Network Centrality

Saptarshi Ghosh
Department of CSE, IIT Kharagpur
Social Computing course, CS60017

Node centrality

- Relative importance of a node in a network

- How influential a person is within a social network
- How important a webpage is in the Web

- There is an analogous concept of edge centrality, but we will focus on node centrality

Node centrality measures

- Many proposed centrality measures
  - Network structure based
  - Activity based (e.g., number of times a user is retweeted or mentioned on Twitter)
  - Temporal (e.g., Test-of-Time awards to publications)
  - Hybrid
  - … and more
- We will focus on the first two types of measures

Degree centrality

- Simply, centrality measured by degree of a node
  - A node of higher degree is more important

- Undirected graphs
  - Number of friends of a user in Facebook
  - Important stations in railway networks

- Directed graphs: usually indegree of node
  - Number of pages linking to a given page in the Web
  - Number of followers of a user in Twitter

Closeness centrality

- Farness of node *s* : sum of its shortest distances to all other nodes
- Closeness of node *s* : inverse of farness

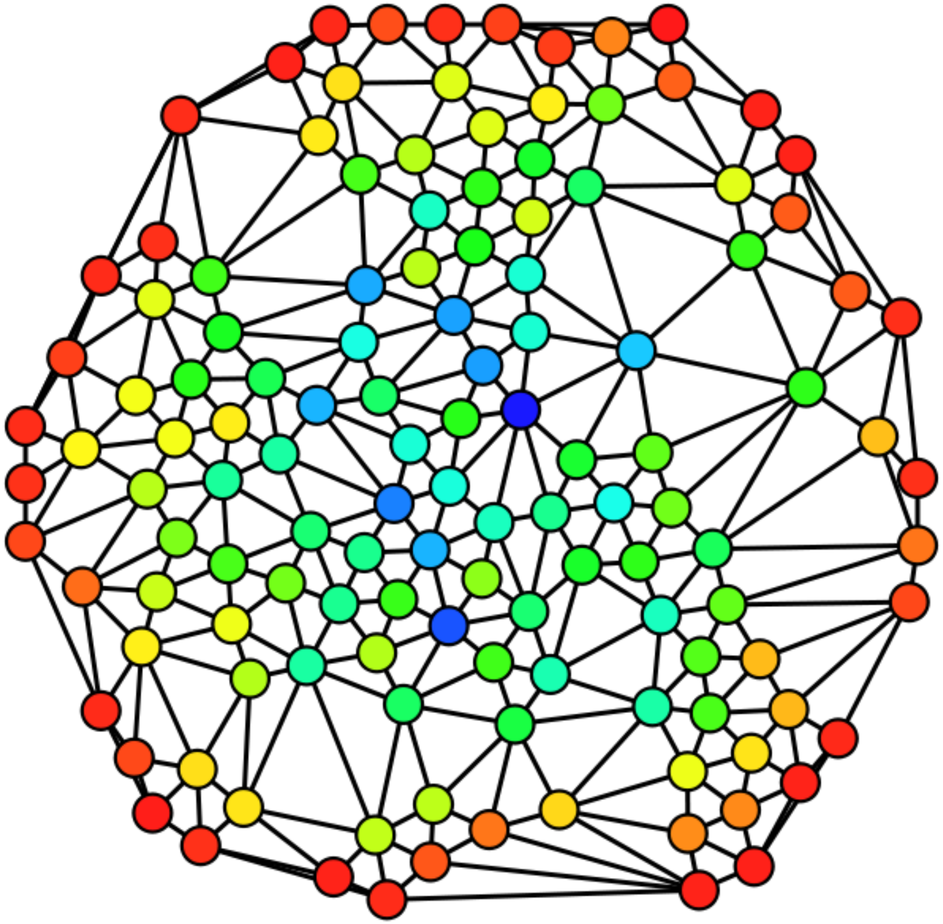
- Higher the closeness centrality of *s*, the lower is its total distance to all other nodes

- Applications
  - Where to set up a hospital in a town?
  - How fast can information spread from *s* to all other nodes?

Betweenness centrality

- Betweenness of node s:
  - For each pair of vertices (u, v), find the shortest paths between them
  - Compute the fraction of these shortest paths which pass through node s
  - Sum this fraction for all pairs of nodes (u, v)

Example of betweenness centrality



Betweenness centrality coded by color

Red: 0 betweenness
Blue: maximum betweenness

CENTRALITY IN DIRECTED GRAPHS (WEB GRAPH)

Node centrality in Web

- Web graph: nodes are webpages, edges are hyperlinks (directed)

- Results of Web search: list of webpages / websites ranked according to
  - Relevance to query
  - <u>Importance / trustworthiness - centrality</u>
  - Location / time of query
  - Recency of page
  - … and many others

Importance of node centrality in Web

- If only relevance used to rank webpages, ranking algorithm can be easily spammed
- Previously, indegree of webpages used to rank pages according to importance
- Easily gamed by spammers creating their own webpages

HITS ALGORITHM

HITS algorithm

- Hyperlink-Induced Topic Search, by Kleinberg

- Two types of important pages on the Web
  - Authority: has authoritative content on a topic
  - Hub: pages which link to many authoritative pages, e.g., a directory or catalog
  - A good hub is one which links to many good authorities
  - A good authority is one which is linked to by many good hubs

HITS

- HITS computes two scores for each page *p*
  - Authority score: sum of hub scores of all pages which point to *p*
  - Hub score: sum of authority scores of all pages which *p* points to

- Iterative algorithm
  - A series of iterations run, until the scores of all pages converge

HITS run on a query-dependent sub-graph

- Meant to run on a (sub)set of pages that are relevant to a given query
  - Top N pages relevant to query retrieved based on content → called the root set
  - Add to the root set all pages that are linked from it or that links to it → base set
  - Sub-graph of all nodes in base set → focused sub-graph

- Motivation of building base set
  - A good authority page may not contain the query term
  - Hubs describe authorities through the anchor text / text surrounding hyperlinks

HITS Algorithm

```
Find focused sub-graph G of pages relevant to given query
for each page p in G:
    p.auth ← 1,  p.hub ← 1
do until convergence
    for each page p in G
        p.auth ←  Σ q.hub  for all pages q which link to p
        p.hub ←  Σ r.auth  for all pages r which p links to

    Normalize hub and auth scores for all pages
    Check convergence of scores
```

Normalization of scores

- Scores need to be normalized after each iteration

- Different normalization schemes proposed
  - Normalize so that score vectors sum to 1

  - Normalization factor F: square root of sum of squares of current scores of all pages; divide score of each page by F at the end of each iteration

Checking for convergence

- Various convergence criteria used
  - Fixed number of iterations
  - Iterate until scores do not change appreciably from one iteration to the next (compute difference of score vectors from previous and current iterations)
  - Iterate until rankings of pages do not change

Matrix version of HITS

- Matrices / vectors
  - A: adjacency matrix of web graph. (u, v)-th element is 1 if page u links to page v
  - h: vector of hub scores of all pages
  - a: vector of authority scores of all pages

- h ← A.a
- a ← A^T .h

HITS not used commonly

- Hubs often transit to authorities
- Search engines themselves become hubs

PAGERANK ALGORITHM

PageRank

- By Larry Page and Sergey Brin

- Problem in measuring importance by indegree
  - Not all in-links are same
  - How important are those pages which link to page *p*?

- PageRank of a page
  - Just one of many factors used by Google to rank pages
  - Independent of query

Idea of PageRank

- PR of page *p* is a function of the PR of pages which link to *p*

- If page *q* links to 4 pages, *q* contributes *PR(q)/4* to the PR of each of those 4 pages

- Iterative algorithm, multiple iterations needed (until convergence)

PageRank computation

```
/* initialization */
for all nodes u in G: $d(u) \leftarrow 1/N$, where $N$ = #nodes
for all nodes u in G: $PR(u) \leftarrow d(u)$
/* iteration */
do until $PR$ vector converges
   for all nodes $u$ in G
         for all nodes $v$ that links to $u$
                  $t = \Sigma PR(v)$ / out-degree($v$)
         $PR(u) \leftarrow \alpha * t + (1 - \alpha) * d(u)$
   normalize scores
   check for convergence
end
```

Theoretical basis of PageRank

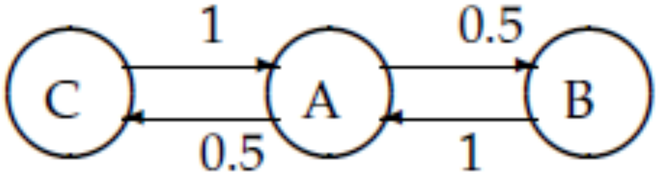
- Random surfer model
  - Start at a node, execute a random walk on Web graph
  - At each step, proceed from current node *u* to a randomly chosen node that *u* links to
  - Teleport: jump to any random node with probability 1/N
  - At a node with no outgoing links, teleport
  - At a node that has outgoing links
    - Follow standard random walk with probability α where $0<\alpha<1$
    - Teleport with probability $(1-\alpha)$
- Nodes visited more frequently in this random walk are web-pages with higher PR

Theoretical basis

- The random walk defines a Markov chain
  - A discrete time stochastic process following Markov property (next state depends only on current state)
  - *N* states corresponding to the *N* nodes; chain is at one of the states at any given time-step
  - *N* x *N* transition probability matrix *P* : P_{ij} is the probability that state at next time-step is *j*, given current state is *i*

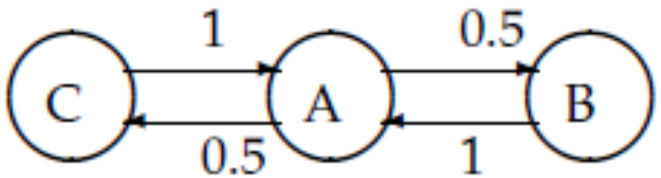
$$\forall i, j, P_{ij} \in [0,1] \qquad \forall i, \sum_{j=1}^{N} P_{ij} = 1.$$

Toy example



$$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Toy example



$$\begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

- *P* is a stochastic matrix
  - Every element is in [0, 1]
  - Sum of every row is 1
  - Largest eigenvalue is 1
  - Has a principal left eigenvector corresponding to its largest eigenvalue

Transition matrix for random surfer

- How to derive the transition matrix for the random surfer on the Web graph?

- Adjacency matrix of Web graph
  - A_{ij} = 1 if there is a hyperlink from page *i* to page *j*
  - A_{ij} = 0 otherwise

- Derive transition matrix *P* of Markov chain from *A*

Transition matrix for random surfer

- Derive transition matrix *P* of Markov chain from *A*
  - If a row of *A* has no 1's, replace each element by *1/N*
  - For all other rows: divide each 1 by the number of 1's in the row
  - Multiply the resulting matrix by α
  - Add (1-α)/*N* to every entry of the resulting matrix

Given P, how to compute PageRank?

- Vector x: probability distribution of surfer's position at any time
  - At $t = 0$: one entry in x is 1, rest are 0
  - At $t = 1$: xP
  - At $t = 2$: $(xP)P = xP^2$
  - …
- Steady-state $x = \Pi$ gives the PageRank scores

Given P, how to compute PageRank?

- Vector x: probability distribution of surfer's position at any time
  - At $t = 0$: one entry in x is 1, rest are 0
  - At $t = 1$: xP
  - At $t = 2$: $(xP)P = xP^2$
  - …

- Steady-state $x = \Pi$ gives the PageRank scores
- PageRank scores obtained as the principal left eigenvector of P (corresponding to eigenvalue 1)

Why teleportation?

