# An Introduction to Social Mining

Vladimir Gorovoy* and Yana Volkovich†

†@yvolkovich
Barcelona Media, Information, Technology & Society Group
Barcelona, Spain

* @vgorovoy
Yandex, Yandex.Uslugi
Saint Petersburg, Russia

August, 15-19 2011

RuSSIR
Russian Summer School in Information Retrieval

# Outline

1. Ranking Twitter
2. Location and social networks

# Ranking Twitter

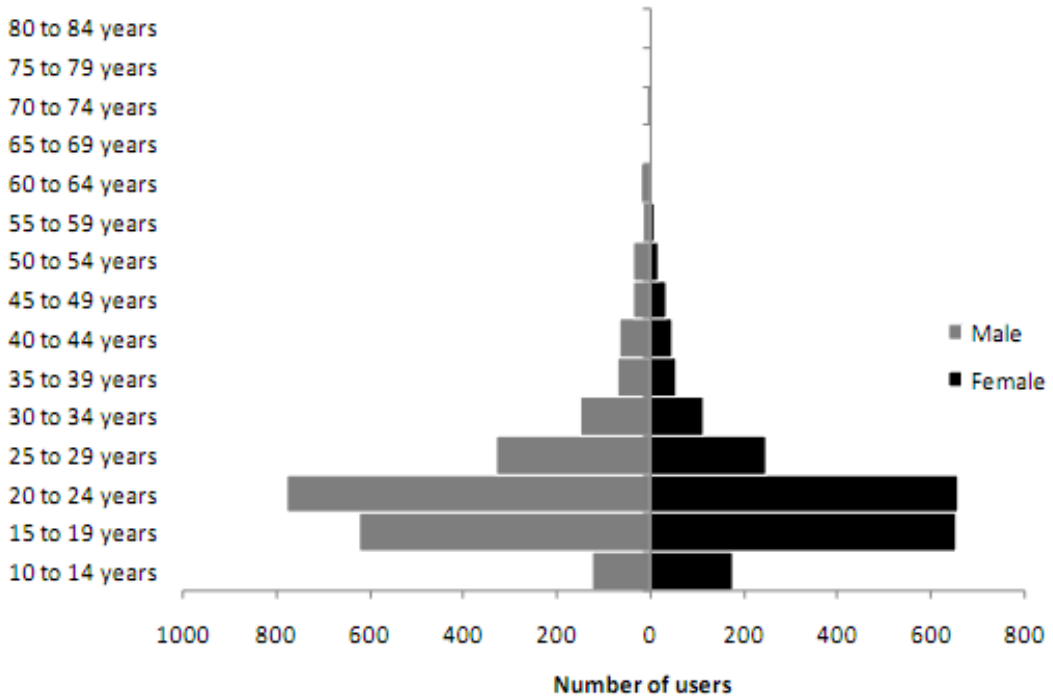
## Twitter

- **Twitter** is an online service that allows users to publish text-based post up to 140 characters (“tweets”).
- Twitter was launched in 2006;
- Now: 200 million users;
  180 million tweets and 1.6 billion search queries per day

RuSSIR
Russian Summer School in Information Retrieval

## Who?

- 5% of twitter users create 75% of the content;
- 54% of Twitter users are female, 46% users are male;



## What?

- 40% of tweets are **pointless babble** (“I’m eating a sandwich”) [Pearanalytics, 2009]

| By Time of Day | 11:00 | 11:30 | 1:30 | 2:00 | 2:30 | 3:00 | 3:30 | 4:00 | 4:30 | 5:00 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| News | 9 | 7 | 5 | 11 | 9 | 5 | 8 | 7 | 5 | 6 | **72** |
| Spam | 9 | 7 | 10 | 4 | 4 | 10 | 7 | 8 | 6 | 10 | **75** |
| Self promotion | 14 | 12 | 11 | 15 | 13 | 11 | 13 | 6 | 12 | 10 | **117** |
| Pointless Babble | 71 | 78 | 79 | 82 | 78 | 96 | 88 | 85 | 73 | 81 | **811** |
| Conversational | 76 | 71 | 76 | 72 | 84 | 66 | 66 | 73 | 84 | 83 | **751** |
| Pass along value | 21 | 25 | 19 | 16 | 12 | 12 | 18 | 21 | 20 | 10 | **174** |
| **Totals** | **200** | **200** | **200** | **200** | **200** | **200** | **200** | **200** | **200** | **200** | **2000** |

