

Extracting Relevant & Trustworthy Information from Microblogs

Joint work with Bimal Viswanath, Farshad Kooti,
Saptarshi Ghosh, Naveen Sharma, Niloy Ganguly,
Fabricio Benevenuto

MPI-SWS, Germany; IIT Kharagpur, India; UFOP, Brazil

Twitter microblogging site

- ❑ An important source for real-time Web content
 - ❑ 200 million active users as of 2011
 - ❑ 150 million tweets posted daily
- ❑ Quality of tweets / content vary widely
 - ❑ Any one can post tweets
 - ❑ Celebrities, politicians, news media, academics, spammers
- ❑ Challenge: Finding relevant & trustworthy content
 - ❑ Trustworthy: Thwart spammers and their spam
 - ❑ Relevance: Identify authoritative experts on specific topics

Part 1

Thwarting Spammers in Twitter

[WWW 2012]

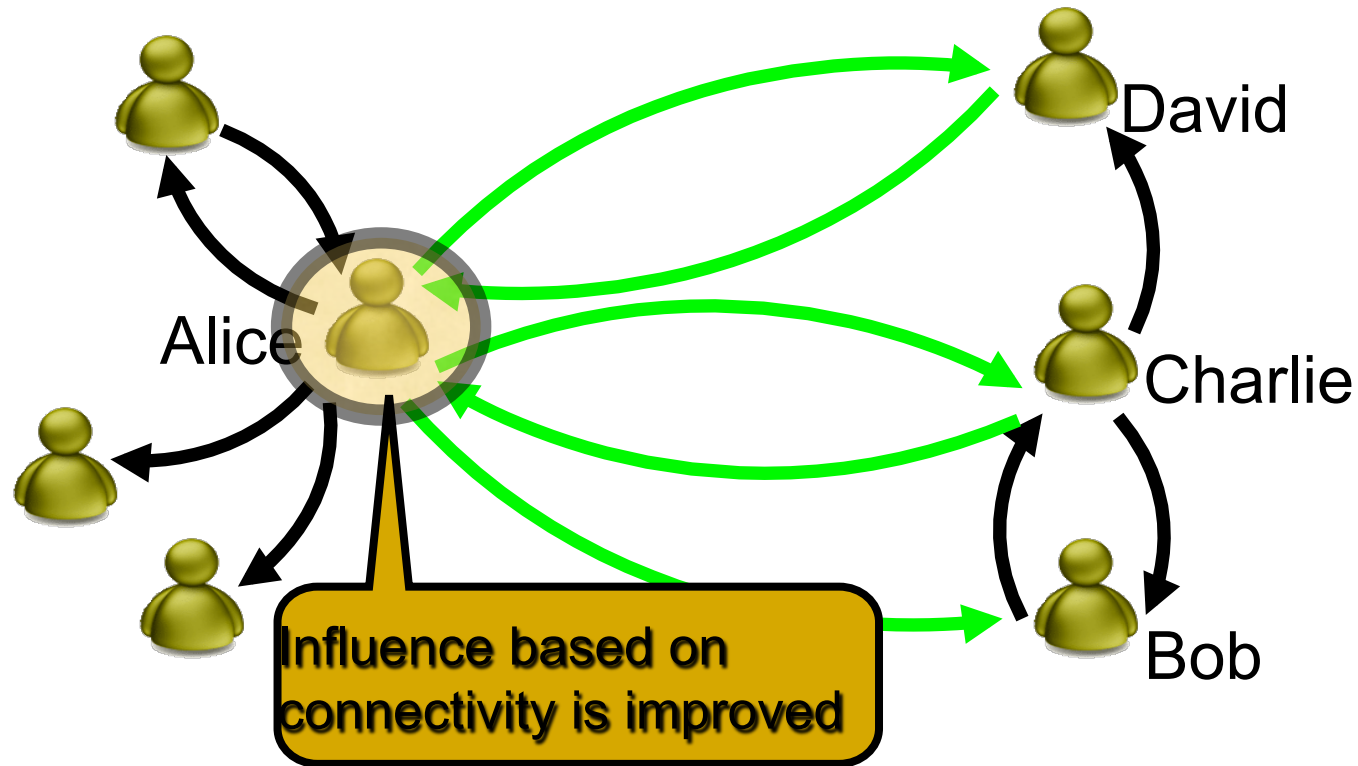
How Twitter spammers operate

- ❑ Spammers try to gain lots of followers
 - ❑ To promote spam directly
 - ❑ To gain influence in the network
- ❑ Search engines rank tweets based on how influential the user is
 - ❑ Most metrics depend on user's network connectivity
 - ❑ More followers help a user to gain influence

Incentivizes spammers to acquire links to gain influence

Acquiring followers via link farming

- Unrelated users exchange links with each other
 - To gain more influence based on network connectivity



To thwart spammers

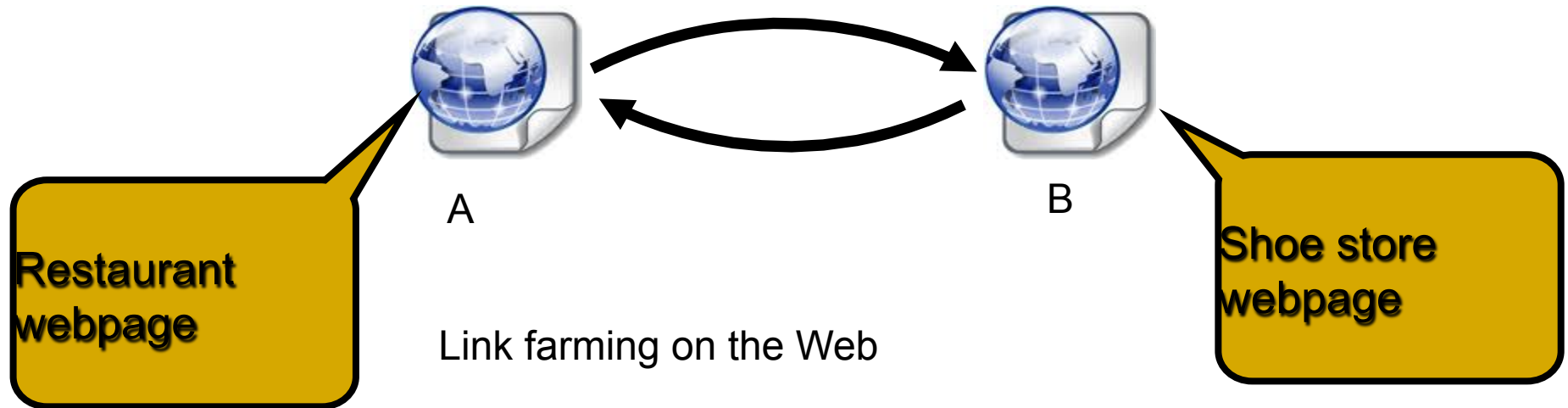
- ❑ We need to
 - ❑ 1. Understand link farming activity in Twitter
 - ❑ 2. Combat link farming activity in Twitter
 - ❑ Prior works: Focused on detecting spammers
 - ❑ Via their characteristics, e.g., follower to following ratios
 - ❑ Rat-race between spammers and spam fighters
 - ❑ We focus on the spammer support network
-