- Convergence of PageRank is guaranteed only if
  - The transition probability matrix P is irreducible, i.e., all transitions have a non-zero probability
  - In other words, if the graph (on which random surfing is taking place) is strongly connected

- To ensure convergence
  - To nodes with out-degree 0, add an outgoing edge to every node
  - Damp the walk by factor α, by adding a complete set of outgoing edges, with weight $(1-\alpha)/N$, to all nodes

PageRank computation

- Need to compute principal left eigenvector of a stochastic matrix
- Several numerical methods, e.g., power iteration
- Difficult to compute for matrices of the size of the Web graph

Practical challenges

- All links *u*→*v* do not signify a vote for *v*
  - E.g., links to a copyright page from all pages in a website

- Attempts to spam PageRank: link spam farms or link farms
  - A target page (whose PR the spammer wants to boost)
  - A number of boosting pages, which link to the target page, link to each other and also to external pages
  - Hijacked links – links accumulated from pages outside the link farm

Example link farm

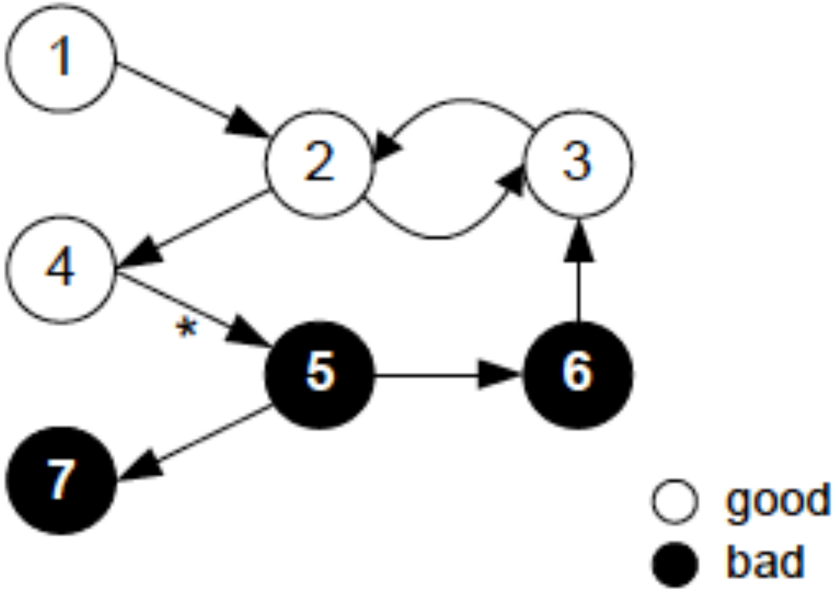


Figure 2: A web of good (white) and bad (black) nodes.

VARIATIONS OF PAGERANK

PageRank computation

```
/* initialization */
for all nodes u in G: $d(u) \leftarrow 1/N$, where $N$ = #nodes
for all nodes u in G: $PR(u) \leftarrow d(u)$
/* iteration */
do until $PR$ vector converges
    for all nodes $u$ in G
        for all nodes $v$ that links to $u$
            $t = \Sigma\, PR(v)$ / out-degree($v$)
        $PR(u) \leftarrow \alpha * t + (1 - \alpha) * d(u)$
    normalize scores
    check for convergence
end
```

Biased PageRank

- Instead of using the uniform vector *d(u)* ← *1/N* for all nodes u, use a non-uniform preference vector:

  d(u) = 1 / |S|, for all u ε S

  = 0 otherwise

- Implication for random surfer:
  - With probability α, follow standard random walk
  - With probability (1-α), teleport to a node in S, where the particular node in S is chosen randomly

Biased PageRank

- Instead of using the uniform vector *d(u)* ← *1/N* for all nodes u, use a non-uniform preference vector:

  d(u) = 1 / |S|, for all u ε S

  = 0 otherwise

- The preference vector biases the ranks towards nodes that are closer to nodes with a larger value in the preference vector

Topic-sensitive PageRank [Haveliwala, WWW 2002]

- Webpages are classified into various topics (16 Open Directory Project high-level categories)
- Computes PageRank for a particular topic of interest

- For category c_j
  - T_j is the set of websites for category c_j
  - Modified teleportation function

$$v_{ji} = \begin{cases} \frac{1}{|T_j|} & i \in T_j, \\ 0 & i \notin T_j. \end{cases}$$

TrustRank [Gyongyi, VLDB 2004]

- Aims to rank trusted pages higher, and push untrusted pages down in the rankings
- Assumes
  - A way of knowing trusted nodes: oracle
  - Trusted (good) nodes will only link to other good nodes but this assumption is violated in the real Web
  - Bad nodes will link to other bad nodes and good nodes

- Run PageRank by biasing the preference vector towards a set of trusted nodes

TrustRank vs. PageRank

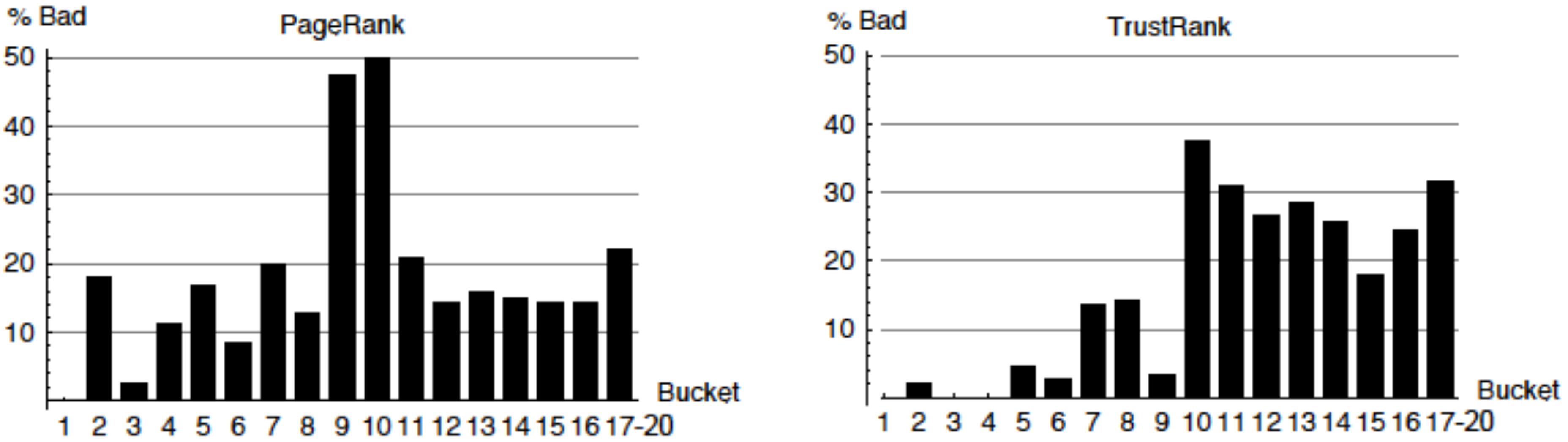


Figure 10: Bad sites in PageRank and TrustRank buckets.

Case Study 1

Measuring User Influence in Twitter: The Million Follower Fallacy, Cha et al., ICWSM 2010

Different influence measures for OSN

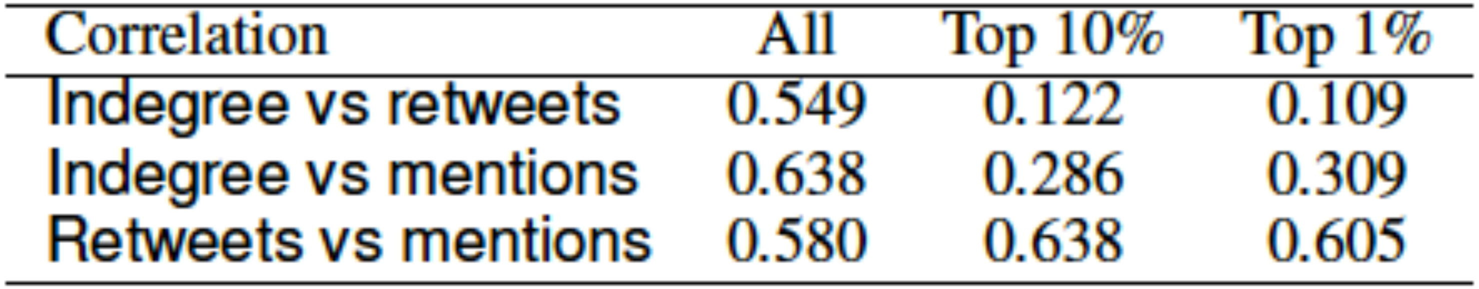
- Compared different influence measures for the Twitter social network
- Network structure based:
  - In-degree (number of followers)
- Activity based:
  - Number of times a user is retweeted
  - Number of times a user is mentioned

- Two measures compared using Spearman’s rank correlation coefficient

Results of comparison

- Across all three measures, top influentials were public figures (politicians, celebrities, …) and websites (news media sites)
- But top influentials according to indegree have low overlap with top influentials according to activity

Table 1: Spearman's rank correlation coefficients



Correlation	All	Top 10%	Top 1%
Indegree vs retweets	0.549	0.122	0.109
Indegree vs mentions	0.638	0.286	0.309
Retweets vs mentions	0.580	0.638	0.605

Case Study 2

Understanding and Combating Link Farming in the Twitter Social Network, Ghosh et al., WWW 2012

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

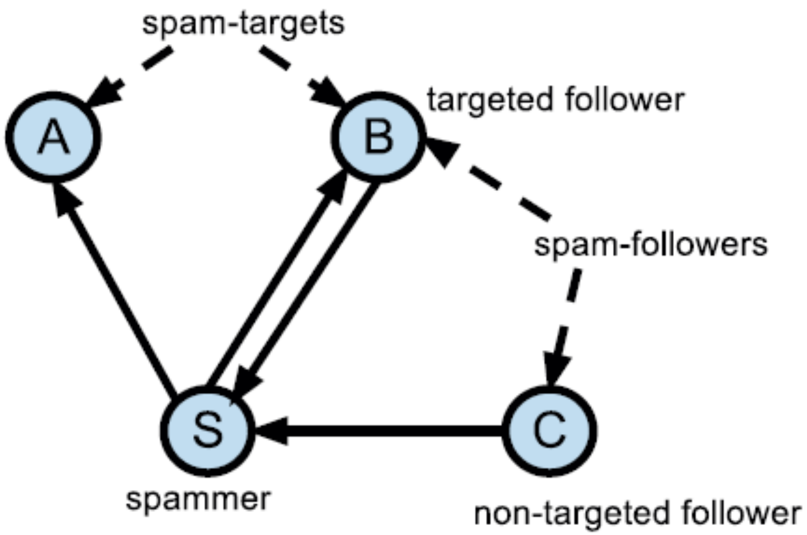
- Link farming in Twitter
  - Spammers follow other users and attempt to get them to follow back

Started by identifying spammers

- Identified 41,352 spammers in Twitter
  - Accounts suspended by Twitter
  - Had posted blacklisted URLs

- Many of the spammer accounts had high number of followers (in-degree)
  - Average in-degree for random user: 36
  - Average in-degree for spammer: 234

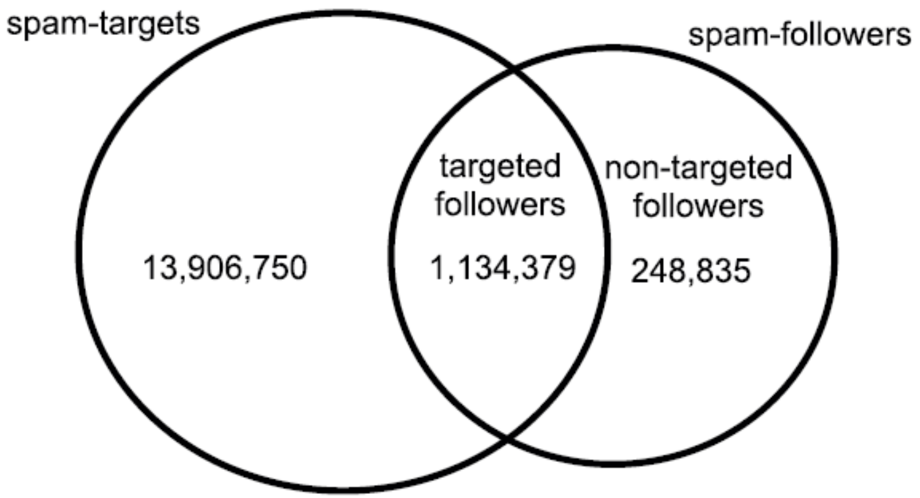
Terminology for spammers' links



- 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



Who are the spam-followers?

