Understanding and Combating Link Farming in the Twitter social network

Complex Network Research Group
Department of CSE, IIT Kharagpur, India

Networked Systems Research Group
Max Planck Institute for Software Systems, Germany

Link farming: a prevalent evil in Web

- Search engines rank websites / webpages based on graph metrics such as Pagerank
  - High in-degree helps to get high Pagerank
- Link farming in Web
  - Websites exchange reciprocal links with other sites to improve ranking by search engines
- A form of spam – heavily penalized by search engines

Why link farming in Twitter?

- Twitter has become a Web within the Web
  - Vast amounts of information and real-time news
  - Twitter search becoming more and more common
  - Search engines rank users by follower-rank, Pagerank to decide whose tweets to return as search results
  - High indegree (#followers) seen as a metric of influence

- Link farming in Twitter
  - Spammers follow other users and attempt to get them to follow back

Link farming in Web & Twitter similar?

- Motivation is similar
  - Higher indegree will give better ranks in search results
- Who engages in link farming?
  - Web – spammers
  - Twitter – spammers + many legitimate, popular users !!!
- Additional factors in Twitter
  - ‘Following back’ considered a social etiquette

- Is link farming in Twitter spam at all?

How to identify link farmers in Twitter?

- Idea: start with spammers
  - Study how spammers acquire social links

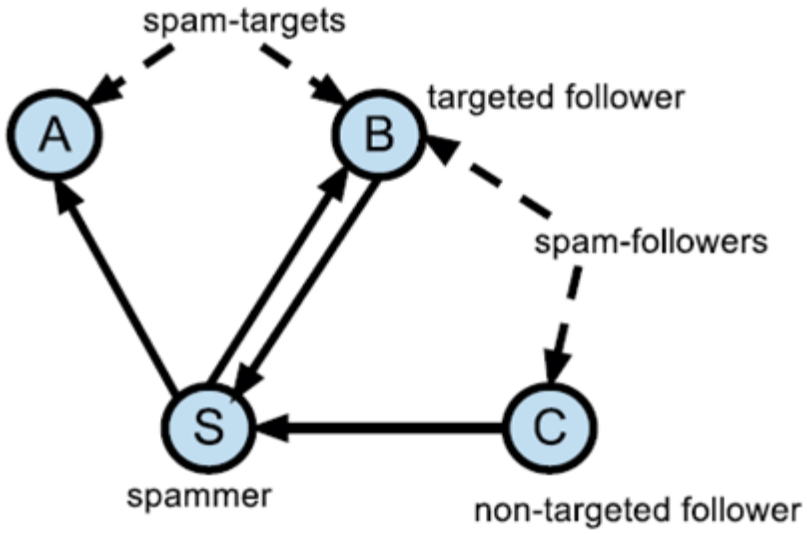
- Reported: large amounts of spam exist in Twitter
  - Spam-URLs in Twitter get much higher clickthrough rates than spam-URLs in email [Grier, CCS 2011]
  - Shows spammers are successfully acquiring social links and social influence

Large scale identification of spammers

- Twitter dataset collected at MPI-SWS, Germany
  - Complete snapshot of Twitter as of August 2009
  - 54 million users, 1.9 billion social links

- Identifying spammers
  - 379,340 accounts suspended during Aug 2009 – Feb 2011
  - Suspension is due to spam-activity or long inactivity
  - 41,352 suspended accounts posted at least one blacklisted URL shortened by bit.ly or tinyurl → spammers

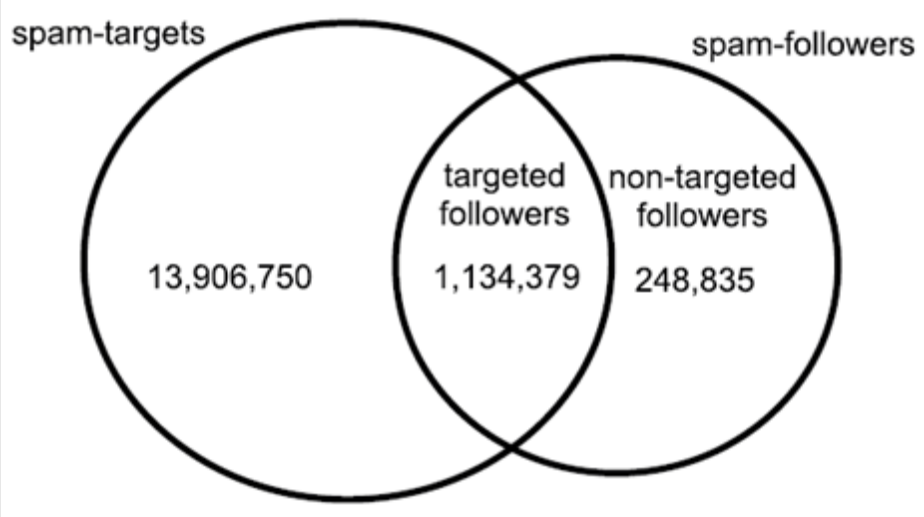
Terminology for spammers' links



- Spam-targets: users followed by spammers
- Spam-followers: users who follow spammers
  - Targeted: spam-target and spam-follower
  - Non-targeted: follow spammers without being targeted