Research Challenge 1

Understanding link farming activity
in Twitter

Challenge: Detecting link farming

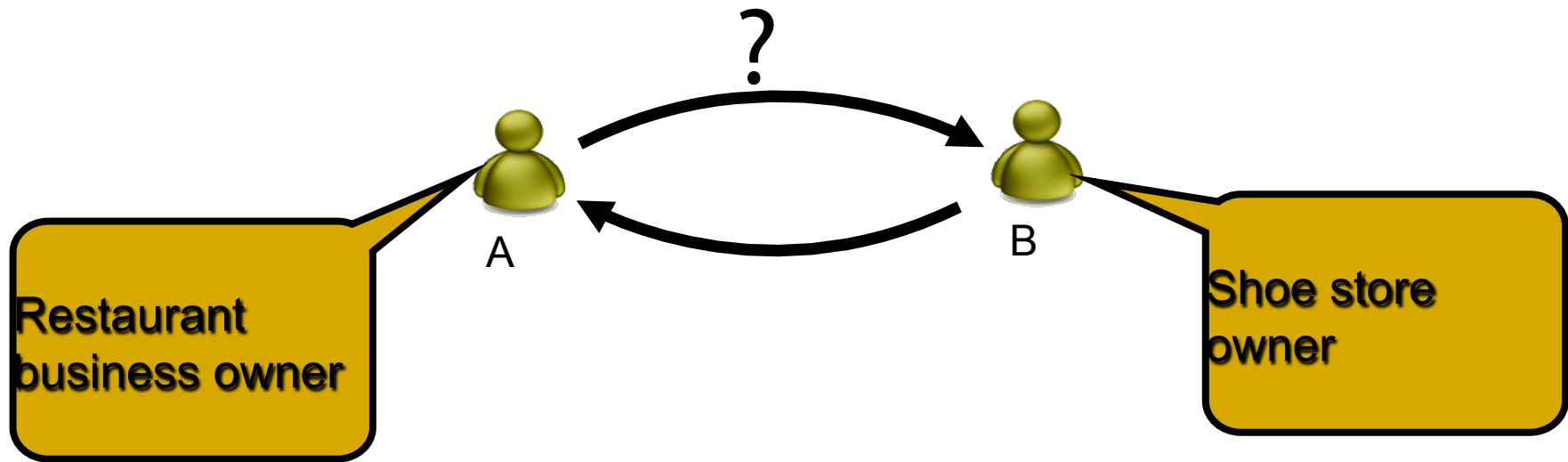
- Link farming in Twitter is more challenging than on the Web



Can detect link farming by analyzing the content of linked pages

Challenge: Detecting link farming

- ❑ Meaning of a social link between two users is unknown



Harder to detect and analyze link farming activity in Twitter

Our idea

- ❑ Analyze the social network of known spammers
 - ❑ Look for evidence of link reciprocation in getting followers