Ranking Twitter

# Ranking Twitter

## # followers (1)

- # followers;

1 **Lady Gaga** (@ladygaga)
mother monster
12,736,208 followers | 142,052 following | 1,001 tweets
Follow

2 **Justin Bieber** (@justinbieber)

NEVER SAY NEVER 3D out on DVD NOW!! SOMEDAY out June 20th...and thank you to the fans...who are always there for me and I will always be there for them
11,956,331 followers | 116,205 following | 10,310 tweets
Follow

3 **Barack Obama** (@BarackObama)
This account is run by #Obama2012 campaign staff. Tweets from the President are signed -BO.
9,717,786 followers | 691,235 following | 1,725 tweets
Follow

4 **Katy Perry** (@katyperry)

i kissed a girl AND diddled her skittle.
9,210,251 followers | 78 following | 3,246 tweets
Follow

5 **Kim Kardashian** (@KimKardashian)

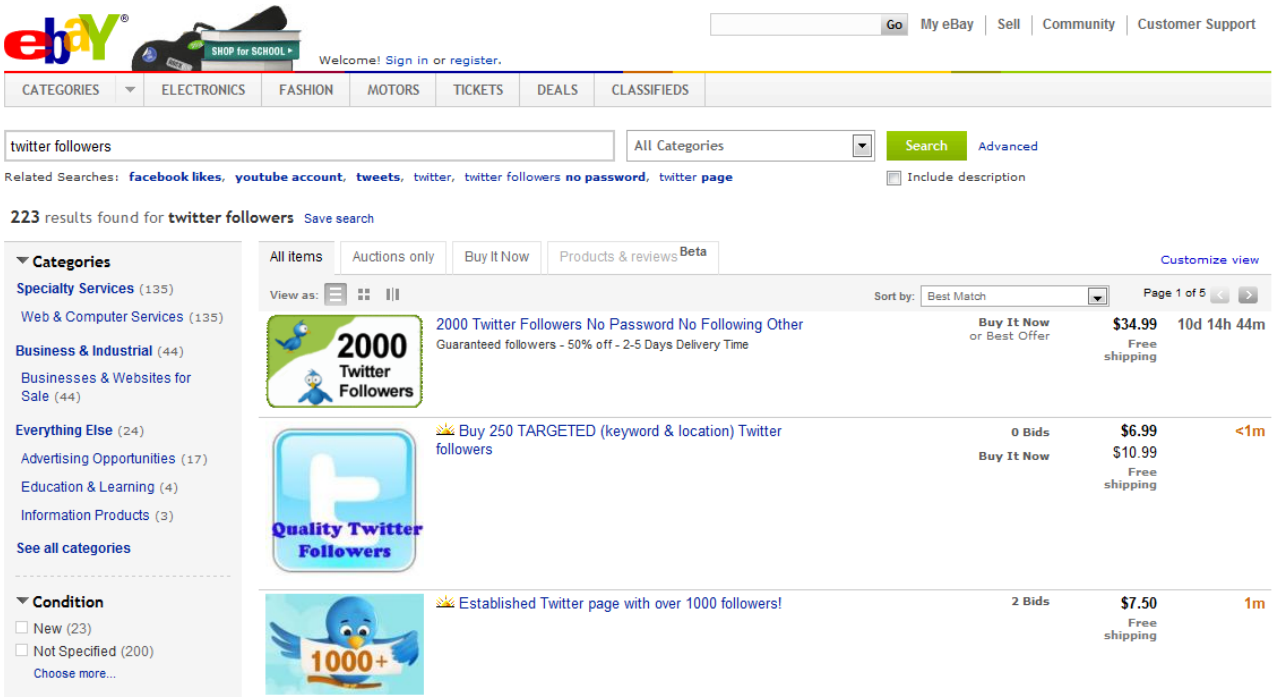
business woman, exec producer, fashion designer, perfumista
9,087,602 followers | 135 following | 8,629 tweets
Follow

twittercounter.com/pages/100

# Ranking Twitter

## # followers (2)

- spammers have far more followers than average users [Yardi et al., 2010];



# Ranking Twitter

## #followers/#followee

- 72% of users follow more than 80% of their followers;
  80% of users have 80% of their friends follow them back. [Weng et al., 2010].