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

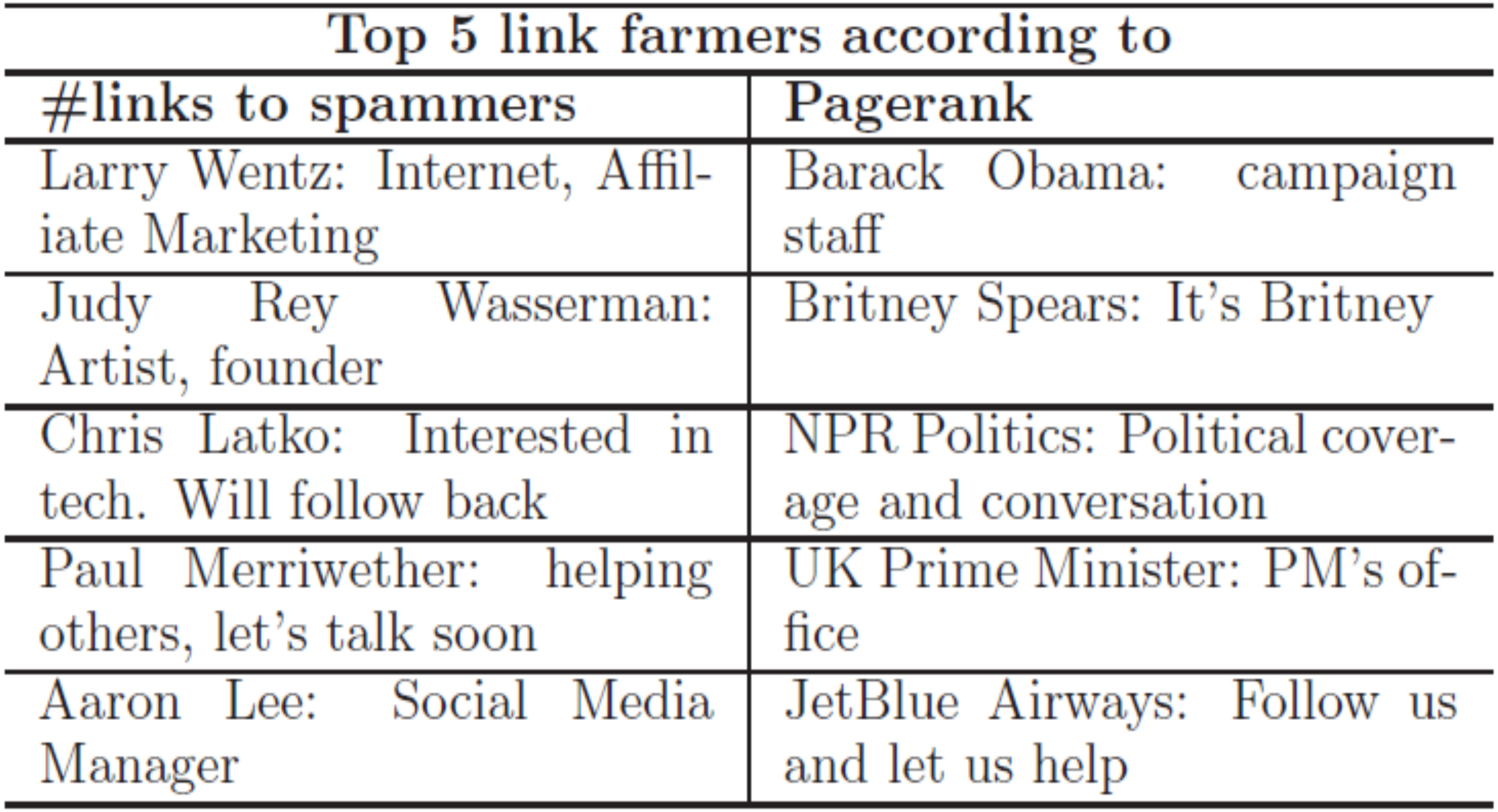
- Targeted spam-followers
  - Ranked on the basis of number of links to spammers
  - 60% of follow-links acquired by spammers come from the top 100,000 targeted followers – LINK FARMERS

- Are the link farmers themselves spammers?

Are link farmers themselves spammers?

- No, over 80% are real, popular, active users
- Most of them are marketers trying to promote their business or some product
- Includes some verified accounts
- Many link farmers within the top 5% users according to PageRank

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

Why are popular users link farming?

- Social etiquette – you follow me, I follow you
- Amass social capital in the network

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 link farmers
  - Within 3 days, 65 reciprocated by following back
  - Our account ranked within the top 9% of all users in Twitter in 3 days !!!

The problem with link farming

- Existence of a set of users from whom social links (hence social influence) can be farmed easily
- Spammers easily gain links from popular users
- Increases the PageRank of spammers as well
- Leads to increases spam in Twitter search results

Combating link farming in Twitter

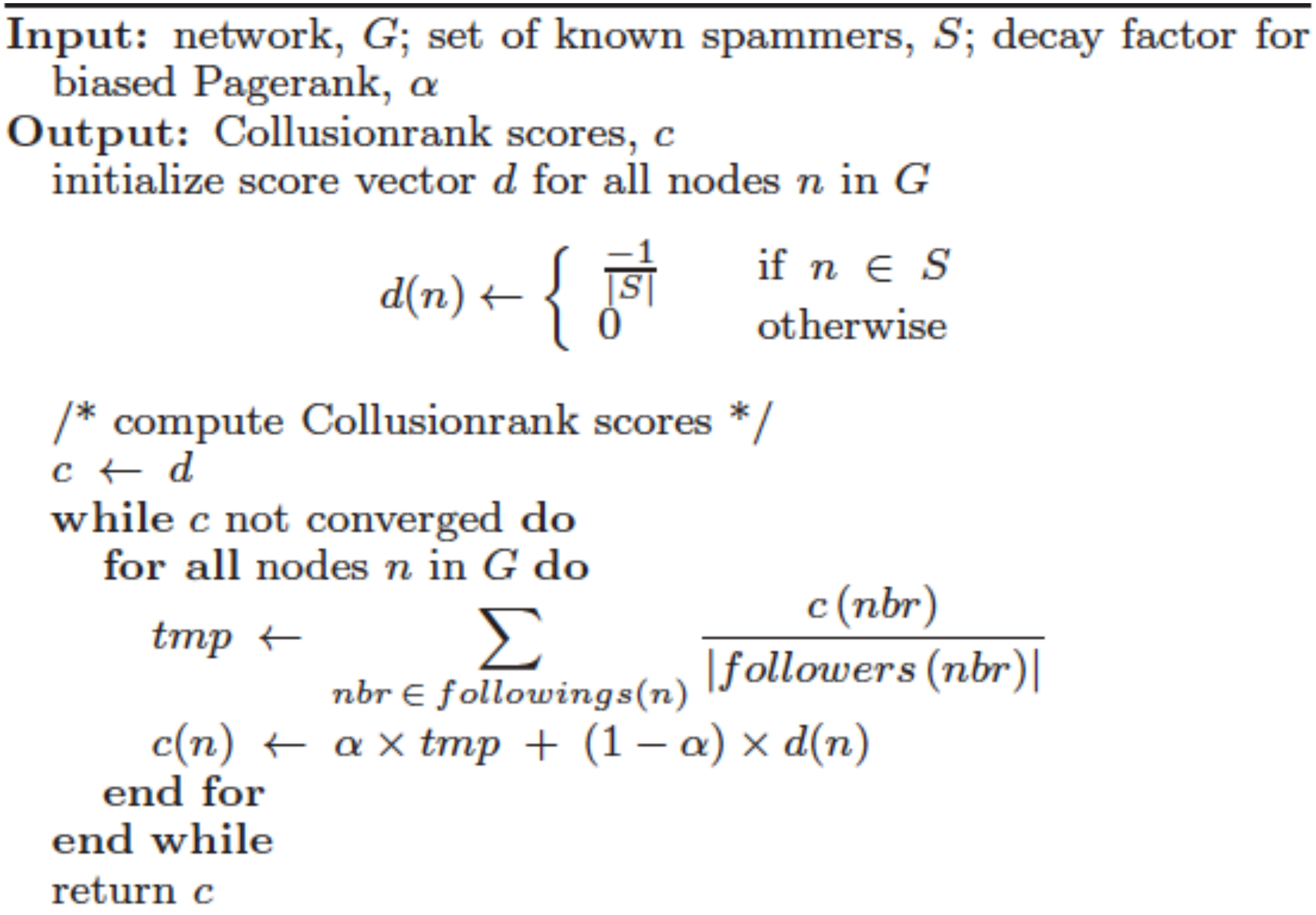
- Key challenges
  - Real, popular users engaged in link farming
  - Detecting and suspending spammers alone will not help

- Discourage users from following others carelessly
  - Penalize users for following someone bad – lower the influence scores of users who follow spammers

CollusionRank

- Identify a seed set of known spammers
- In PageRank style
  - Negatively bias initial scores towards the known spammers
  - Iteratively penalize users who follow spammers, or those who follow spam-followers

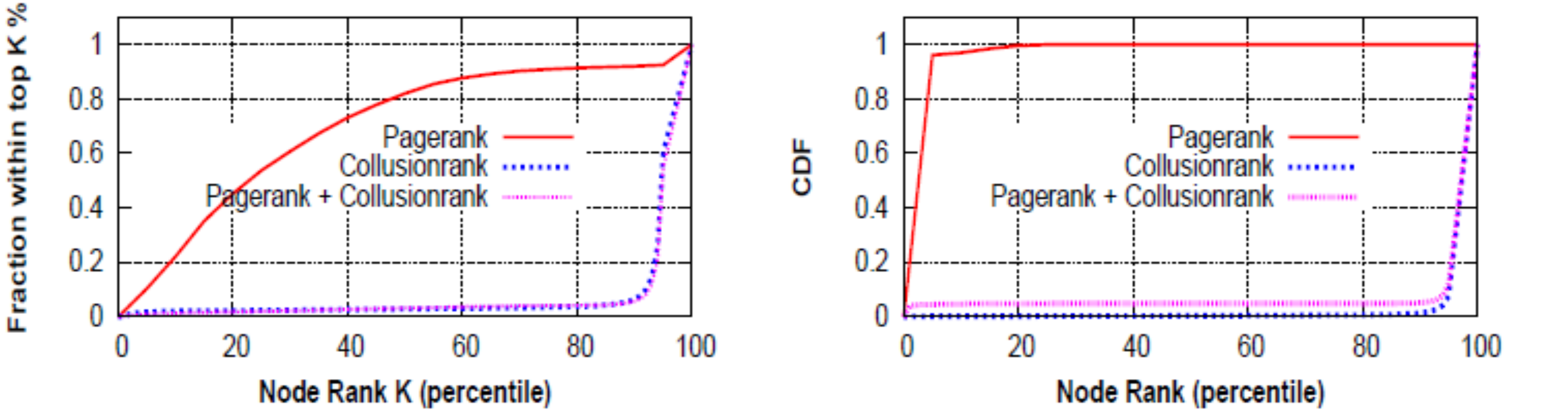
CollusionRank



```
Input: network, $G$; set of known spammers, $S$; decay factor for
  biased Pagerank, $\alpha$
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
      $tmp \leftarrow \sum_{nbr \in followings(n)} \frac{c(nbr)}{|followers(nbr)|}$
      $c(n) \leftarrow \alpha \times tmp + (1 - \alpha) \times d(n)$
    end for
  end while
  return $c$
```

How effective is CollusionRank?