Identifying spammers

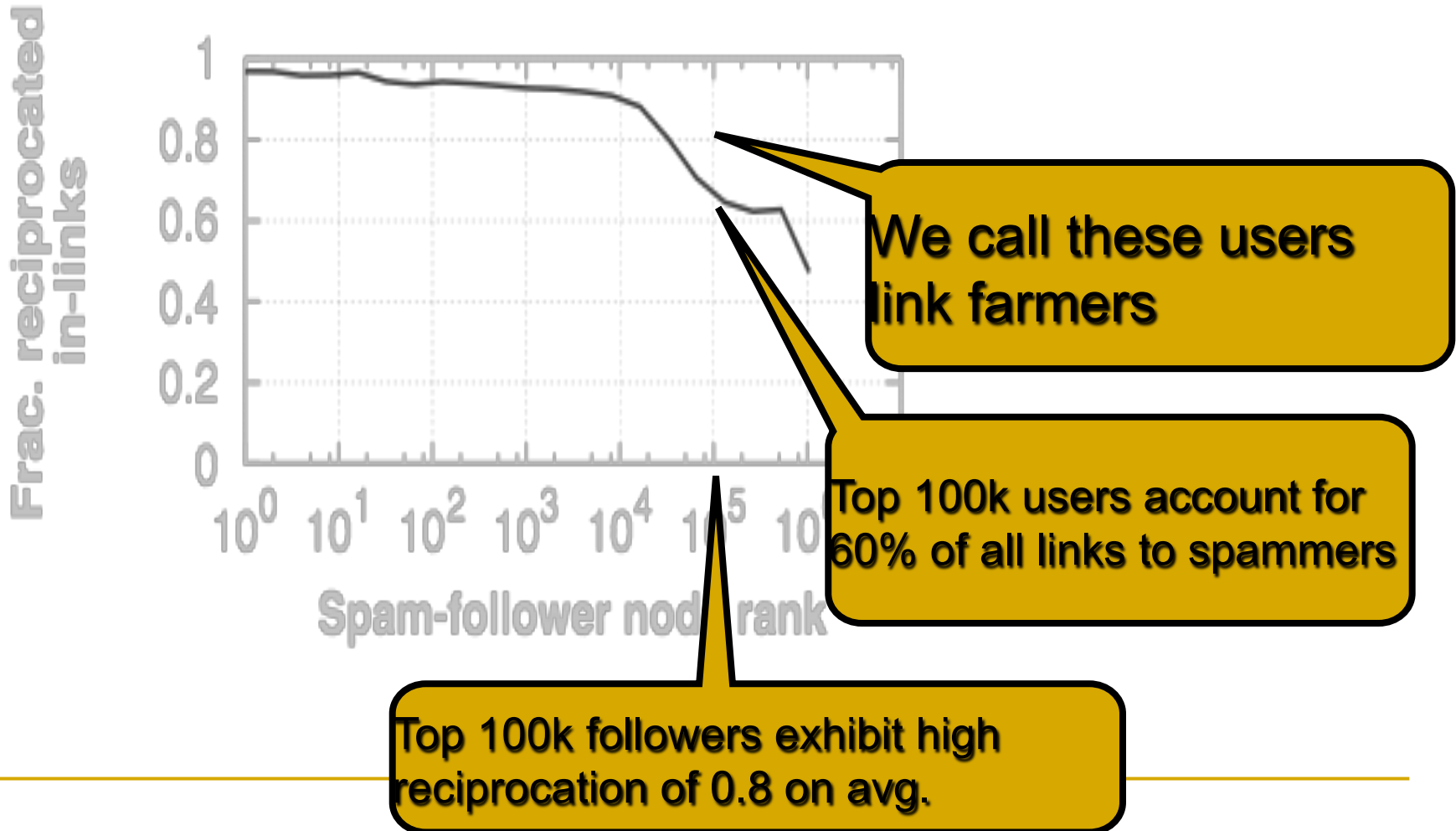
- ❑ Used Twitter network gathered from previous study [ICWSM' 10]
 - ❑ Data collected in August 2009
 - ❑ 54M nodes, 1.9B links, 1.7B Tweets
- ❑ Identified accounts suspended by Twitter
 - ❑ Account could be suspended for various reasons
- ❑ Found suspended users that posted blacklisted URLs
 - ❑ Includes 41,352 such spammers

Spammers farm links at large-scale

- ❑ Spam-targets: 15M users followed by spammers
 - ❑ 27% of all users!
 - ❑ Spam-followers: 1.4M followers
 - ❑ 82% of all followers have been targeted
 - ❑ Spammers have more followers than random users
 - ❑ Avg follower count for Spammers: 234, Random users: 36
-

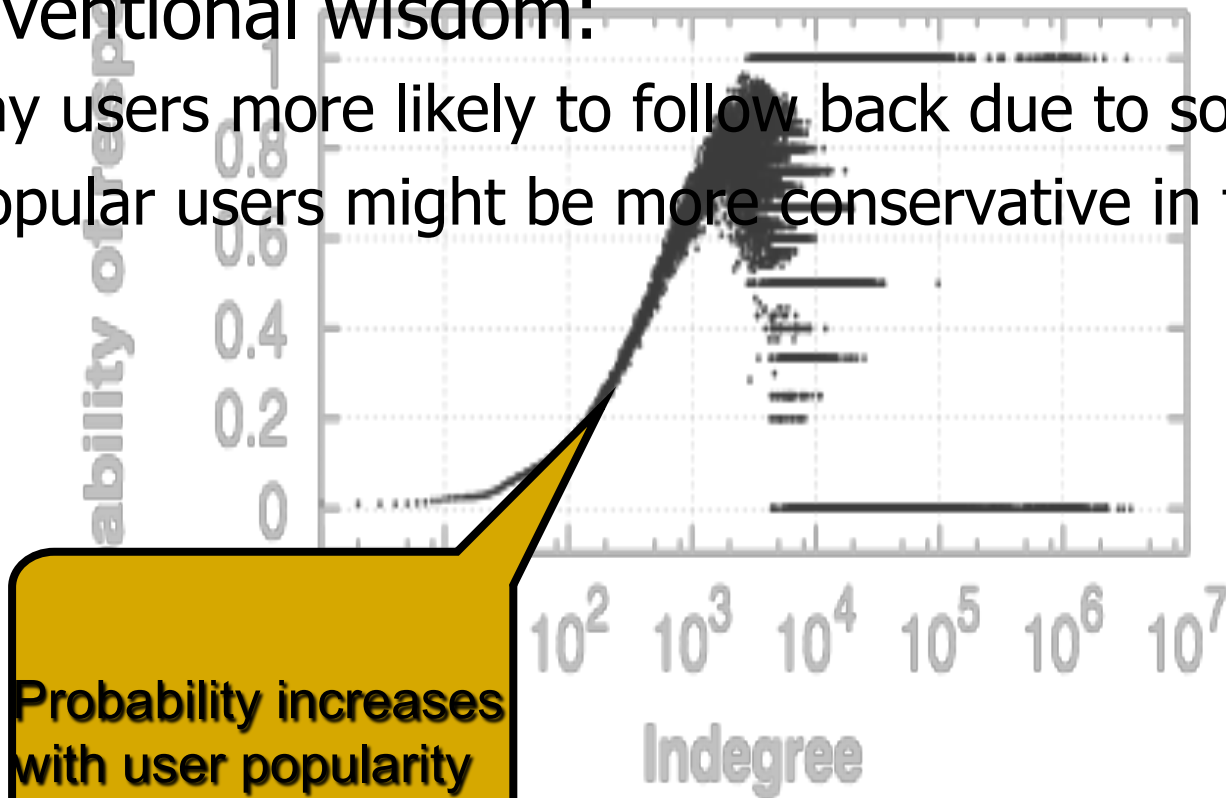
Who responds to links from spammers?

- Small number of followers respond most of the time



Are link farmers lay or popular users?

- Conventional wisdom:
 - Lay users more likely to follow back due to social etiquette
 - Popular users might be more conservative in following others



Link farmers are popular users with lots of followers

Analysis of link farmers

- ❑ Are link farmers real users or spammers?
 - ❑ Who are the link farmers?
 - ❑ What motivates them to engage in link farming?
-

Are link farmers real users or spammers?