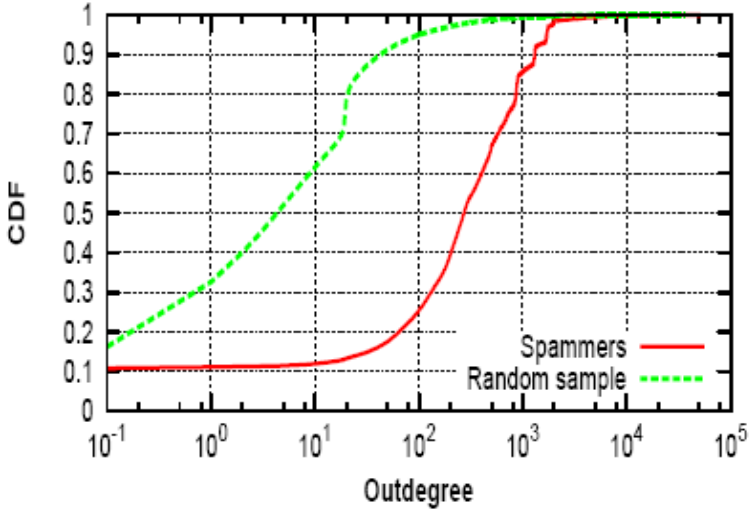
Link farming by spammers

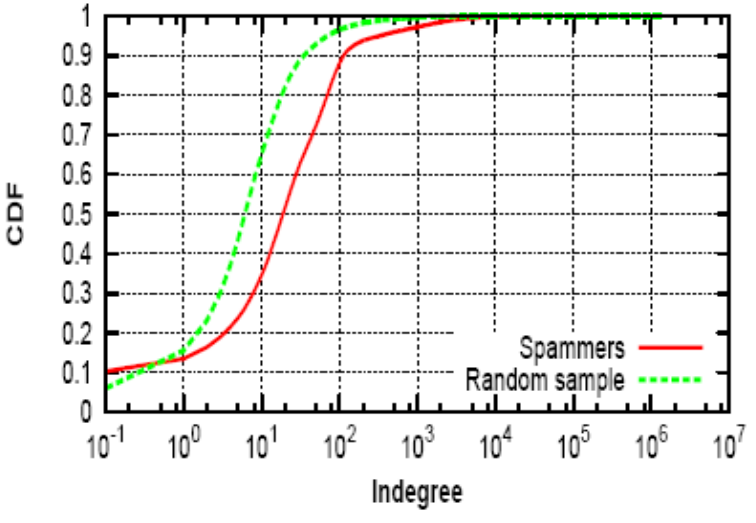
- Spammers farm links at large scale
  - Over 15 million users (27% of total) targeted by 41,352 spammers (0.08% of total)
- 1.3 million spam-followers
  - 82% are targeted → spammers get most links by reciprocation



Link farming makes spammers influential

- Spammers get more followers than an average Twitter user
- Some spammers acquire very high Pageranks
  - 304 within top 100,000 (0.18% of all users)



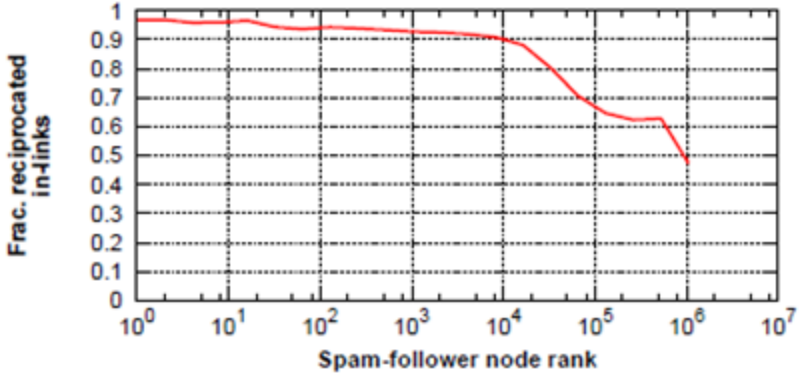


Who are the spam-followers?