- Compare ranks of spammers and link farmers
  - PageRank
  - CollusionRank
  - PageRank + CollusionRank

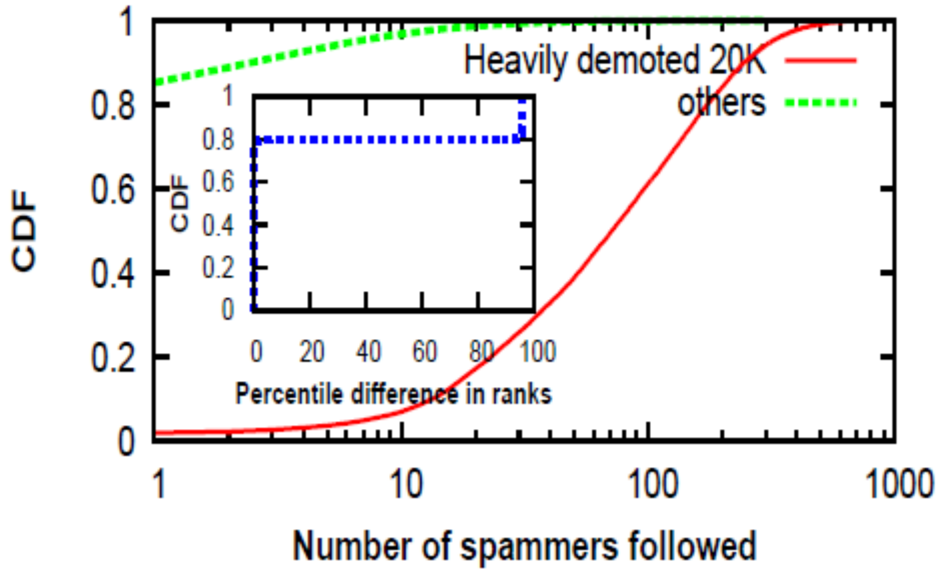


(a) Rankings of all 41,352 spammers

(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)



Case Study 3

Cognos: Crowdsourcing Search for Topic Experts in Microblogs, Ghosh et al., SIGIR 2012

Topical search on Twitter

- Twitter has emerged as an important source of information & real-time news
  - Most common search in Twitter: search for trending topics and breaking news

- Topical search
  - Identifying topical attributes / expertise of users
  - Searching for topical experts
  - Searching for information on specific topics

Prior approaches to find topic experts

- Research studies
  - Pal et. al. (WSDM 2011) uses 15 features from tweets, network, to identify topical experts
  - Weng et. al. (WSDM 2010) uses ML approach

- Application systems
  - Twitter Who To Follow (WTF), Wefollow, …
  - Methodology not fully public, but reported to utilize several features

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, …

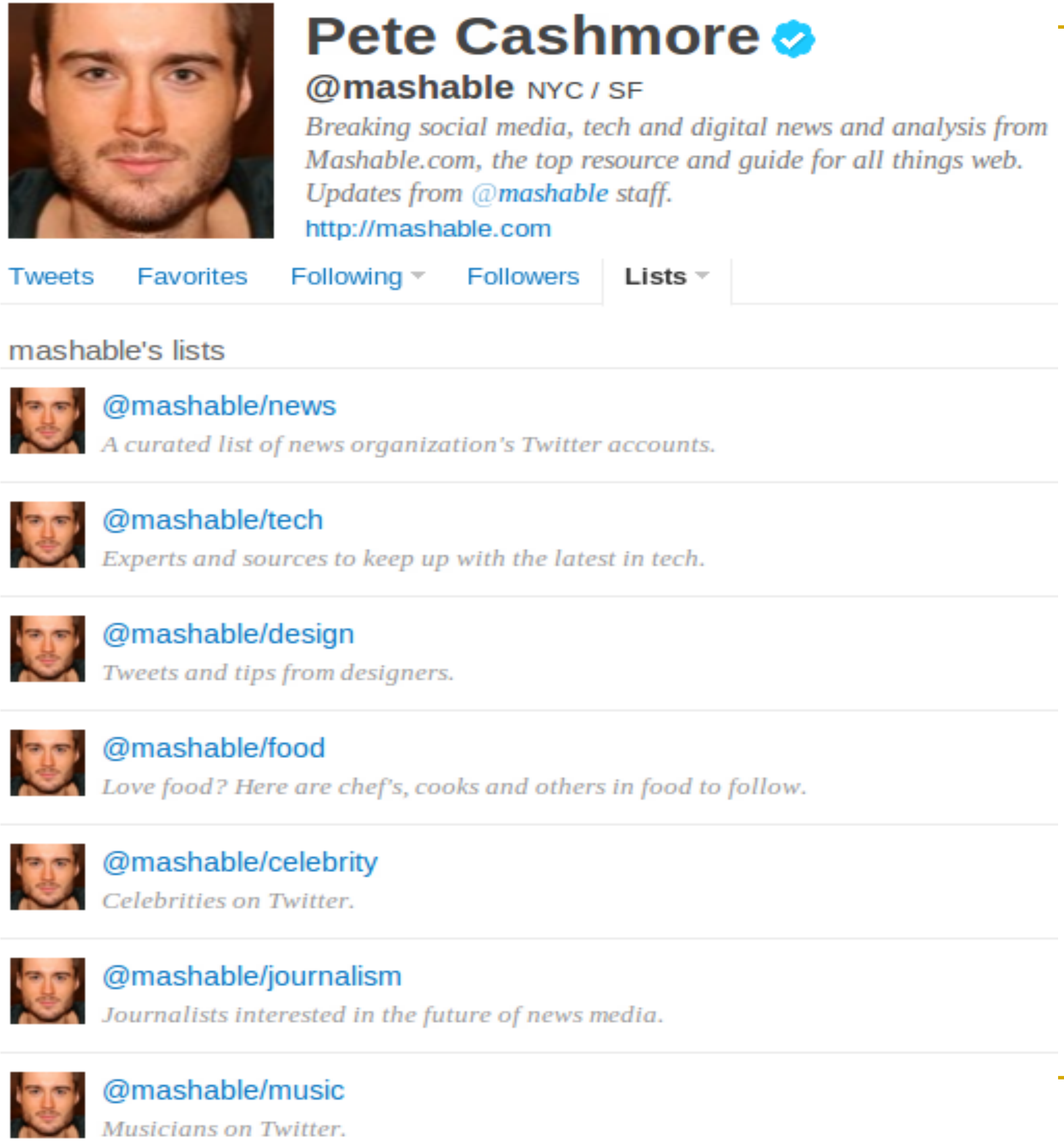
Problems with prior approaches

- User profiles – screen-name, bio, …
  - Bio often does not give meaningful information
  - Information in users profiles mostly unvetted
- Tweets posted by a user
  - Tweets mostly contain day-to-day conversation
- Social graph of a user – #followers, PageRank
  - Does not provide topical information

We proposed

- Use crowdsourcing
  - How does the Twitter crowd describe a user?
  - Social annotations

- Crowdsourced information collected using a feature called Twitter Lists



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

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.

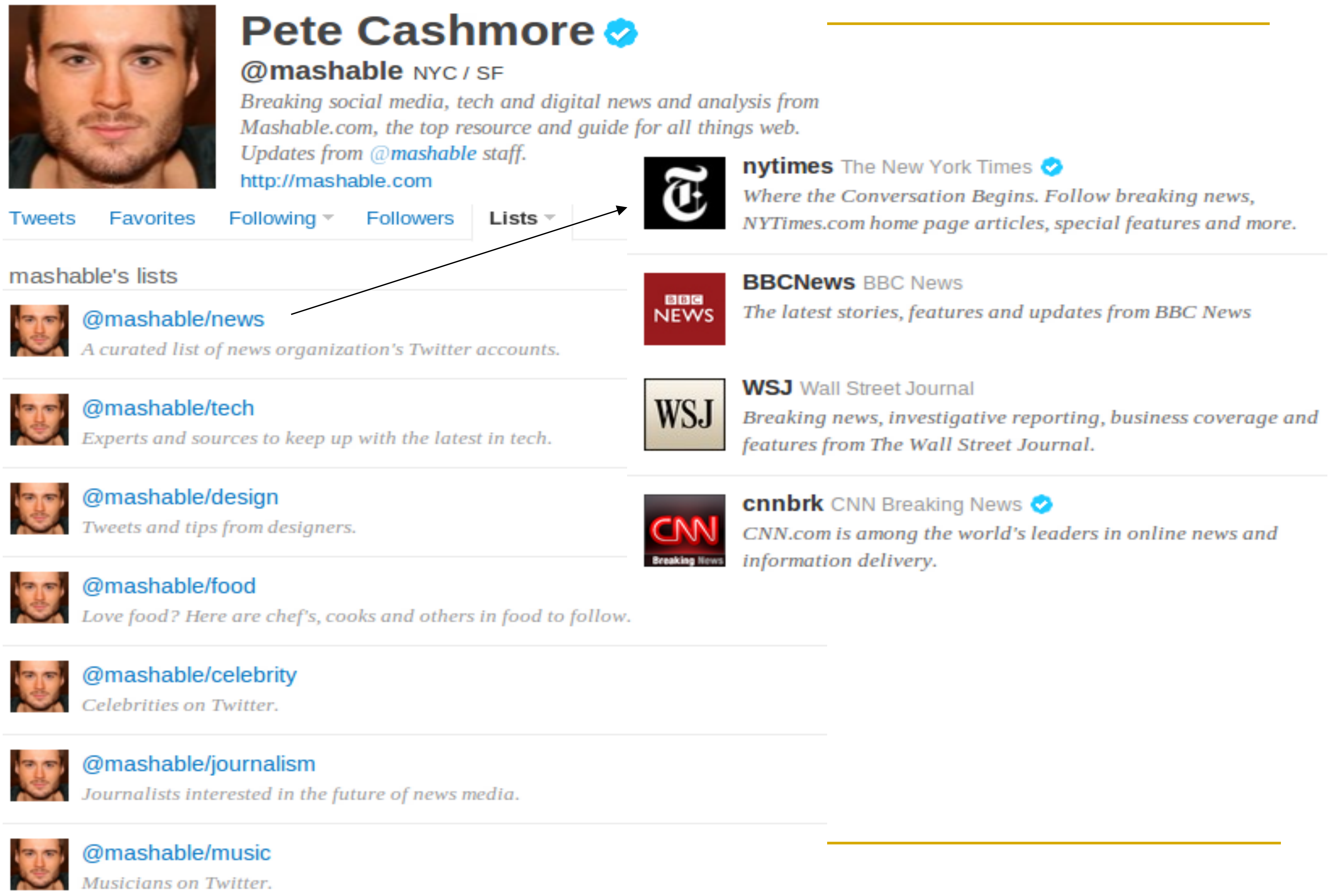
@mashable/design
Tweets and tips from designers.

@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.



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

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.