To find out if they are spammers or real users, we

- ❑ 1. Checked if they were suspended by Twitter
 - ❑ 76% users not suspended, 235 of them verified by Twitter
 - ❑ 2. Manually verified 100 random users
 - ❑ 86% users are real with legitimate links in their Tweets
 - ❑ 3. Analyzed their profiles
 - ❑ More active in updating their profiles than random users
-

Who are the link farmers?



- ❑ Link farmers are mostly interested in promoting their business or
- ❑ Tweeting about trends in a particular domain

Who are the link farmers?

- ❑ Top 5 link farmers according to Pagerank:
 - ❑ 1. Barack Obama: Obama 2012 campaign staff
 - ❑ 2. Britney Spears
 - ❑ 3. NPR Politics: Political coverage and conversation
 - ❑ 4. UK Prime Minister: PM' s office
 - ❑ 5: JetBlue Airways

Link farmers include legitimate users & organizations

What possibly motivates link farmers?

- ❑ One explanation:
 - ❑ Link farmers have similar incentives as spammers
 - ❑ They seek to amass social capital & influence in the network
 - ❑ Link farmers rank among top 5% influential Twitter users
 - ❑ In terms of various metrics like Pagerank & Followerrank
-

Summary, so far

- ❑ Spammers farm links at large-scale
 - ❑ Some have gained high influence in the network
 - ❑ They are helped by a set of link farmers
 - ❑ Who are legitimate, popular & active users in the network
 - ❑ Have high influence in the network
 - ❑ Link farmers are social capitalists
 - ❑ Seeking to amass social capital & influence in the network
-

Research Challenge 2

**How to combat link farming activity
in Twitter?**

Key challenge and insight

❑ Key challenge:

- ❑ Real, popular and active users are involved in link farming
- ❑ Detecting and suspending spammers alone will not help

❑ Insight:

- ❑ Discourage users from following others carelessly
- ❑ Penalize users following anyone found to be bad
 - ❑ Lower the influence scores of users following spammers

Incentivizes users to be more careful about who they link to

Collusionrank

- ❑ Borrows ideas from spam defense strategies for Web [WWW' 05]
 - ❑ Low Collusionrank score for a user indicates
 - ❑ heavy linking to spammers or spam-followers
 - ❑ Requires a seed set of known spammers
 - ❑ Twitter operator periodically identifies and updates spammers
-

Collusionrank

Algorithm:

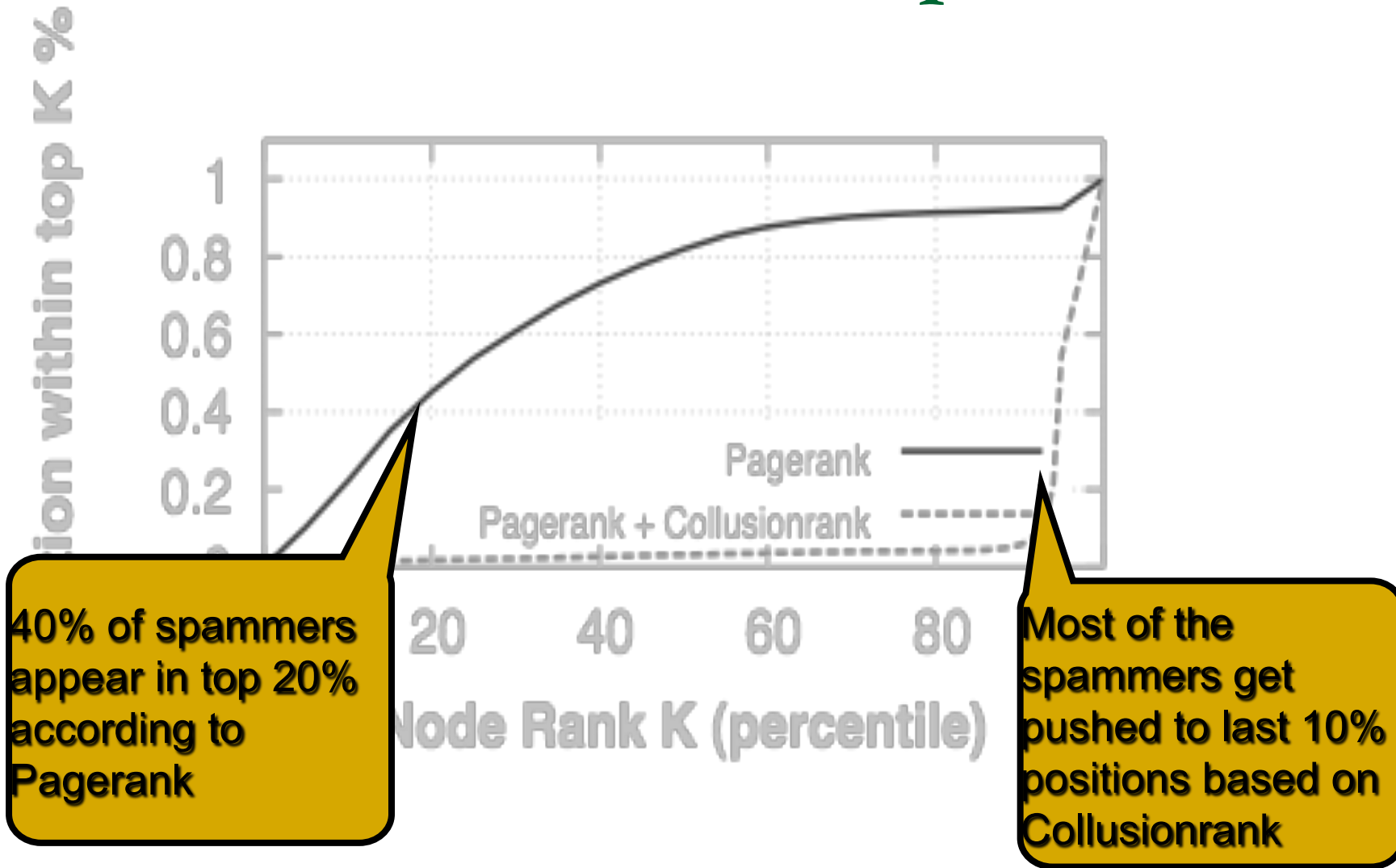
1. Negatively bias the initial scores to the set of spammers
2. In Pagerank style, iteratively penalize users who follow spammers or those who follow spam-followers

