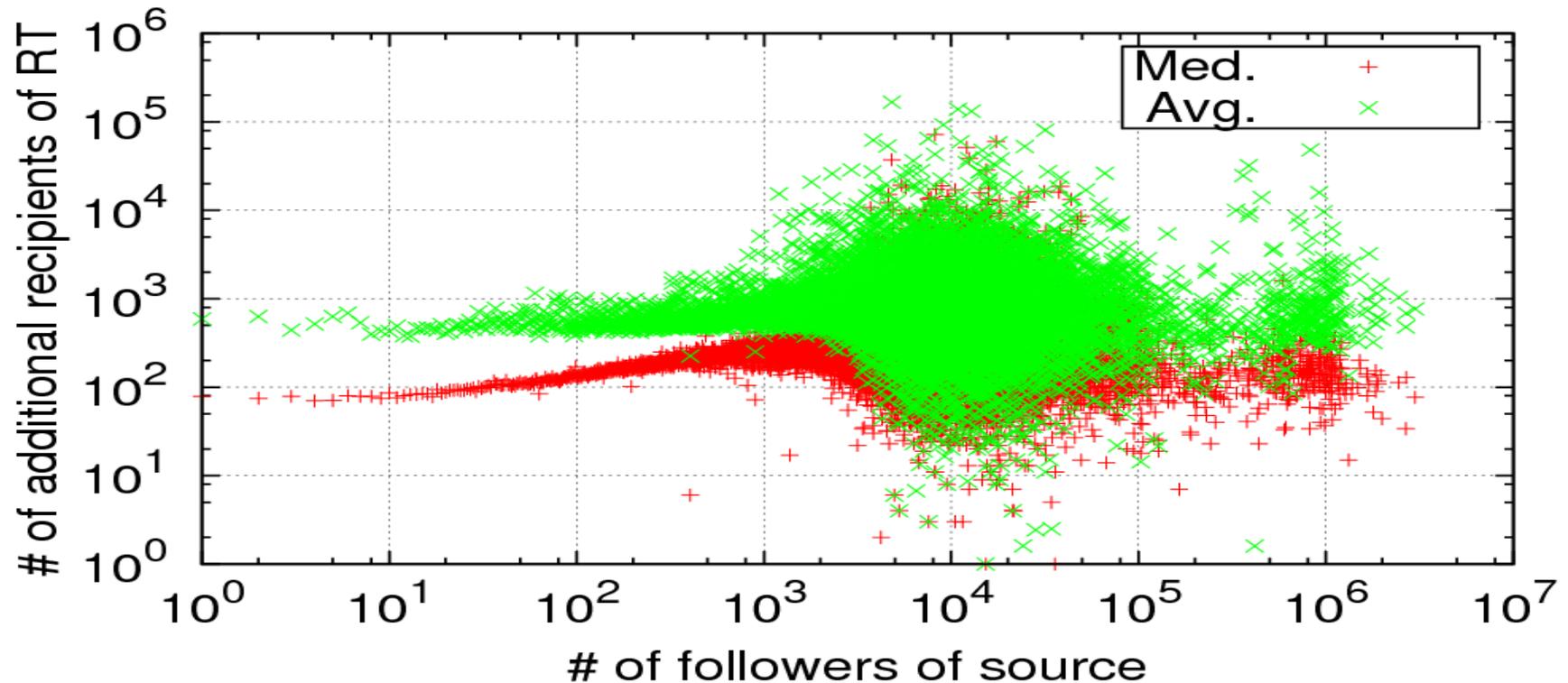
Active Period of Trends

- They consider a trending topic inactive if there is no tweet on the topic for 24 hours.
- 73% topics have a single active period. About 15% of topics have 2 active periods and 5% have 3. Very few have more than 3 active periods.
- Most of the active periods are a week or shorter. 31% of periods are 1 day long, and only 7% of periods are longer than 10 days.
- When classifying the trending topics, the result reveals that half of the trending topics are timely breaking news.