- Non-targeted spam-followers
  - Mostly sybils / hired helps of spammers
  - Most have now been suspended by Twitter

- Targeted spam-followers
  - Ranked on the basis of number of links to spammers
  - 60% of follow-links acquired by spammers come from the top 100,000 targeted followers

Top spam-followers tend to reciprocate almost all links established to them by spammers



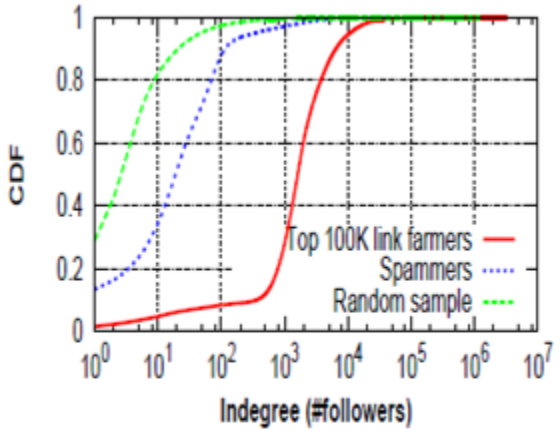
Is it easy to farm links in Twitter?

- We created a Twitter account and followed some of the top targeted spam-followers
  - Followed 500 randomly selected users out of the top 100K spam-followers
  - Within 3 days, 65 reciprocated by following back
  - Our account ranked within the top 9% of all users in Twitter in 3 days !!!

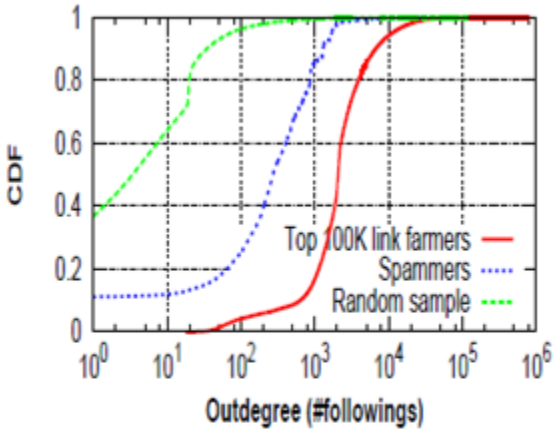
- Existence of a set of users from whom social links (hence social influence) can be farmed easily
  - Referred to as the top link-farmers

Who are the top link-farmers?

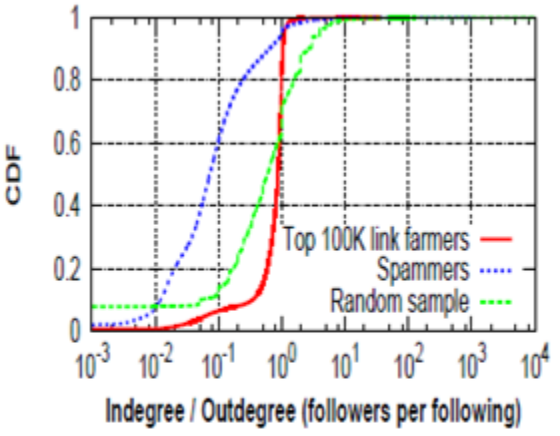
- Not spammers themselves
  - 76% not suspended by Twitter in the last two years
  - 235 verified by Twitter to be real, well-known users
  - Have much higher indegree as well as outdegree compared to spammers
  - Most of their tweets contain valid URLs



(a) Indegree



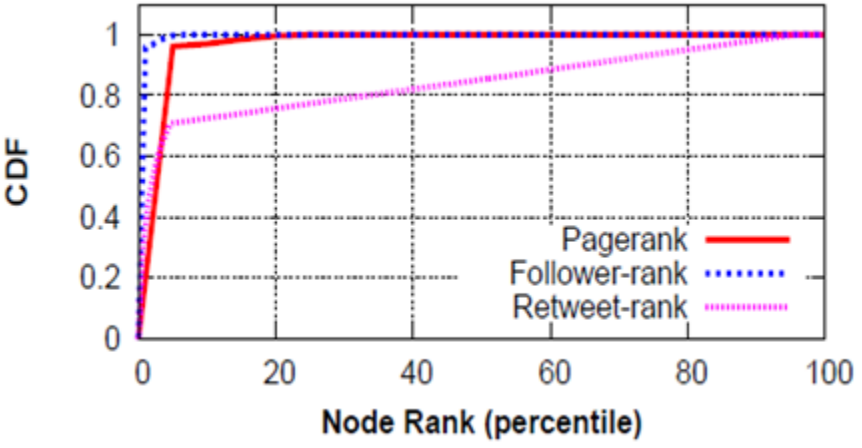
(b) Outdegree



(c) Indegree/outdegree ratio

Who are the top link-farmers?