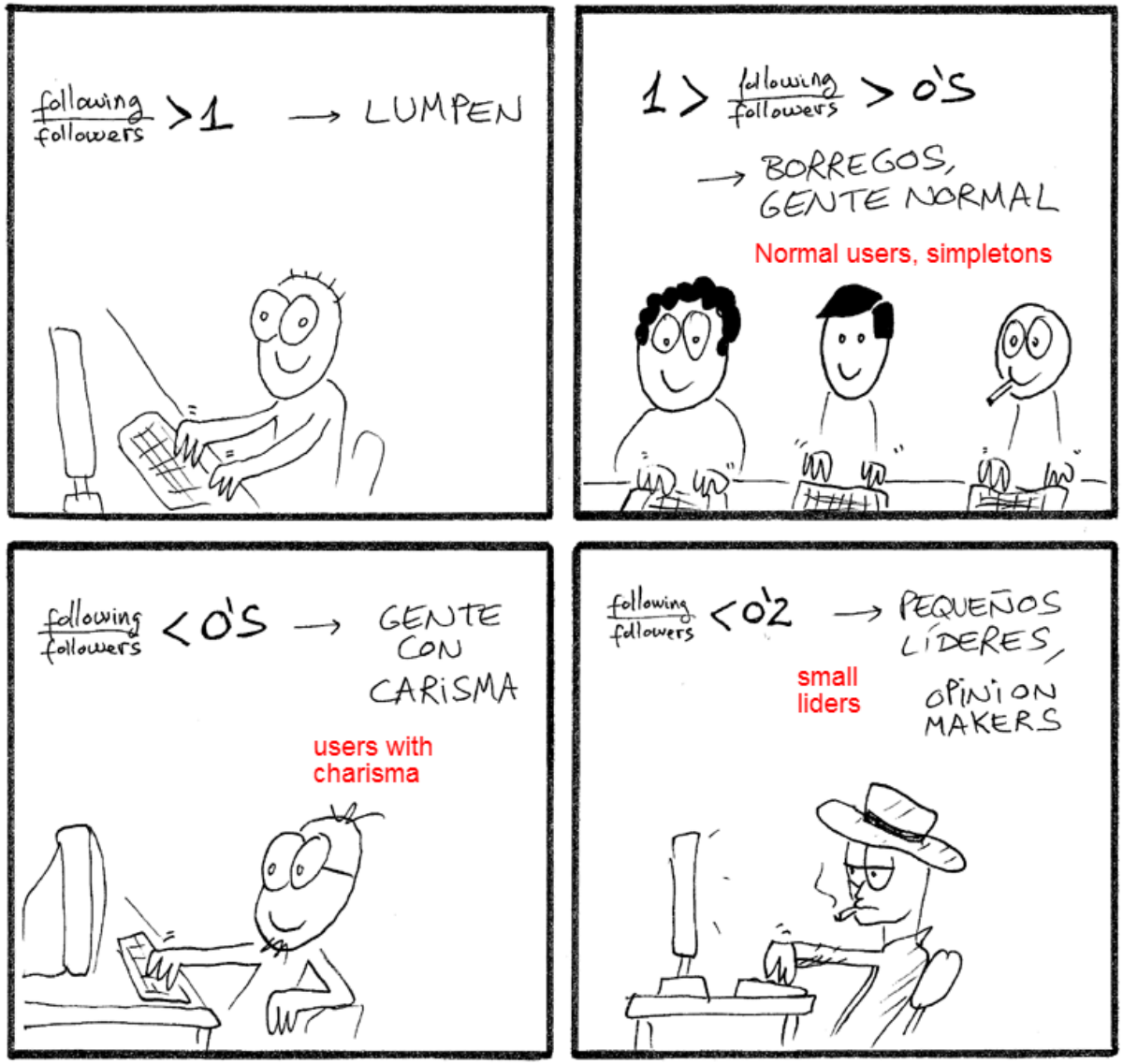
# Ranking Twitter

## #followers/#followee

- $\frac{\#followers}{\#followee}$ ratio;