Data Analysis

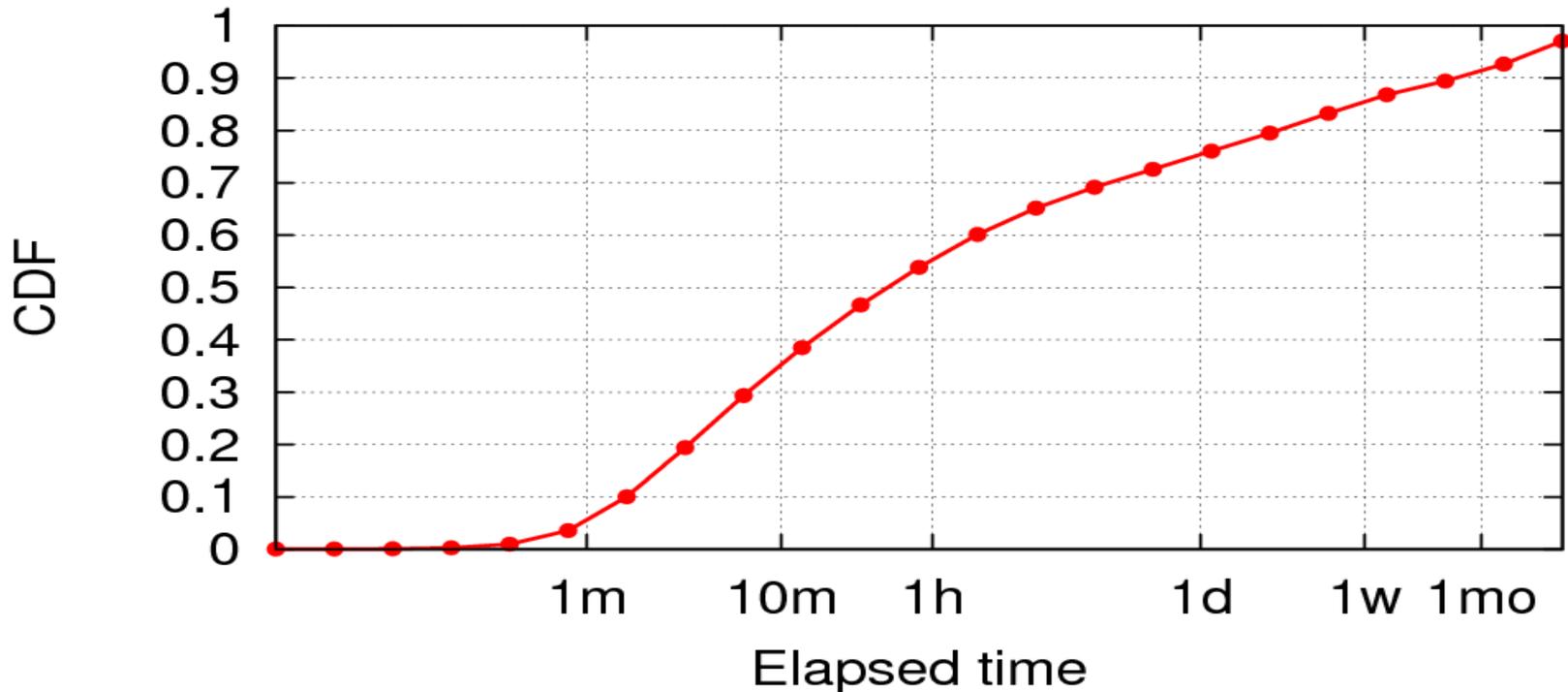
- Topological Analysis
- Ranking Twitter Users
- Trending the Trends
- **Impact of Retweet**