- Highly influential users
  - Rank within top 5% according to Pagerank, follower-rank, retweet-rank
- Mostly social marketers, entrepreneurs, ...
  - Want to promote some online business / website
  - Heavily interconnect with each other – density of subgraph is 0.018 (for whole graph: 10^{-7})
  - Aim: to acquire social capital



Collusionrank

Algorithm 1 Collusionrank

Input: network, G; set of known spammers, S; decay factor for biased Pagerank, α

Output: Collusionrank scores, c

initialize score vector d for all nodes n in G

$$d(n) \leftarrow \begin{cases} \frac{-1}{|S|} & \text{if} n \in S \\ 0 & \text{otherwise} \end{cases}$$

/* compute Collusionrank scores */

$c \leftarrow d$

while c not converged **do**

for all nodes n in G **do**

$$tmp \leftarrow \sum_{nbr \in followings(n)} \frac{c(nbr)}{|followers(nbr)|}$$

$$c(n) \leftarrow \alpha * tmp + (1 - \alpha) * d(n)$$

end for

insert leaked scores uniformly across all nodes such that $\sum_n c(n) = -1$

end while

return c

Top link-farmers: examples

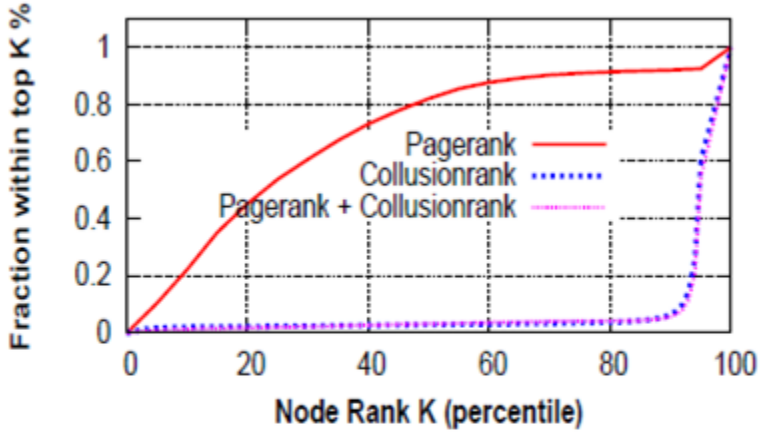
Top 5 link farmers according to	
#links to spammers	**Pagerank**
Larry Wentz: Internet, Affiliate Marketing	Barack Obama: campaign staff
Judy Rey Wasserman: Artist, founder	Britney Spears: It's Britney
Chris Latko: Interested in tech. Will follow back	NPR Politics: Political coverage and conversation
Paul Merriwether: helping others, let's talk soon	UK Prime Minister: PM's office
Aaron Lee: Social Media Manager	JetBlue Airways: Follow us and let us help

Combating the problem

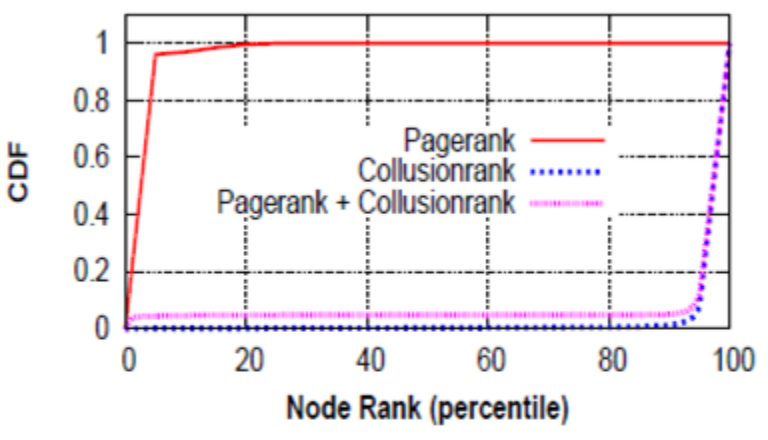
- Not practical for Twitter to suspend / blacklist top link-farmers
- Solution
  - Strategy to disincentivize users from following / reciprocating to unknown people
  - Penalize users for following spammers
- Algorithm that is inverse of Pagerank
  - Negatively bias a small set of known spammers
  - Propagate negative scores from spammers to spam-followers

Pagerank + Collusionrank

- Computed Collusionrank considering 600 known spammers
- Rank users by Pagerank + Collusionrank
  - Effectively filters out spammers and link-farmers (top spam-followers) from top ranks