Collusionrank is based on the score of followings of a user
Because user is penalized based on who he follows

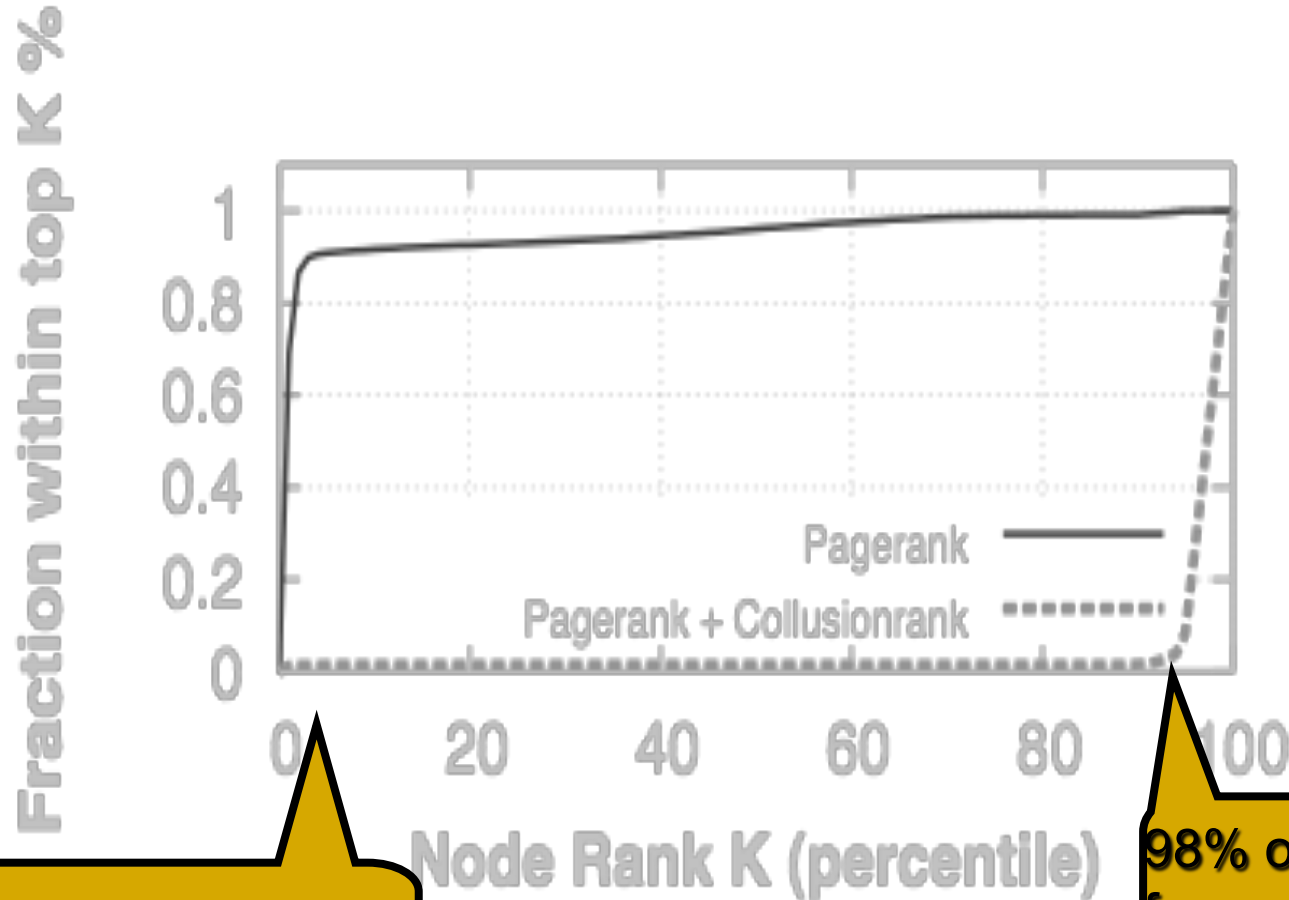
Evaluating Collusionrank

- ❑ Goal:
 - ❑ To penalize spammers and spam-followers
 - ❑ Should not penalize users who are not following spammers
- ❑ Used a small subset of 600 spammers as seed set
- ❑ Compare ranks between
 - ❑ Pagerank
 - ❑ Pagerank + Collusionrank
 - ❑ Measures influence after accounting for link farming activity