BBCNews BBC News

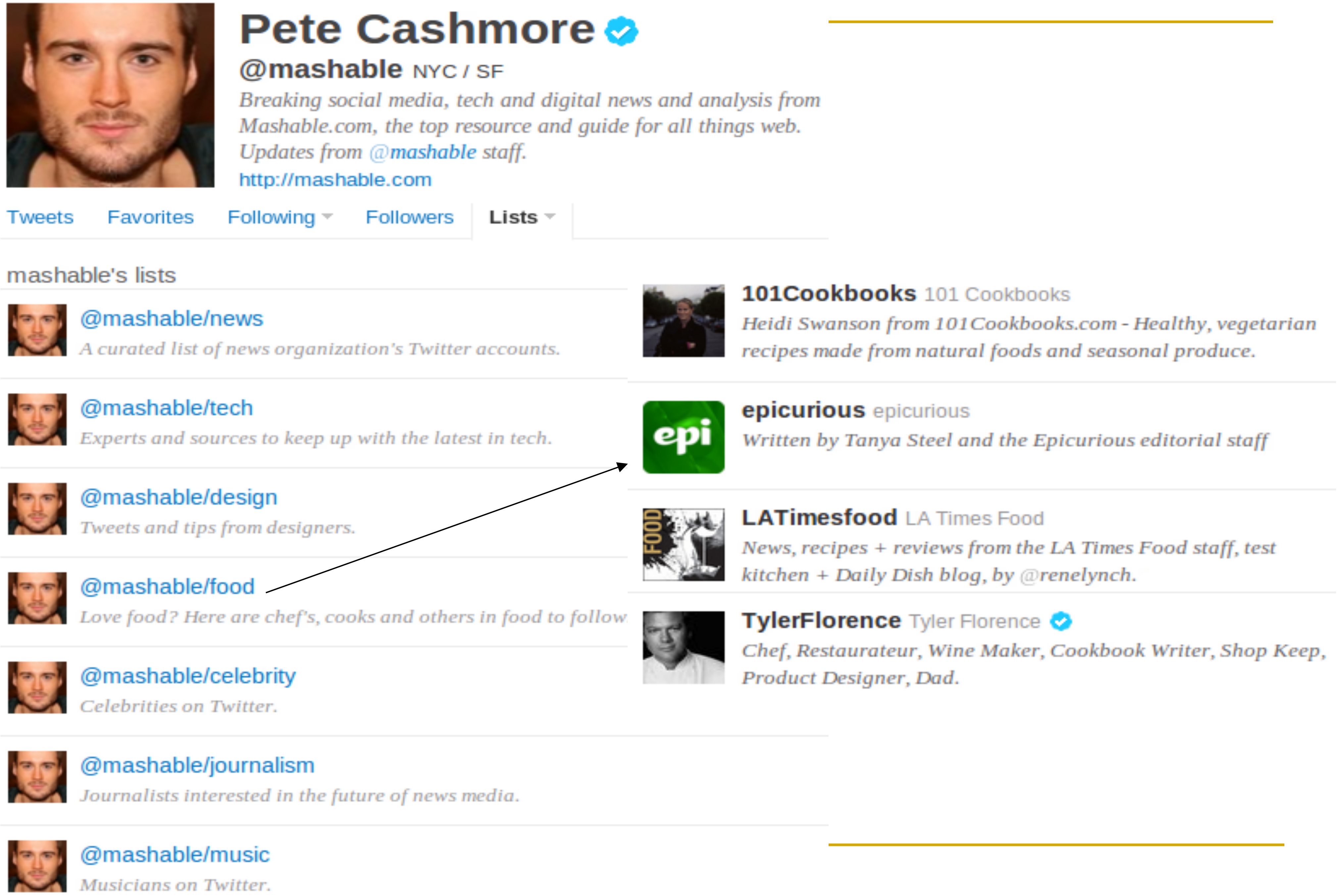
The latest stories, features and updates from BBC News

WSJ Wall Street Journal

Breaking news, investigative reporting, business coverage and features from The Wall Street Journal.

cnnbrk CNN Breaking News

CNN.com is among the world's leaders in online news and information delivery.



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

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.

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

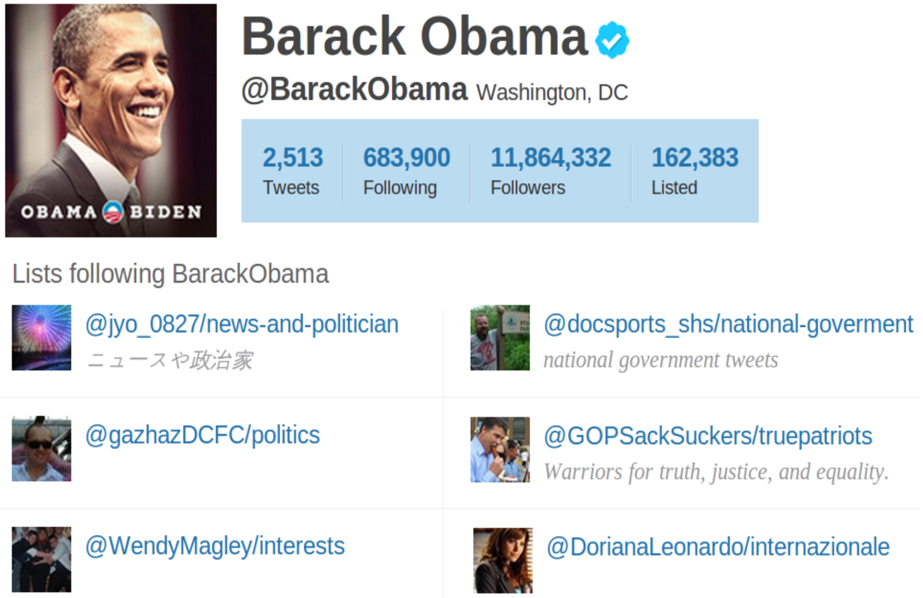
epicurious epicurious
Written by Tanya Steel and the Epicurious editorial staff

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.

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

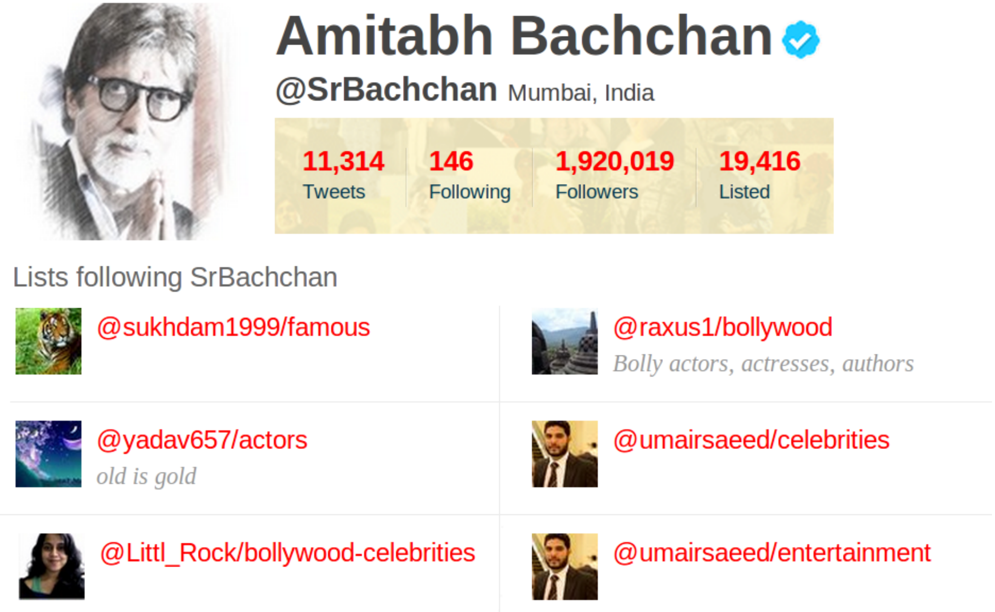
@docsports_shs/national-goverment
national government tweets

@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

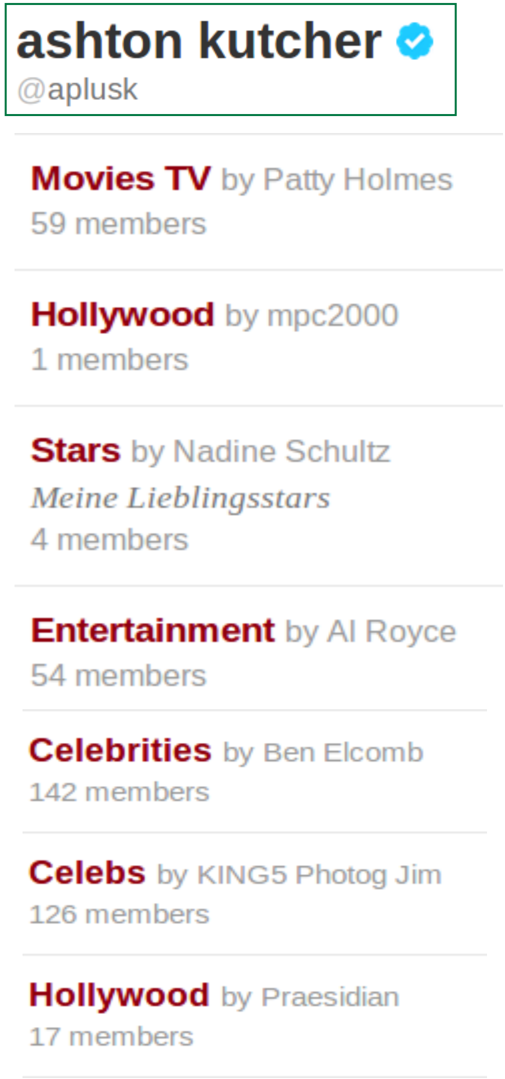
@yadav657/actors
old is gold

@umairsaeed/celebrities

@Littl_Rock/bollywood-celebrities

@umairsaeed/entertainment

Mining Lists to infer expertise



ashton kutcher
@aplusk

Movies TV by Patty Holmes
59 members

Hollywood by mpc2000
1 members

Stars by Nadine Schultz
Meine Lieblingsstars
4 members

Entertainment by Al Royce
54 members

Celebrities by Ben Elcomb
142 members

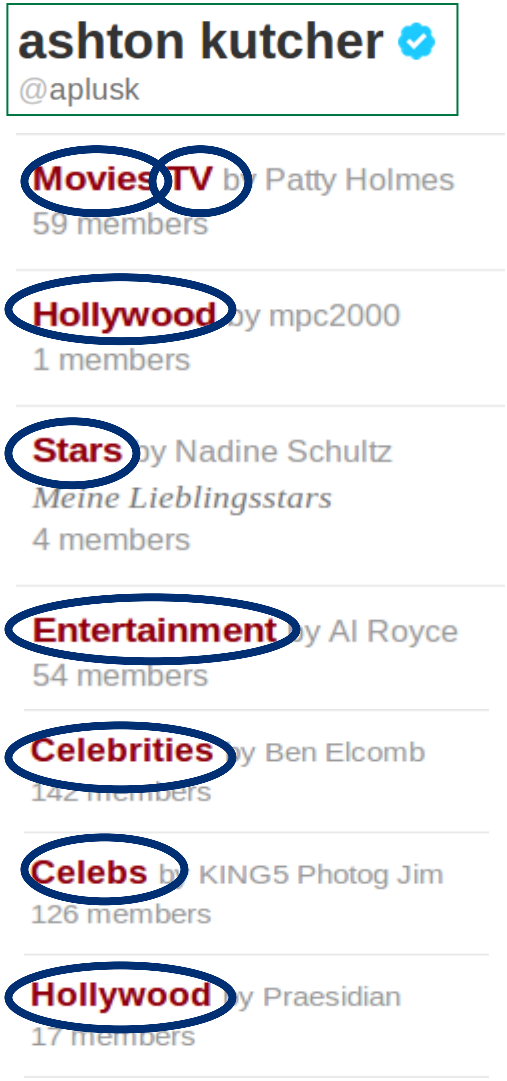
Celebs by KING5 Photog Jim
126 members

Hollywood by Praesidian
17 members

- Collect Lists containing a given user U
- Merge List meta-data to get a ‘topic document’ T_U for U

- Identify U’s topics from T_U
  - Basic IR techniques: case-folding, remove domain-specific stopwords
  - Extract nouns and adjectives using part-of-speech tagger
  - Topics for U: the extracted words along with their frequencies