# Ranking Twitter

## #followers/#followee



# Ranking Twitter

## #followers/#followee

- ratio: Oprah: $1.67 * 10^5$;
  CNN Breaking News: $1.04 * 10^5$;
  Lady Gaga 18.08;
  from [Gayo-Avello and Brenes, 2010];
- discounted ratio $= \frac{\#followers - reciprocal}{\#followee - reciprocal}$

RuSSIR
Russian Summer School in Information Retrieval

# Ranking Twitter

## Other techniques

- PageRank;
- TunkRank;
- TwitterRank;

# Ranking Twitter

## TunkRank

- **TunkRank** by Daniel Tunkelang (`tunkrank.com`)
- assumptions:

  (1) every user has a given *influence* that is a numerical estimator of the number of people who will read his tweets;
  (2) users' attention to their followees is equally distributed;
  (3) user $X$ will retweet a tweet by user $Y$ with defined probability $p^{retweet}$;
- just 2% of tweets are retweets (Dan Zarrella, "The science of ReTweets report");
  2,87% in [Gayo-Avello and Brenes, 2010].

RuSSIR
Russian Summer School in Information Retrieval

# Ranking Twitter

## TunkRank

- TunkRank:

$$Influence(X) = \sum_{Y \in Followers(X)} p^{notice} \left( \frac{1 + p^{retweet} \cdot Influence(Y)}{|Following(Y)|} \right)$$

- $p^{notice}$ is total attention of the user devoted to Twitter;
- $p^{retweet}$ is retweet probability;

# Ranking Twitter

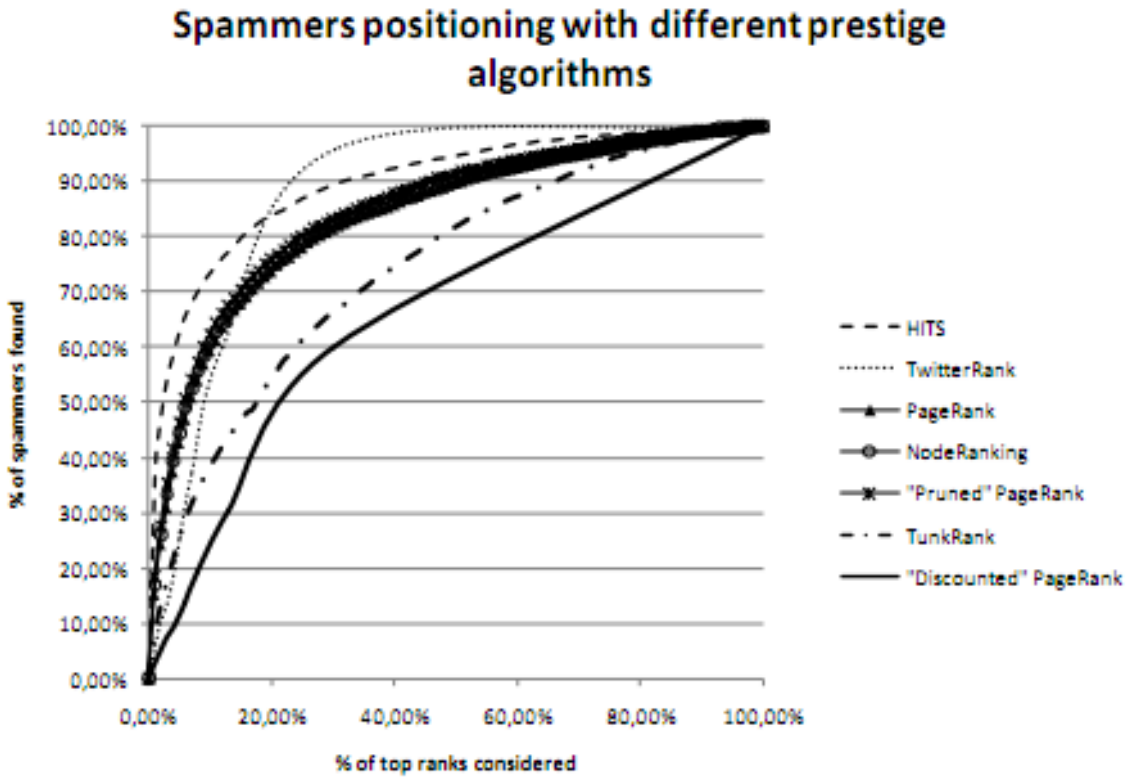
## Twitter rank

- TwitterRank [Weng et al., 2010]: to rank users separately for different topics
- PageRank + topical similarity between users;

# Ranking Twitter

## Comparison

- comparison [Gayo-Avello and Brenes, 2010]:



# Twitter study

## Spanish revolution

- **15-M Movement**: series of peaceful demonstrations in Spain.
- several weeks of sit-ins in 58 cities.