Effect of Collusionrank on spammers



Effect on link farmers

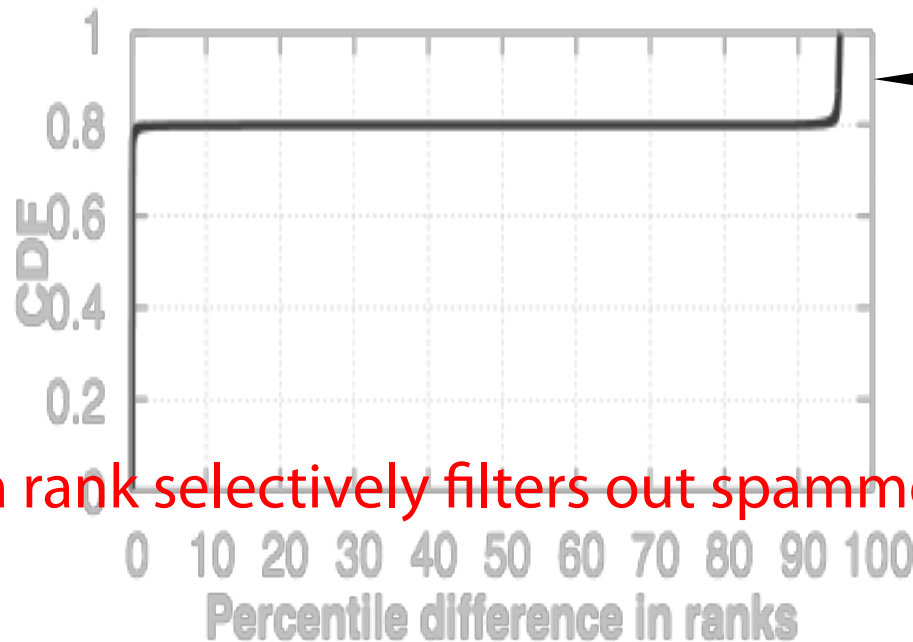


87% of link farmers in top 2% users according to Pagerank

98% of the link farmers get pushed to last 10% positions based on Collusionrank

Effect on normal users

- Focus on top 100,000 users according to Pagerank
 - Analyze the percentile difference in ranks between
 - Pagerank (P) & Pagerank + Collusionrank (PC)
 - Percentile Difference = $(|PC-P|/N) \times 100$



Heavily demoted users follow many more spammers than others

Collusion rank selectively filters out spammers and spam-followers

Conclusion: Thwarting spammers

- ❑ Spammers infiltrate the Twitter network by farming links
 - ❑ Link farming helps them gain influence to promote spam
 - ❑ Search involves ranking users based on connectivity & influence
 - ❑ Analyzed link farming in Twitter by studying spammers
 - ❑ Top link farmers are real, active and popular users
 - ❑ Proposed an algorithm Collusionrank to limit link farming
 - ❑ Incentivizes users to be careful about who they connect with
-

Part 2

Finding Topic Experts in Twitter

[SIGIR 2012]

Prior approaches to find topic experts

- ❑ Research studies

- ❑ Pal et. al. (WSDM 2011)
- ❑ Weng et. al. (WSDM 2010)

- ❑ Application systems

- ❑ Twitter WTF
- ❑ Wefollow

Prior approaches use features extracted from

- ❑ User profiles
 - ❑ Screen-name, bio, ...
- ❑ Tweets posted by a user
 - ❑ Hashtags, others retweeting a given user, ...
- ❑ Social graph of a user
 - ❑ #followers, Pagerank, links with other topic experts, ...

Problems with prior approaches

- ❑ User profiles – screen-name, bio, ...
 - ❑ Information in users profiles mostly unvetted
- ❑ Tweets posted by a user – hashtags, others retweeting a given user, ...
 - ❑ Tweets mostly contain day-to-day conversation
- ❑ Social graph of a user – #followers, Pagerank, links with other topic experts, ...
 - ❑ Followers / influence can easily be acquired by creating Sybil accounts, link farming

Research challenges for search engine for topic experts

- ❑ How to infer topics of expertise of an individual Twitter user?
 - ❑ How to rank the relative expertise of users identified as experts on a topic?
 - ❑ How to keep the system up-to-date as thousands of new users join Twitter daily?
-

Research Challenge 1

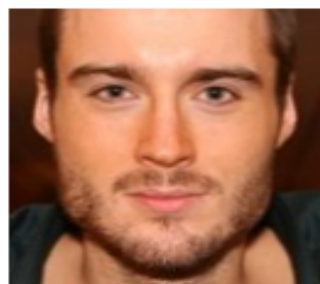
**How to infer topics of expertise of
Twitter users?**

Our proposal

- ❑ Use a different way to infer topics of expertise for an individual Twitter user
 - ❑ Existing approaches primarily rely on information provided by the user herself
 - ❑ We utilize “wisdom of the Twitter crowd”, i.e., how other users describe this user
-

Twitter Lists

- ❑ A feature to organize tweets received from the people whom a user is following
 - ❑ Create a List, add name & description, add other people to the list
 - ❑ Tweets from all listed people will be available as a separate List stream
-



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

Use Lists to associate topics to users

- ❑ If U is an expert on a certain topic
 - ❑ U likely to be included in several Lists by other people
 - ❑ List names / descriptions indicate topics of expertise of U



Barack Obama ✓
@BarackObama Washington, DC

9,245,117
Followers

149,504
Listed



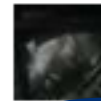
@flyingpackman/politics



@RobertKunz/celebrities



@marcelbedard/personalities



@LisaHathaway/government



@matthew_paul/politicians

Mining Lists to infer expertise

- ❑ Identify frequently occurring terms in List names and descriptions
 - ❑ Handle CamelCase words
 - ❑ Ignore domain-specific stopwords
 - ❑ Identify nouns and adjective
 - ❑ Unify similar words based on edit-distance, e.g., *journalists* and *jornalistas*, *politicians* and *politicos*
 - ❑ Consider unigrams and bigrams as topics
- ❑ Gives for a user – a topic vector, with frequency of occurrence of each topic in List meta-data

Topics extracted from List meta-data



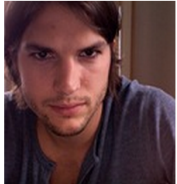
Barack Obama ✓

@BarackObama Washington, DC

This account is run by #Obama2012 campaign staff. Tweets from the President are signed -BO.

<http://www.barackobama.com>

politics, celebs, government, famous, president, media, current events



ashton kutcher ✓

@aplusk Los Angeles, California

I make stuff, actually I make up stuff, stories mostly, collaborations of thoughts, dreams, and actions. Thats me.

<http://www.facebook.com/Ashton>

celebs, actors, famous, movies, stars, comedy, hollywood, pop culture



The Linux Foundation

@linuxfoundation San Francisco, CA

A nonprofit consortium dedicated to fostering the growth of Linux.

<http://www.linux-foundation.org/>

linux, tech, open, software, libre, gnu, computer, developer, ubuntu, unix



ChuckGrassley ✓

@ChuckGrassley Iowa

U.S. Senator born, raised and still living in New Hartford, IA.