Mining Lists to infer expertise



ashton kutcher
@aplusk

MoviesTV by Patty Holmes
59 members

Hollywood by mpc2000
1 members

Stars by Nadine Schultz
Meine Lieblingsstars
4 members

Entertainment by Al Royce
54 members

Celebrities by Ben Elcomb
142 members

Celebs by KING5 Photog Jim
126 members

Hollywood by Praesidian
17 members

- Collect Lists containing a given user U
- Merge List meta-data to get a 'topic document' T_U for U

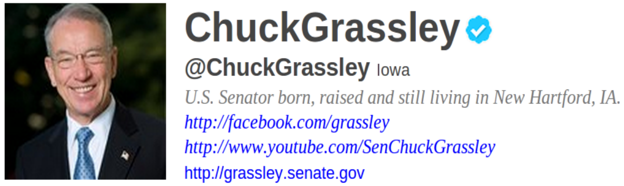
- Identify U's topics from T_U
  - Basic IR techniques: case-folding, remove domain-specific stopwords
  - Extract nouns and adjectives using part-of-speech tagger
  - Topics for U: the extracted words along with their frequencies

Dataset

- Crawled the List-data for all users in our Twitter dataset, in November 2011

- 1.3 million users are included in 10 or more Lists
  - Includes a large majority of the most popular users
  - Our studies focus on this set of users

Topics inferred from Lists



ChuckGrassley

@ChuckGrassley Iowa

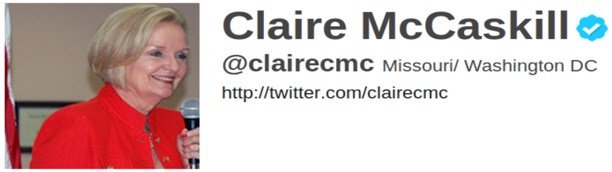
U.S. Senator born, raised and still living in New Hartford, IA.

http://facebook.com/grassley

http://www.youtube.com/SenChuckGrassley

http://grassley.senate.gov

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

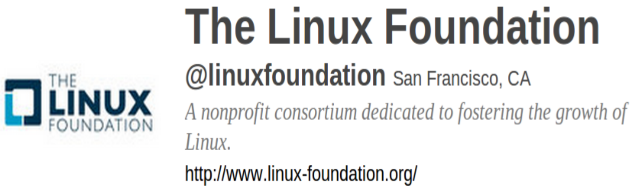


Claire McCaskill

@clairecmc Missouri/ Washington DC

http://twitter.com/clairecmc

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



The Linux Foundation

@linuxfoundation San Francisco, CA

A nonprofit consortium dedicated to fostering the growth of Linux.

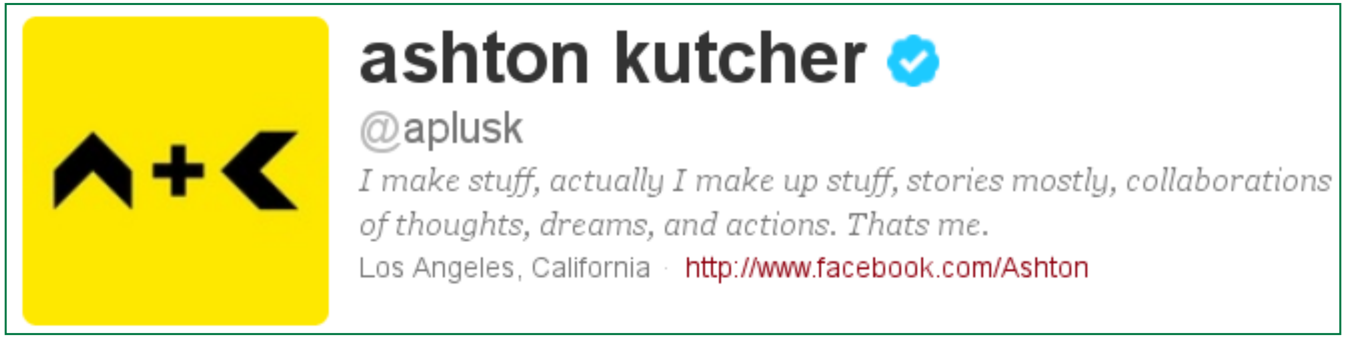
http://www.linux-foundation.org/

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

Evaluating the List-based methodology

- Are the inferred topics (i) accurate (ii) informative?
- Evaluated using feedback through a user-survey
- More than 93% evaluators judged the topics to be both accurate and informative
  - The few negative judgments were a result of subjectivity

Lists work better than other features



ashton kutcher

@aplusk

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

Los Angeles, California · http://www.facebook.com/Ashton

Profile bio

love, daily, people, time, GUI, movie, video, life, happy, game, cool

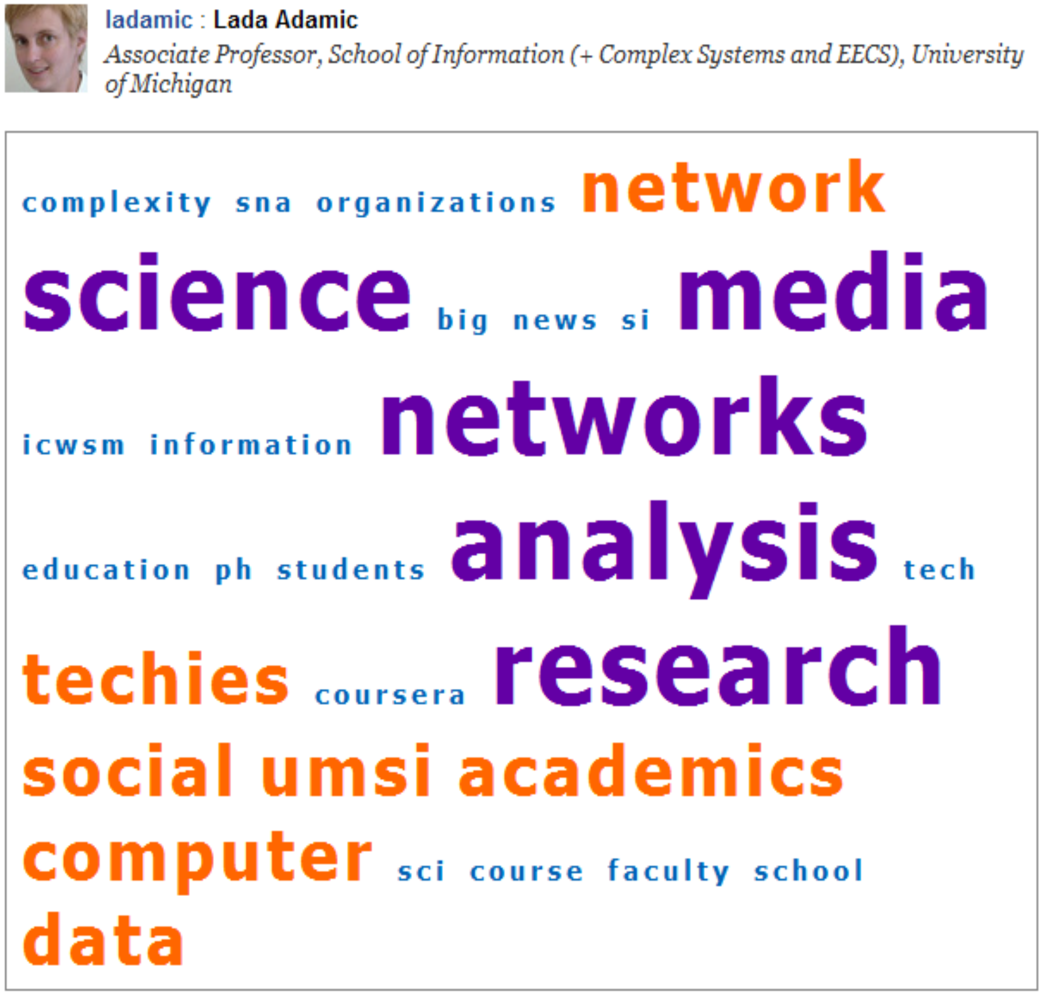
Most common words from tweets

celeb, actor, famous, movie, stars, comedy, music, Hollywood, pop culture

Most common words from Lists

Who-is-who service

- Developed a Who-is-Who service for Twitter
  - Shows word-cloud for major topics for a given user
  - http://twitter-app.mpi-sws.org/who-is-who/



ladamic : Lada Adamic

Associate Professor, School of Information (+ Complex Systems and EECS), University of Michigan

complexity sna organizations network
science big news si media
icwsm information networks
education ph students analysis tech
techies coursera research
social umsi academics
computer sci course faculty school
data

Search system for topic experts

- Given a query (topic)
  - Identify users related to the topic using Lists
  - Rank identified users

Ranking experts

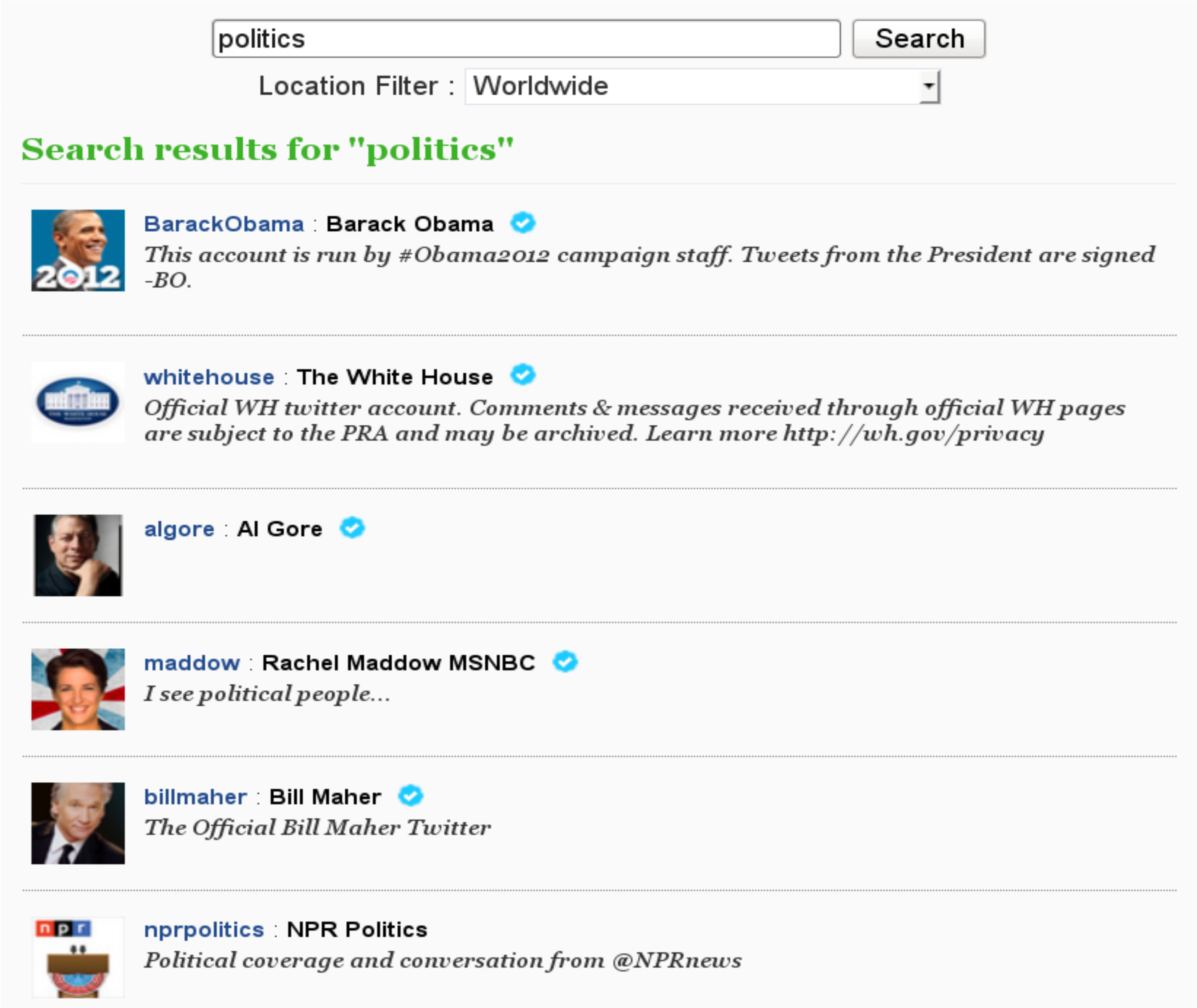
- Used a ranking scheme solely based on Lists
- Two components of ranking user U w.r.t. query Q
  - Relevance of user to query – cover density ranking between topic document T_U of user and Q
  - Popularity of user – number of Lists including the user

$$\text{Topic relevance}(T_U, Q) \times \log(\#\text{Lists including } U)$$

Search system for topic experts

Cognos, a search system for topic experts

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



Search results for "politics"