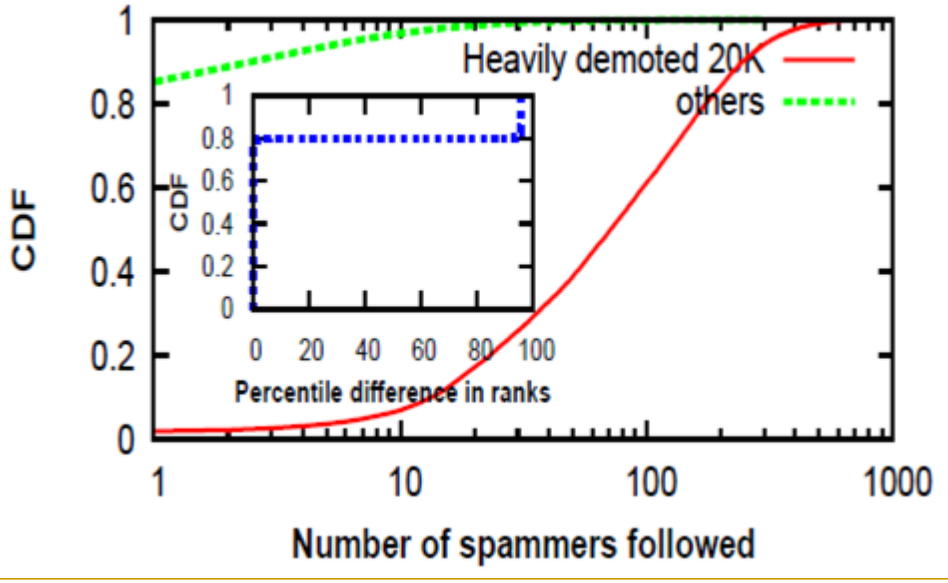
(a) Rankings of all 41,352 spammers



(b) Rankings of Top 100,000 capitalists

Pagerank + Collusionrank

- Selectively penalizes spammers & link-farmers
  - Out of top 100K according to Pagerank, 20K demoted heavily, rest 80% not affected much (inset)
  - The heavily demoted 20K follow many more spammers than the rest (main figure)



Related Publications

- Preliminary version: Poster at ACM World Wide Web Conference 2011, Hyderabad, India
- Complete study: Paper accepted at ACM World Wide Web Conference 2012, Lyon, France

Who is who in Twitter: Crowdsourcing expertise inference of Twitter users

Complex Network Research Group
Department of CSE, IIT Kharagpur

Networked Systems Research Group
Max Planck Institute for Software Systems

Motivation for who-is-who service

- Twitter has emerged as an important source of information & real-time news
- Need to know the credentials / expertise of a user to trust the content posted by her
- Knowledge of users' topical expertise can be used to identify experts in specific topics

How to know expertise of a user

- Use content provided by the user herself
  - Bio of Twitter account, tweets posted by user, ...
- Problems:
  - Many popular users do not have bio, or bio does not give topical information

Name	Bio	Major Topics obtained from List
Jimmy fallon	astrophysicists	celebs, comedy, funny, actors, famous, humor
Danecook	When I tweet, I tweet to kill	celebs, comedy, funny, actors, famous
ScreenOrigami	Web developer from Germany	Webdesign, webkraut, html, designer

How to know expertise of a user

- Use content provided by the user herself

Name	Bio	Major Topics obtained from List
Jimmy fallon	astrophysicists	celebs, comedy, funny, actors, famous, humor
Danecook	When I tweet, I tweet to kill	celebs, comedy, funny, actors, famous
ScreenOrigami	Web developer from Germany	Webdesign, webkraut, html, designer

  - Tweets often contain daily conversation
- Alternative: use crowdsourcing
  - How does the Twitter crowd describe a user?
  - Crowdsourced information collected using Twitter Lists

Twitter Lists

- A feature used to organize the people one is following on Twitter
  - Create a named list, add an optional List description
  - Add related users to the List
  - Tweets posted by these users will be grouped together as a separate stream

List Name	Description	Members
News	News media accounts	nytimes, BBCNews, WSJ, cnnbrk, CBSNews
Music	Musicians	Eminem, britneyspears, ladygaga, rihanna, BonJovi
Politics	Politicians and people who talk about them	BarackObama, nprpolitics, whitehouse, billmaher

Pete Cashmore

@mashable NYC / SF

Breaking social media, tech and digital news and analysis from Mashable.com, the top resource and guide for all things web. Updates from @mashable staff.

http://mashable.com

Tweets | Favorites | Following | Followers | **Lists**

mashable's lists


@mashable/news
A curated list of news organization's Twitter accounts.


@mashable/tech
Experts and sources to keep up with the latest in tech.


@mashable/design
Tweets and tips from designers.


@mashable/food
Love food? Here are chef's, cooks and others in food to follow


@mashable/celebrity
Celebrities on Twitter.


@mashable/journalism
Journalists interested in the future of news media.