<http://facebook.com/grassley>

<http://www.youtube.com/SenChuckGrassley>

<http://grassley.senate.gov>

politics, senator, congress, government, republicans, iowa, gop, conservative



Claire McCaskill ✓

@clairecmc Missouri/ Washington DC

<http://twitter.com/clairecmc>

politics, senate, government, congress, democrats, missouri, progressive, women

Comparison with Twitter WTF

- ❑ Obtained top 20 WTF results for about 200 queries
→ 3495 distinct users
- ❑ Topics inferred from Lists include query-topic for 2916 users (83.4%)
- ❑ For the rest
 - ❑ Case 1 – inferred topics include semantically very similar words, but not exact query-word (18%)
 - ❑ Case 2 – wrong results by WTF, unrelated to query (58%)

Comparison with Twitter WTF

Case 1

- ❑ Restaurant *dineLA* for query “dining”
 - ❑ Inferred topics – food, restaurant, recipes, los angeles
- ❑ Space explorer *HubbleHugger77* for query “hubble”
 - ❑ Inferred topics – science, tech, space, cosmology, nasa
- ❑ Comedian jimmyfallon for query “astrophysicist”
 - ❑ Inferred topics – celebs, comedy, humor, actor
- ❑ Web developer ScreenOrigami for query “origami”
 - ❑ Inferred topics – webdesign, html, designers

Case 2

Research Challenge 2

How to rank experts on a topic?

Ranking experts

- ❑ Used a ranking scheme solely based on Lists
- ❑ For given query, identify experts on the query-topic
- ❑ Compute topical similarity score sim_t for each user
 - ❑ Cover density ranking between topic vector for user and query vector
 - ❑ Queries are short – mostly unigrams, few bigrams
- ❑ Multiply sim_t by logarithm of number of Lists containing the user

Cognos

- ❑ Search engine for topic experts in Twitter
 - ❑ Initially populated with 1.3 million users
 - ❑ From among the 54 million users in Twitter as of 2009
 - ❑ Included in at least 10 Lists
 - ❑ Collected 88 million Lists in total
-

Location Filter :

Search results for "web search"

**sengineland** : **Search Engine Land** ✓*Follow us for news about Google, Bing, Yahoo, search marketing (SEM), search engine optimization (SEO), paid search (PPC) & how to use search engines better!***SEOMoz** : **SEOMoz** ✓*The Web's Best SEO Software, Tools, Resources and Community.***webseoanalytics** : **Web SEO Analytics***Web SEO Analytics - Professional SEO Tools, Search Engine Optimization News, Search Engine Marketing & Online Marketing Blog***sejournal** : **SEJournal***Search Engine Journal : Internet Marketing News, Tutorials and Features***mattcutts** : **Matt Cutts***I'm the head of the webspam team at Google.***rustybrick** : **Barry Schwartz***Search Geek*

Cognos
results for
query “web
search”

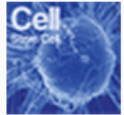
Location Filter :

Search results for "stem cell"



StemCellTracker : Stem Cell Tracker

The #1 resource for up-to-date stem cell news.



CellStemCell : Cell Stem Cell

A monthly journal at the forefront of stem cell science and policy. Editor: Dr. Deborah Sweet



StemCellNetwork : Stem Cell Network

SCN supports cutting-edge projects that translate Canadian stem cell research discoveries into new and better treatments.



iPSCellNews : Stem Cell (iPS) News

Latest updates about reprogramming & induced pluripotent stem cell research. -Jordan Kho, PhD student at Baylor College of Medicine-



ATStemCell : All Things Stem Cell

Blog discusses stem cells in multifaceted manner: history, apps, probs, news, & more. Run by PhD stem cell grad student/science writer- Need something written?

Cognos
results for
query
“stem cell”

User evaluation of Cognos

- ❑ Publicly deployed at
<http://twitter-app.mpi-sws.org/whom-to-follow/>
 - ❑ Evaluators: people at the three home institutions of authors
 - ❑ An evaluator shown top 10 results, gives binary relevant / non-relevant judgment on each result
-



User-screen for evaluation

Please rate the top 10 results for "science news"



bbcscitech : BBC SciTech



Set up by @mario, supported by backstage.bbc.co.uk

Relevant  



Reuters_Science : Reuters Science News


From newly charted astronomical anomalies at the far reaches of the universe to the rise of nanotechnology, nobody covers science like Reuters.com.

Relevant  



newscientist : New Scientist



New Scientist is the world's leading science and technology weekly

Relevant  



NatGeo : National Geographic



Since 1888, we've traveled the Earth, sharing its amazing stories with new generations. Official Twitter account of National Geographic.

Relevant  



science : science

Science news from ScienceNewsBlog.com.

Relevant  



NASA : NASA

Relevant  

55 sample queries for evaluation

Whom to Follow? — *Discover topic authorities on Twitter!*

Location Filter :

Sample Queries

News: politics sports entertainment science technology business

Journalists: politics sports entertainment science technology business

Politics: conservative news liberal politicians USA / German / Brazilian / Indian politicians