Audience Size of Retweet



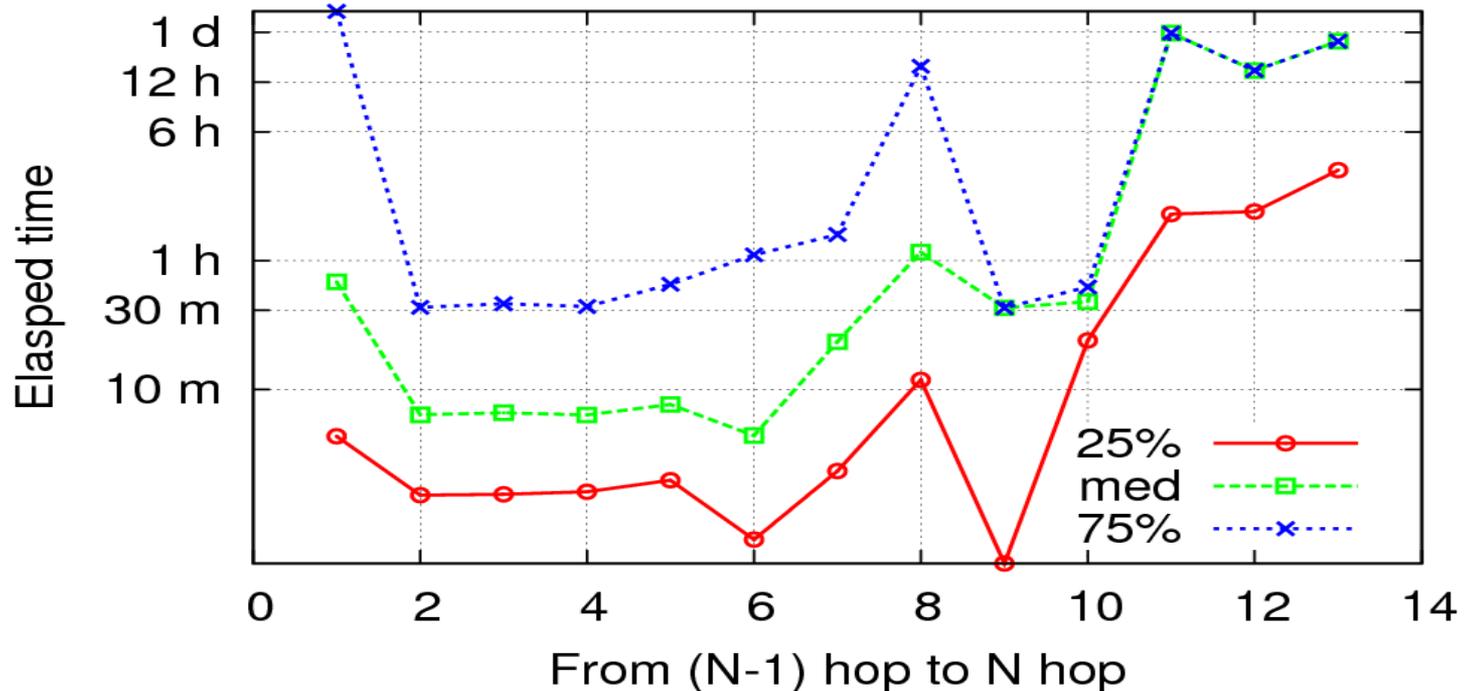
- Up to about 1,000 followers, the average number of additional recipients (those who are not immediate followers of the original tweet owner) is not affected by the number of followers of the tweet source. That is, no matter how many followers a user has, the tweet is likely to reach a certain number of audience, once the user's tweet starts spreading via retweets.

Temporal Analysis of Retweet



- The time lag from a tweet to its retweet: Half of the retweeting occurs within an hour, and 75% under a day. However, 10% of retweets take place a month later.

Temporal Analysis of Retweet (Cont'd)



- The time lag between two nodes on a retweet tree: What is interesting is from the second hop and on the retweets are much more responsive and basically occur back to back up to 5 hops away.

Conclusion

- In twitter's follower-following topology analysis we have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks.
- Among reciprocated users we observe some level of homophily.
- We have ranked users by the number of followers and by PageRank and found two rankings to be similar. But if we rank by the number of retweets, then the ranking differs from the previous two rankings, indicating a gap in influence inferred from the number of followers and that from the popularity of one's tweets.
- We have analyzed the tweets of top trending topics and reported on the temporal behavior of trending topics and user participation.
- Any retweeted tweet is to reach an average of 1,000 users no matter what number of followers is of the original tweet. Once retweeted, a tweet gets retweeted instantly on the 2nd, 3rd, and 4th hops away from the sources, signifying fast diffusion of information after the 1st retweet.

Questions?