@mashable/music
Musicians on Twitter.


nytimes The New York Times
Where the Conversation Begins. Follow breaking news, NYTimes.com home page articles, special features and more.


101Cookbooks 101 Cookbooks
Heidi Swanson from 101Cookbooks.com - Healthy, vegetarian recipes made from natural foods and seasonal produce.

epicurious epicurious
Written by Tanya Steel and the Epicurious editorial staff

LATimesfood LA Times Food
News, recipes + reviews from the LA Times Food staff, test kitchen + Daily Dish blog, by @renelynch.


TylerFlorence Tyler Florence
Chef, Restaurateur, Wine Maker, Cookbook Writer, Shop Keep, Product Designer, Dad.


It's Britney Bitch!


ladygaga Lady Gaga
mother monster

Using Lists to infer topics for users

- If U is an expert / authority in a certain topic
  - U likely to be included in several Lists
  - List names / descriptions provide valuable semantic cues to the topics of expertise of U

Barack Obama

@BarackObama Washington, DC

2,513	683,900	11,864,332	162,383
Tweets	Following	Followers	Listed

Lists following BarackObama

@jyo_0827/news-and-politician
ニュースや政治家

@docsports_shs/national-goverment
national government tweets

@gazhazDCFC/politics

@GOPSackSuckers/truepatriots
Warriors for truth, justice, and equality.

@WendyMagley/interests

@DorianaLeonardo/internazionale

Amitabh Bachchan

@SrBachchan Mumbai, India

11,314	146	1,920,019	19,416
Tweets	Following	Followers	Listed

Lists following SrBachchan

@sukhdam1999/famous

@raxus1/bollywood
Bolly actors, actresses, authors

@yadav657/actors
old is gold

@umairsaeed/celebrities

@Littl_Rock/bollywood-celebrities

@umairsaeed/entertainment

Identify topics from List meta-data

- Consider the Lists in which U is included

- Process List names and descriptions
  - Common language processing techniques such as removal of stopwords, case-folding, ...
  - Identify nouns and adjective (part-of-speech tagging)
  - Get a (term, frequency) vector for user U
  - Consider most frequent unigrams and bigrams as topics

Examples of topics inferred

Twitter Accounts	Top Tags (extracted from List meta-data)
	politics, celebs, government, famous, president, media, leaders, news, current events
	celebs, actors, famous, movies, stars, comedy, funny, music, hollywood, pop culture
	linux, tech, open, software, libre, gnu, computer, developer, ubuntu, unix
	yoga, health, fitness, wellness, magazines, media, mind, meditation, body, inspiration
	politics, senator, congress, government, republicans, iowa, gop, officials, conservative, house
	politics, senate, government, congress, democrats, missouri, dems, officials, progressive, women

Topics inferred from Lists

- Topics inferred are almost always accurate
  - Topics for well-known users (e.g. celebrities, US Senators) verified from Wikipedia pages on these people
  - Conducted a user-survey – more than 80% evaluators found the topics to be accurate and informative

- Depth of information: For US Senators, could identify
  - Political party (democrat / republican), state, gender, ...
  - Political ideologies (e.g. conservative / liberal), ...
  - even Senate committees they are members of