RuSSIR
Russian Summer School in Information Retrieval

Location and social networks.

People will to share their location.

# Location-based social networks

## Obama joins Foursquare

### President Obama Joins Foursquare

10 hours ago by Ben Parr

48

Like Send 969 people like this. Be the first of your friends.

Foursquare has just gained its highest-profile user yet: President Barack Obama.

"The White House is now on Foursquare, a location-based social networking website, which is the latest way for you to engage with the administration," The White House said on its blog. "There are over 10 million people already "checking in" around the world, and now you'll be able to discover "tips" from the White House featuring the places President Obama has visited, what he did there, plus historical information and more."

# Location-based social networks

## Obama joins Foursquare

# Location-based social networks

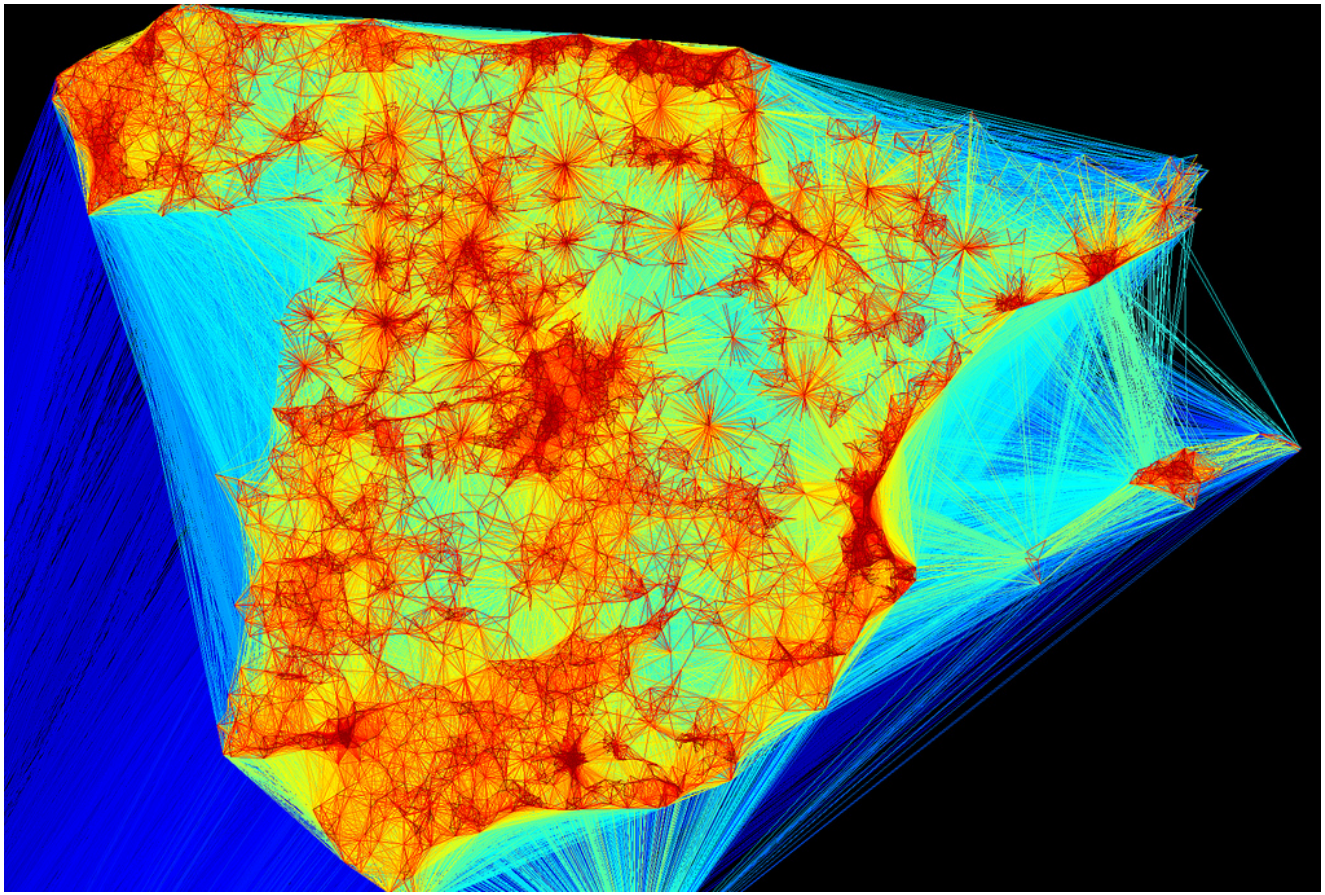
## Facebook connections

http://paulbutler.org/archives/visualizing-facebook-friends/