Sports: F1 baseball soccer poker tennis NFL NBA Bundesliga LALakers

Entertainment: celebrities movie reviews theater music

Hobbies: hiking cooking chefs traveling photography

Lifestyle: wine dining book clubs health fashion

Science: biology astronomy computer science complex networks

Technology: iPhone mac linux cloud computing

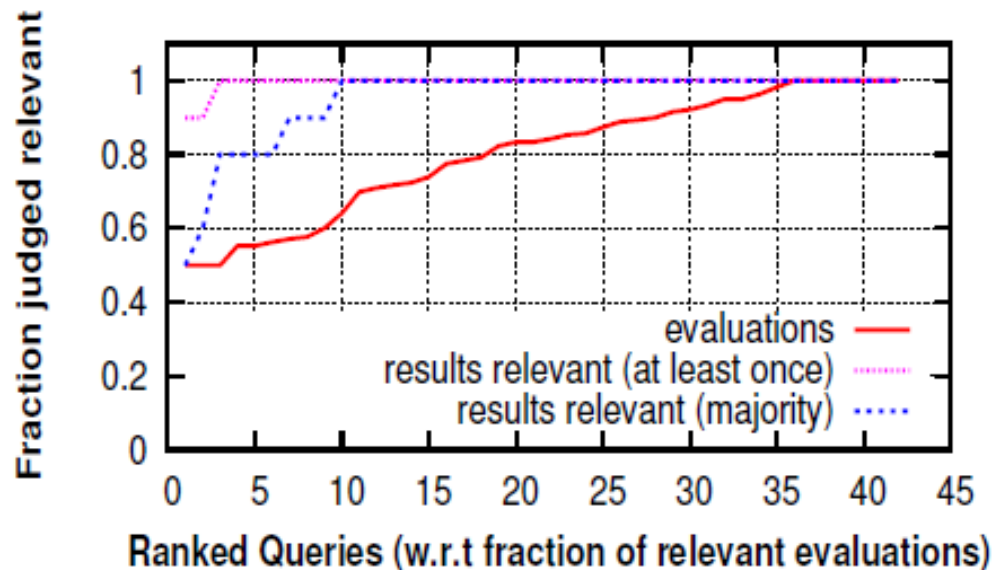
Business: markets finance energy

Evaluation results

- ❑ Overall 2136 relevance judgments over 55 queries
 - ❑ 1680 said relevant (78.7%)
 - ❑ Top 4 results judged relevant in 80% evaluations
 - ❑ Results ranked 5-10 judged relevant in 75% evaluations
 - ❑ Large amount of subjectivity in evaluations
 - ❑ Same result for same query received both relevant and non-relevant judgments
 - ❑ E.g., Werner Vogels for query “cloud computing” got 4 relevant judgments, 6 non-relevant judgments
-

Evaluation results (contd.)

- ❑ Consider only the results evaluated at least twice
- ❑ For each query, measure
 - ❑ What fraction of evaluations judged a result as relevant?
 - ❑ What fraction of top 10 results judged relevant in (i) at least one evaluation? (ii) the majority of evaluations?



Comparison with Twitter WTF

- ❑ Evaluator shown top 10 results by Cognos and Twitter WTF
 - ❑ Result-sets anonymized
 - ❑ Evaluator judges which is better / both good / both bad
 - ❑ Queries chosen by evaluators themselves
 - ❑ 27 distinct queries were asked at least twice
 - ❑ In total, asked 93 times
 - ❑ Judgment by majority voting
-

Please compare the anonymized search results for "iphone"

☐ A is better ☐ B is better ☐ Both are equally good ☐ Both are equally bad

Search Engine A



iPhone_News : iPhone News

iPhone news and notes from around the web.



AppStore : App Store

Follow us for official App Store tweets including our featured apps, exclusive offers, and more.



iphone_dev : iphone_dev

The iPhone Dev Team. We liberate your iPhones!



freeiphoneapps : Free iPhone Apps

We're currently giving away the new iPad! We also give away lots of cool paid iPhone, iPad, and Mac applications on a daily basis. FOLLOWUS.



iphonehackx : iPhone Hacks

Unlock the potential of your iPhone, iPad and iPod Touch

Search Engine B



p0sixninja : Joshua Hill

iPhone Hacker



iH8sn0w : iH8sn0w

I'm the guy that made f0recast, iREB, iFaith, and sn0wbreeze. I also make #pie and dj on the side.



chronicdevteam : Chronic Dev

Hax



MuscleNerd : MuscleNerd

iPhone hacker



iPhone_News : iPhone News

iPhone news and notes from around the web.

Cognos vs Twitter WTF

- ❑ Cognos judged better on 12 queries
 - ❑ Computer science, Linux, mac, Apple, ipad, India, internet, windows phone, photography, political journalist
- ❑ Twitter WTF judged better on 11 queries
 - ❑ Music, Sachin Tendulkar, Anjelina Jolie, Harry Potter, metallica, cloud computing, IIT Kharagpur
 - ❑ Mostly names of individuals or organizations
- ❑ Tie on 4 queries
 - ❑ Microsoft, Dell, Kolkata, Sanskrit as an official language

Cognos vs Twitter WTF

- ❑ Low overlap between top 10 results
 - ❑ ... In spite of same topic being inferred for 83% experts
 - ❑ Major differences are due to List-based ranking
 - ❑ Top Twitter WTF results – mostly business accounts
 - ❑ Top Cognos results – mostly personal accounts
-

music

Search

Please compare the anonymized search results for "music"

☐ A is better ☐ B is better ☐ Both are equally good ☐ Both are equally bad

Submit

Search Engine A



itunesmusic : iTunes Music

Official music updates for the U.S. iTunes Store including new releases, pre-orders, iTunes LP, exclusive offers and more.



SonyMusicGlobal : Sony Music Global

The home of Sony Music on Twitter!



guardianmusic : Guardian music

Squashing music into 140 characters since 2008



yahoo_music : Yahoo! Music

The official Twitter account of Yahoo! Music. We tweet about music news, concerts, performances, videos, and all the things that make us yodel!



justinbieber : Justin Bieber

#BELIEVE out this coming Tues JUNE 19th! - I GOT SO MUCH LOVE FOR THE FANS

Search Engine B



katyperry : Katy Perry

i kissed a girl AND diddled her skittle.



ladygaga : Lady Gaga

mother monster



taylorswift13 : taylorswift13



jtimberlake : Justin Timberlake

Official Justin Timberlake Twitter.



Pink : P!nk

it's all happening

Research challenges for search engine for topic experts

- ❑ How to infer topics of expertise of an individual Twitter user?
 - ❑ How to rank the relative expertise of users identified as experts on a topic?
 - ❑ How to keep the system up-to-date as thousands of new users join Twitter daily?
-

Conclusion: Finding topic experts

- ❑ Developed and deployed Cognos
 - ❑ Uses only crowdsourced Lists to infer topics of expertise and rank experts
 - ❑ Competes favorably with official Twitter WTF and state-of-the-art research system
 - ❑ Future work – make the inference methodology robust against List spam
-

The big picture

- ❑ Microblogs are an important source for real-time Web content
 - ❑ But, quality of tweets / content vary widely
 - ❑ Finding relevant & trustworthy content in Twitter
 - ❑ Part 1: Thwarting spammers and their spam
 - ❑ Part 2: Identifying authoritative experts on specific topics
-

Thank You

You can try Cognos at:

<http://twitter-app.mpi-sws.org/whom-to-follow/>