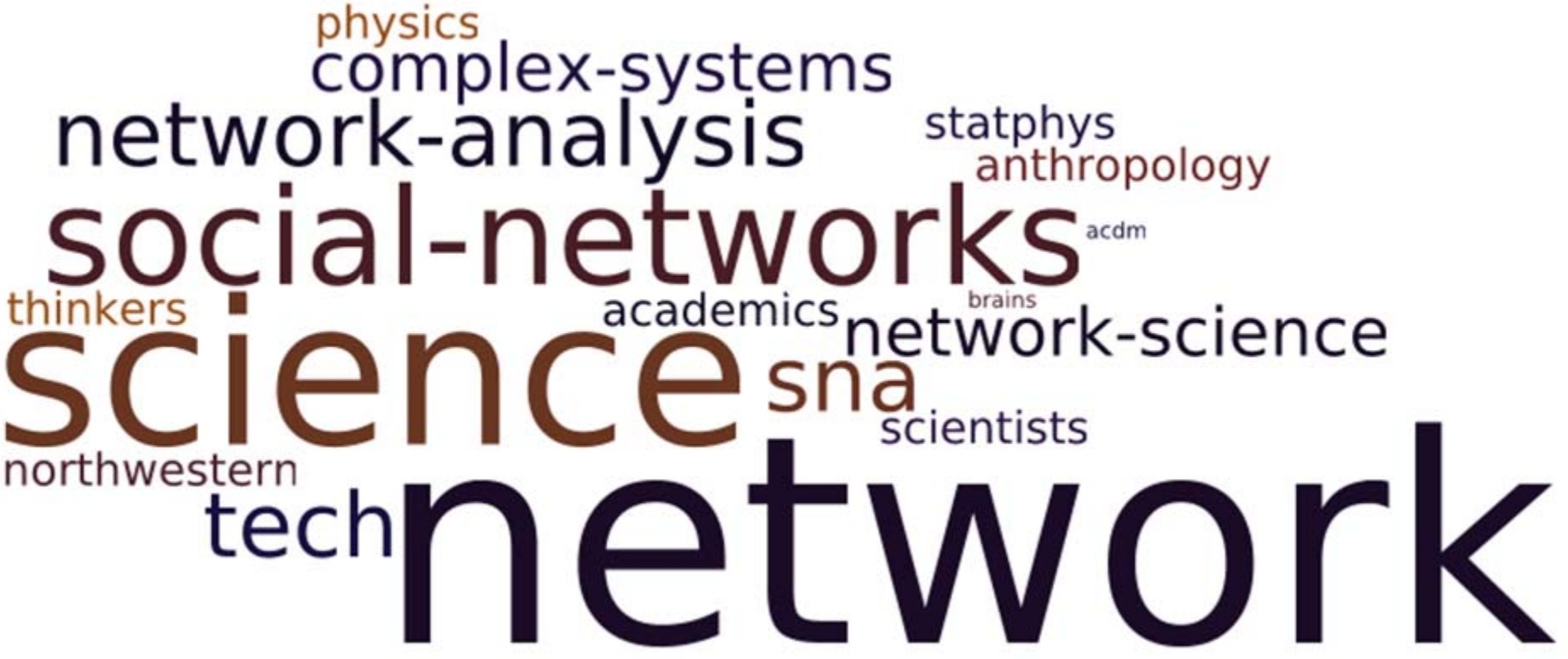
Who-is-who service

- Our who-is-who service for Twitter:

http://twitter-app.mpi-sws.org/who-is-who/

- Given a Twitter user, shows word-cloud for some of the major topics for the user

Topics for Barabasi



Topics for Barack Obama

Twitter as a source of information

- Characterizing the experts in Twitter → characterizing Twitter platform as a whole
- What are the topics on which information can be available in Twitter?
- Do topical experts connect to each other?
- Do topical experts mostly tweet about their own topics of expertise?

Topics in Twitter – major topics to niche ones

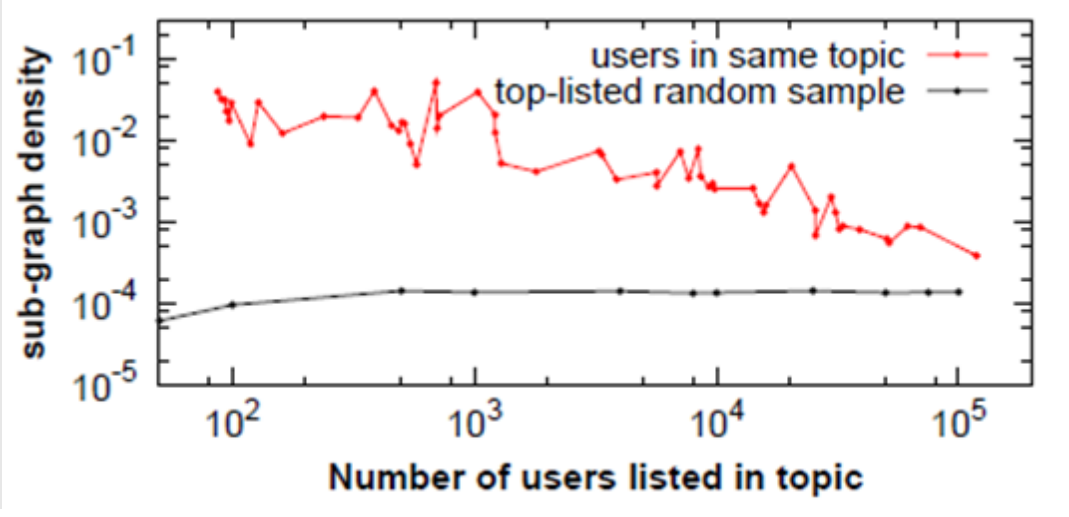
Nos. of listed users in topic	Nos. of topics	General topics	Technology	Sports
> 30K	31	media, music, bloggers, business, artists, politics, writers, celebrities, companies, education, fashion, travel, journalists	tech, developers	sports
10K – 30K	114	health, news, food, books, government, startups, money, club, marketers, *conference*	search-engine, internet, programmers	soccer
5K – 10K	125	economics, environment, religion, librarians, charities, hotels, wine, theatre, comedy, follow-back, *festival*	hackers, podcast, webdesign	athletes
1K – 5K	1,375	history, doctors, police, military, scientists, psychology, philosophy, astronomy, theology, job search	iphone, xbox, HTML, Java, Python, Photoshop	baseball, hockey, tennis, cricket, golf, *olympics*
500 – 1K	1,346	biology, mathematics, geography, astrology, neuroscience, anthropology, classical music, sociology,*earthquake*	Unix, Ubuntu, Perl, javascript	wrestling
100 – 500	8,113	paediatrics, neurology, forestry, forensics, geology, chemistry, homoeopathy, *iranelection, tsunami, hurricane*	django, intranet, broadband, SQL server, malware	volleyball, horse-race, body-building, boxing
< 100	> 78K	astrophysics, audiology, network analysis, malaria, addiction recovery, *oil spill*	cyberwarfare, Solaris	billiards, jiu jitsu, karate, kungfu, *winter olympics*