BarackObama : Barack Obama ✓
This account is run by #Obama2012 campaign staff. Tweets from the President are signed -BO.

whitehouse : The White House ✓
Official WH twitter account. Comments & messages received through official WH pages are subject to the PRA and may be archived. Learn more http://wh.gov/privacy

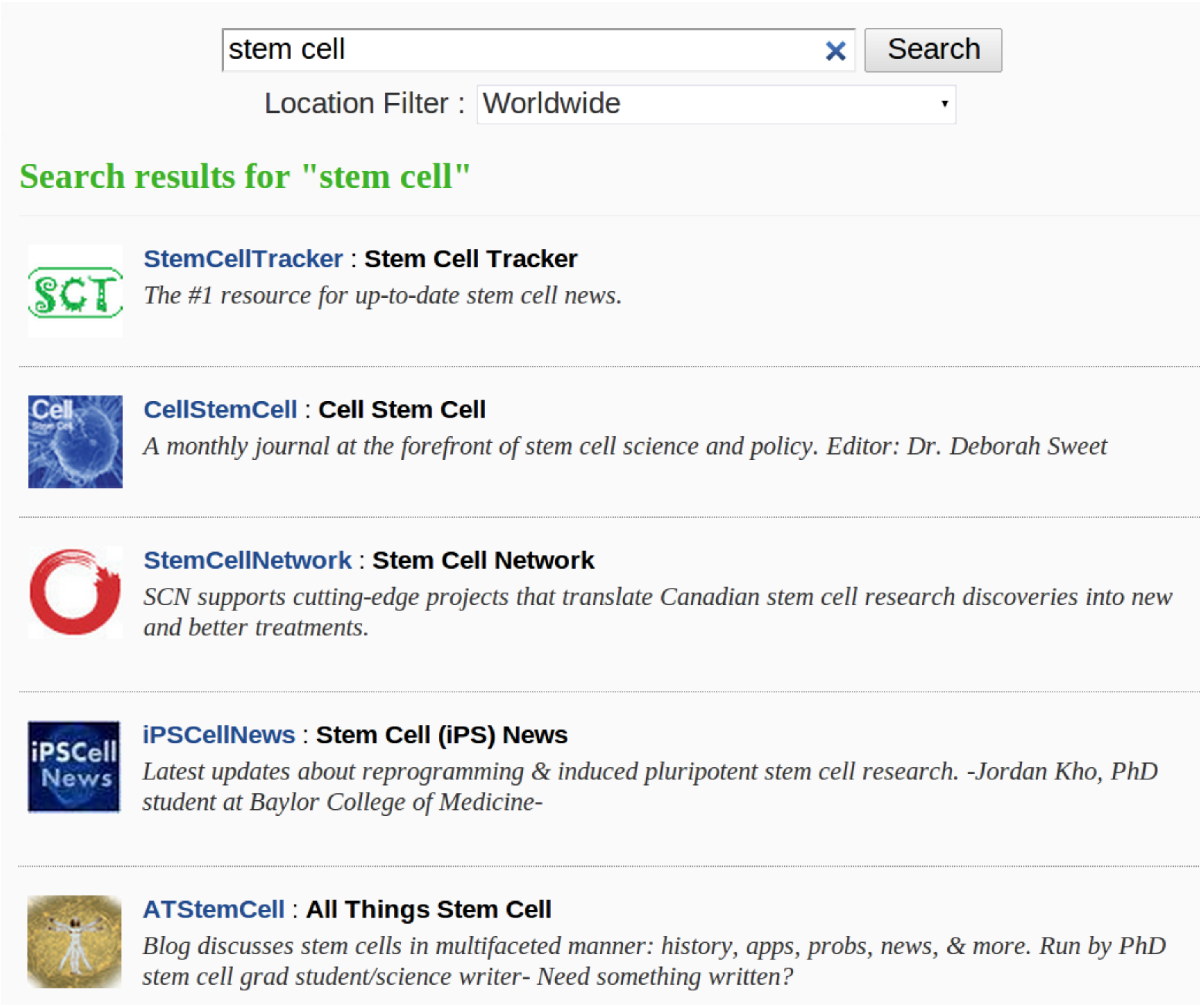
algore : Al Gore ✓

maddow : Rachel Maddow MSNBC ✓
I see political people...

billmaher : Bill Maher ✓
The Official Bill Maher Twitter

nprpolitics : NPR Politics
Political coverage and conversation from @NPRnews

Cognos results for “politics”



Location Filter : Worldwide

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

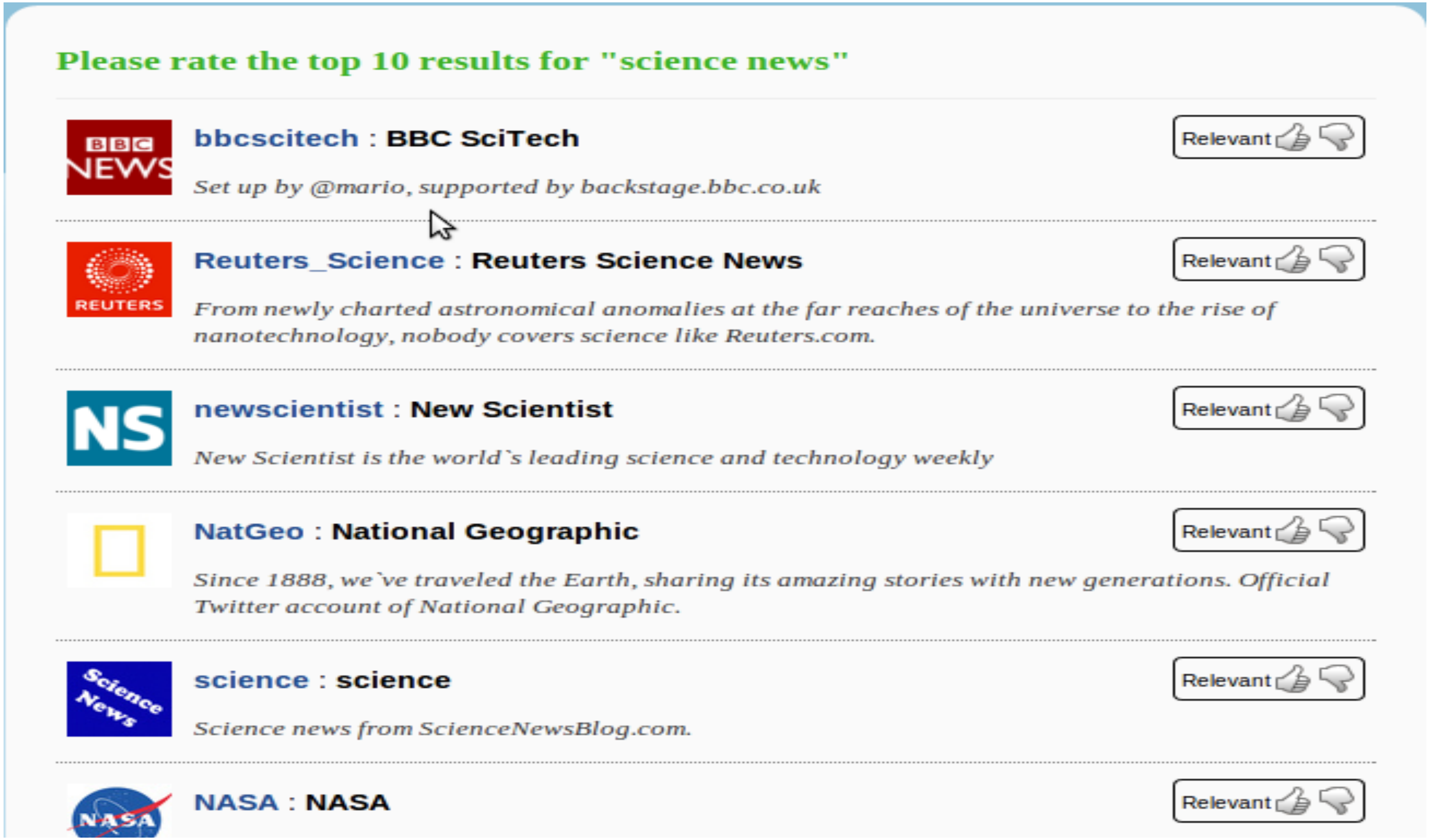
StemCellNetwork : **Stem Cell Network**
SCN supports cutting-edge projects that translate Canadian stem cell research discoveries into new and better treatments.

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 “stem cell”

User-evaluation of Cognos



Please rate the top 10 results for "science news"

bbcscitech : BBC SciTech

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

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.

newscientist : New Scientist

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

NatGeo : National Geographic

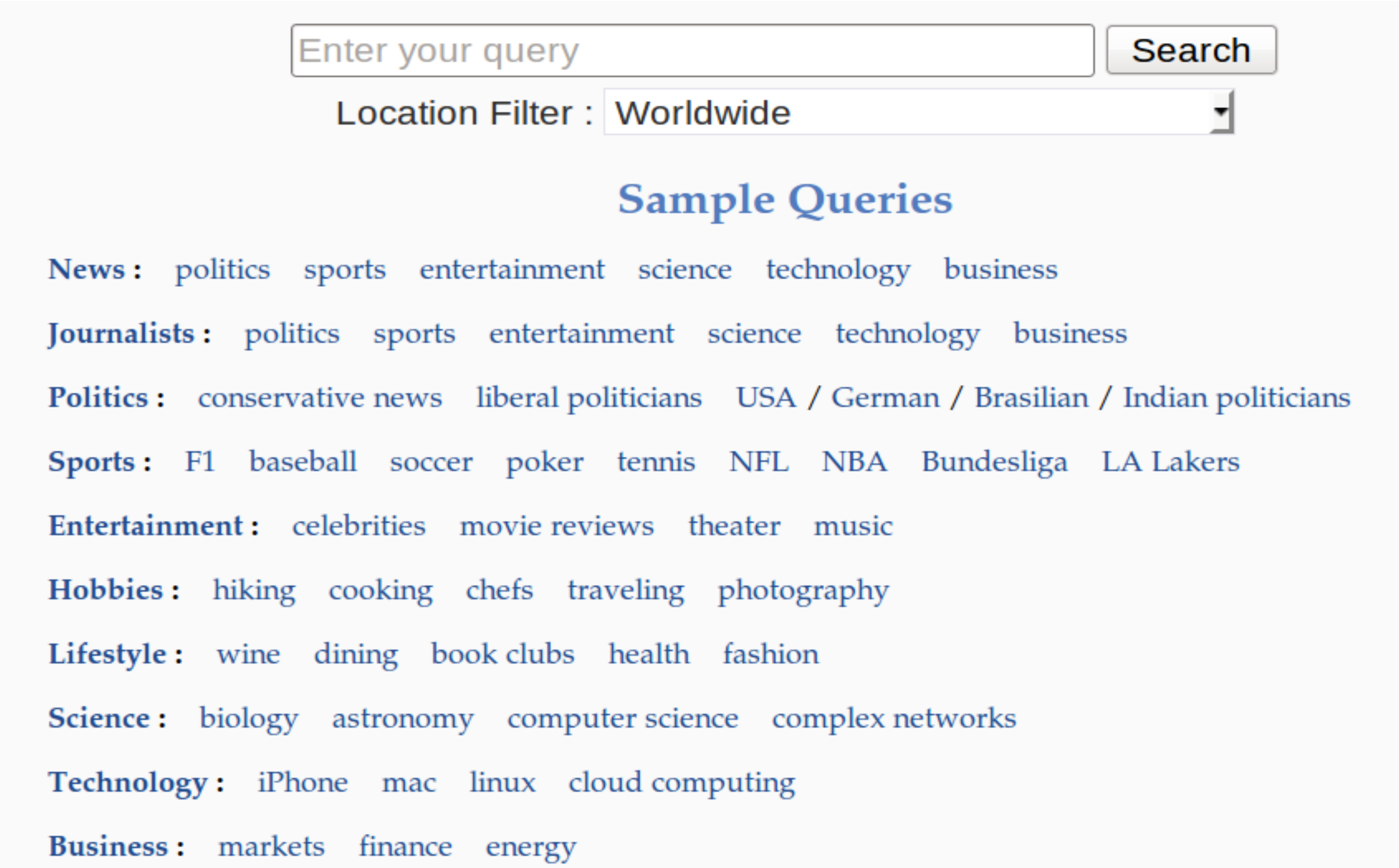
Since 1888, we`ve traveled the Earth, sharing its amazing stories with new generations. Official Twitter account of National Geographic.

science : science

Science news from ScienceNewsBlog.com.