Major topics in Twitter

Niche topics in Twitter

Topical experts connect to each other

- Density of entire Twitter network: 10^{-7}
- Density of subgraph among experts (those who are Listed at least 10 times): 10^{-4}
- Density of subgraph among experts in same topic even higher
  - Higher for niche topics (with fewer experts) than for major topics
  - Experts in niche topics form densely connected knowledge communities



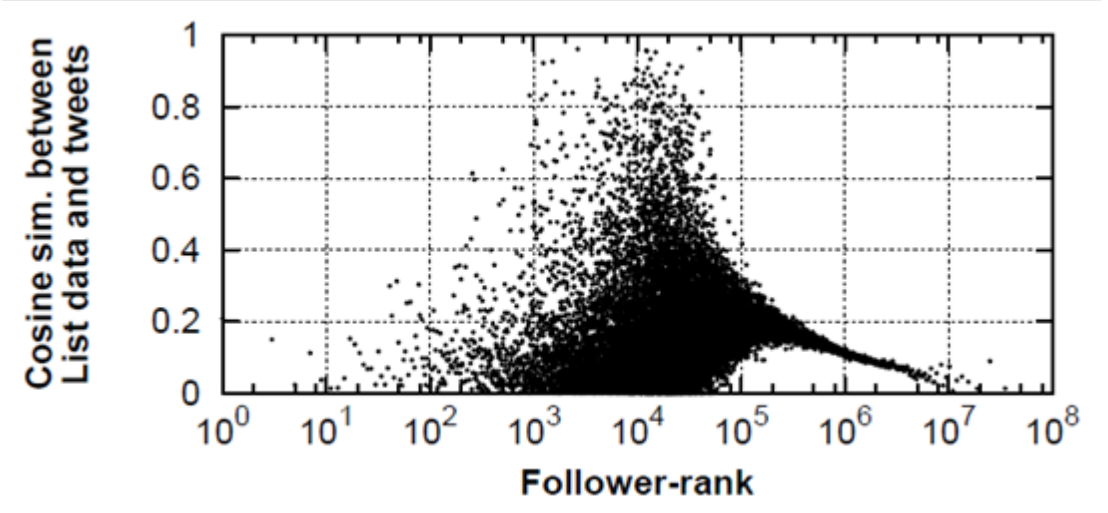
Do experts tweet on their topic of expertise?

- Method
  - (Term, frequency) vector extracted from Lists of U
  - (Term, frequency) vector extracted from tweets posted by U
  - Cosine similarity between the vectors

- Observations
  - Business accounts tweet primarily on their topics of expertise, e.g. Linux Foundation, Yoga journal
  - Most personal accounts tend to tweet on a wide variety of topics, some are more topical

Do experts tweet on their topic of expertise?

- The celebrities (having top follower-ranks) usually tweet on a wide variety of topics
- Some of the users having follower-ranks around 10K mostly tweet on their topics of expertise



Conclusion

- Paper submitted to AAAI ICWSM Conference 2012
- Ongoing work
  - building a topical expert search / who-to-follow service

Thank You

Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Krishna Gummadi

Contact: niloy@cse.iitkgp.ernet.in

Complex Network Research Group (CNeRG)
CSE, IIT Kharagpur, India
http://cse.iitkgp.ac.in/resgrp/cnerg/

Thank You

Contact: niloy@cse.iitkgp.ernet.in

Complex Network Research Group (CNeRG)
CSE, IIT Kharagpur, India
http://cse.iitkgp.ac.in/resgrp/cnerg/

Name	Bio	Major Topics obtained from List
Jimmy fallon	astrophysicts	celebs, comedy, funny, actors, famous, humor
Danecook	When I tweet, I tweet to kill	celebs, comedy, funny, actors, famous
ScreenOrigami	Web developer from Germany	Webdesign, webkraut, html, designer