NASA : NASA

Sample queries for evaluation



Location Filter : Worldwide

Sample Queries

News : politics sports entertainment science technology business

Journalists : politics sports entertainment science technology business

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

Sports : F1 baseball soccer poker tennis NFL NBA Bundesliga LA Lakers

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

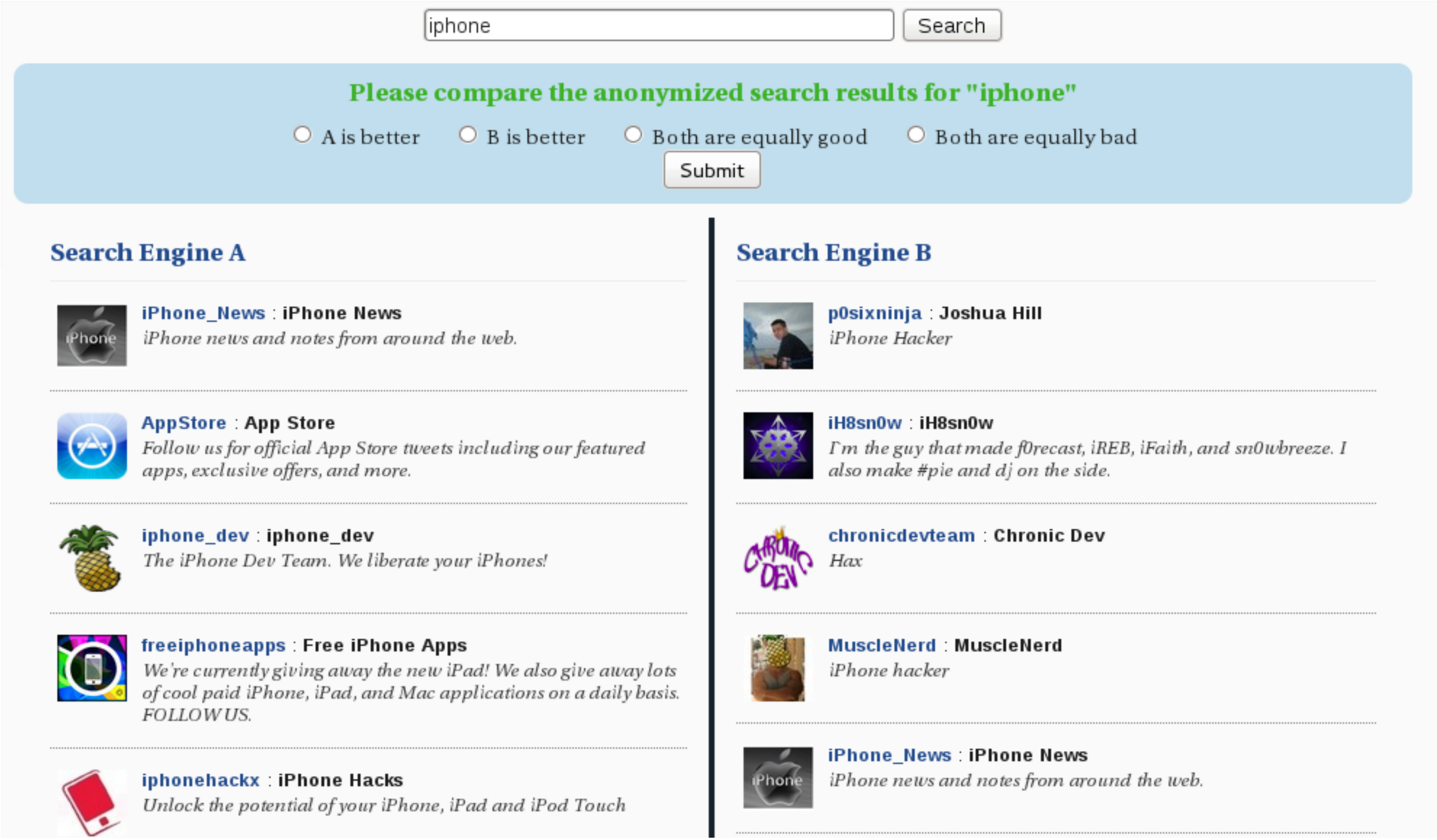
Business : markets finance energy

Evaluation results

- Overall 2136 relevance judgments
  - 1680 said relevant (78.7%)

- Large amount of subjectivity in evaluations
  - Same result for same query received both relevant and non-relevant judgments
  - E.g., for query “cloud computing”, Werner Vogels got 4 relevant judgments, 6 non-relevant judgments

Cognos vs. Twitter Who-To-Follow



iphone Search

Please compare the anonymized search results for "iphone"

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

Submit

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. FOLLOW US.

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.

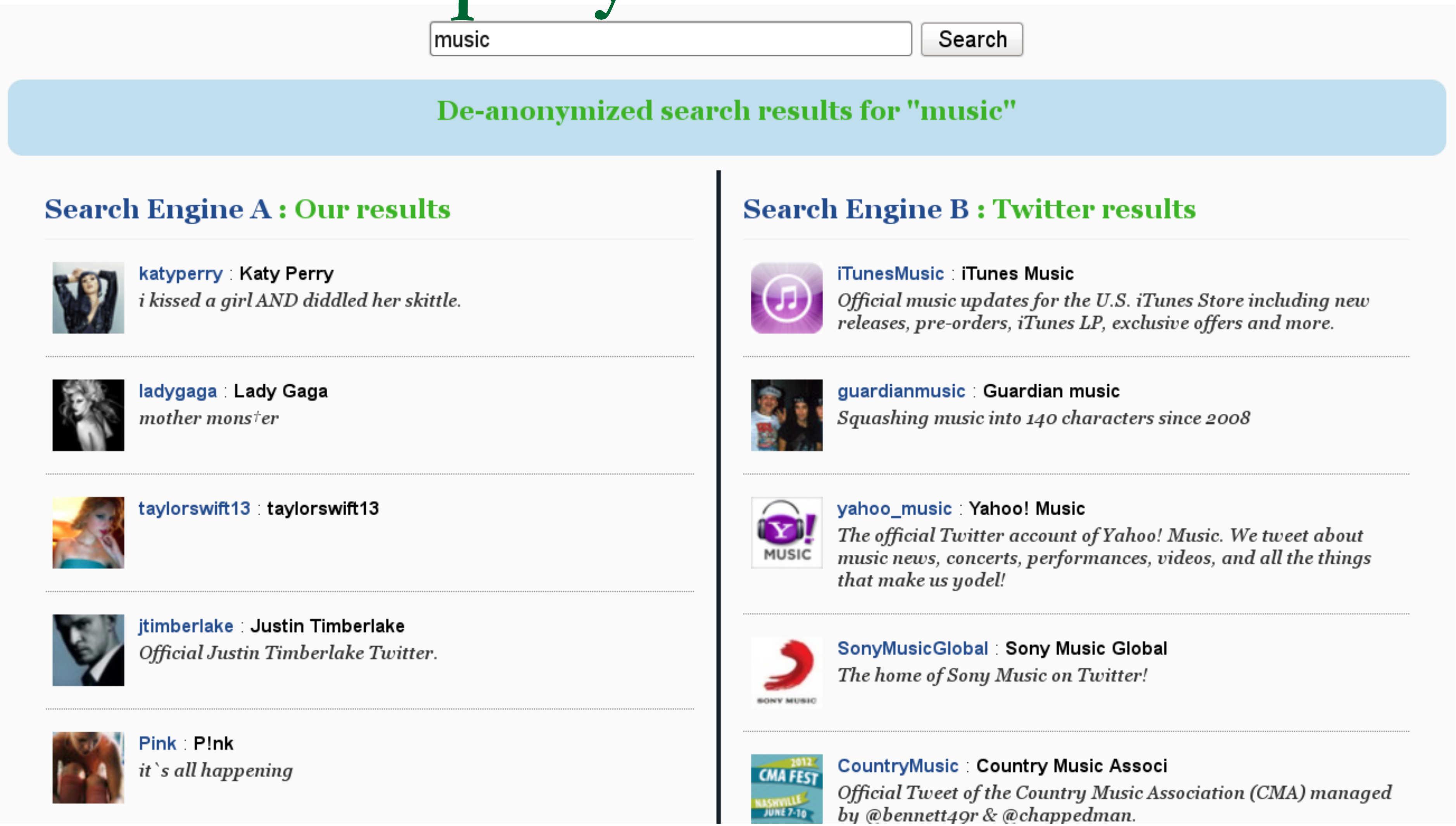
Cognos vs. Twitter Who-To-Follow

- Considering 27 distinct queries asked at least twice
- Judgment by majority voting

- Cognos judged better on 12 queries
  - Computer science, Linux, Mac, Apple, Ipad, Internet, Windows phone, photography, political journalist, …

- Twitter Who-To-Follow judged better on 11 queries
  - Music, Sachin Tendulkar, Anjelina Jolie, Harry Potter, metallica, cloud computing, IIT Kharagpur, …

Results for query music



De-anonymized search results for "music"

Search Engine A : Our results

katyperry : Katy Perry
i kissed a girl AND diddled her skittle.

ladygaga : Lady Gaga
mother mons†er

taylorswift13 : taylorswift13

jtimberlake : Justin Timberlake
Official Justin Timberlake Twitter.

Pink : P!nk
it`s all happening

Search Engine B : Twitter results

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

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!

SonyMusicGlobal : Sony Music Global
The home of Sony Music on Twitter!

CountryMusic : Country Music Associ
Official Tweet of the Country Music Association (CMA) managed by @bennett49r & @chappedman.

Scalability problem

- Twitter now has around 500 million users
- 740K new users join daily
- How to keep the system up-to-date by discovering newly joining experts?
- Twitter restricts crawling through API
    - Brute-force crawl of all users is infeasible

Solution

- Only 1.1% users are listed 10 or more times
  - If experts can be identified efficiently, possible to crawl their Lists
- Used hubs to identify authorities / experts
  - Hubs – users who selectively List many experts
  - Identify hubs using HITS, crawl Lists created by top hubs
  - 50% of users listed by top 2% hubs listed 10 or more times

Details in paper