RuSSIR
Russian Summer School in Information Retrieval

# Location-based social networks

## Tuenti connections



http://beautyofsocialnetworks.blogspot.com/2011/02/visualizing-spains-friendship.html

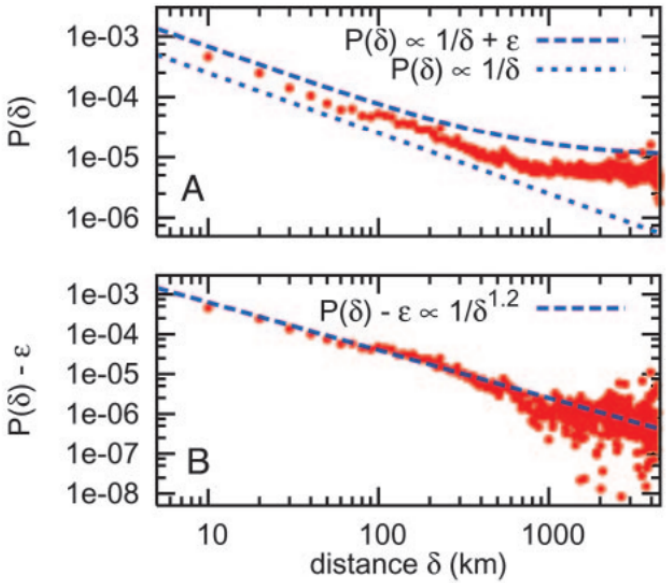
- Social ties and geographic distances
- **Popular assumption:** individuals try to minimize the efforts to maintain a friendship by interacting more with their spatial neighbors.
- online tools and long-distance travel might result in the ‘death of distance’.

- Flickr [Crandall et al., 2010];
  **in 60% cases:** users are friends if they have 5 co-occurrences within a day (in distinct cells with sides equal to 1 latitude-longitude degree).

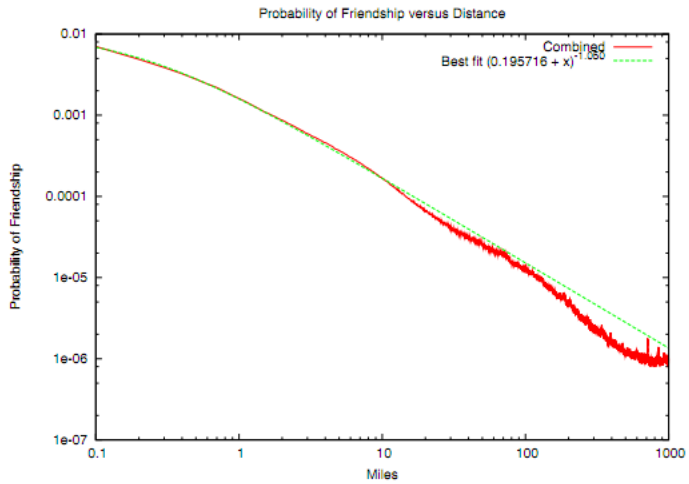
# Location-based social networks

## Social ties and geographic distances

Probability of a friendship between two individuals as a function of their geographic distance.




(a) livejournal [Liben-Nowell et al., 2005]




(b) facebook [Backstrom et al., 2010]

# Location-based social networks

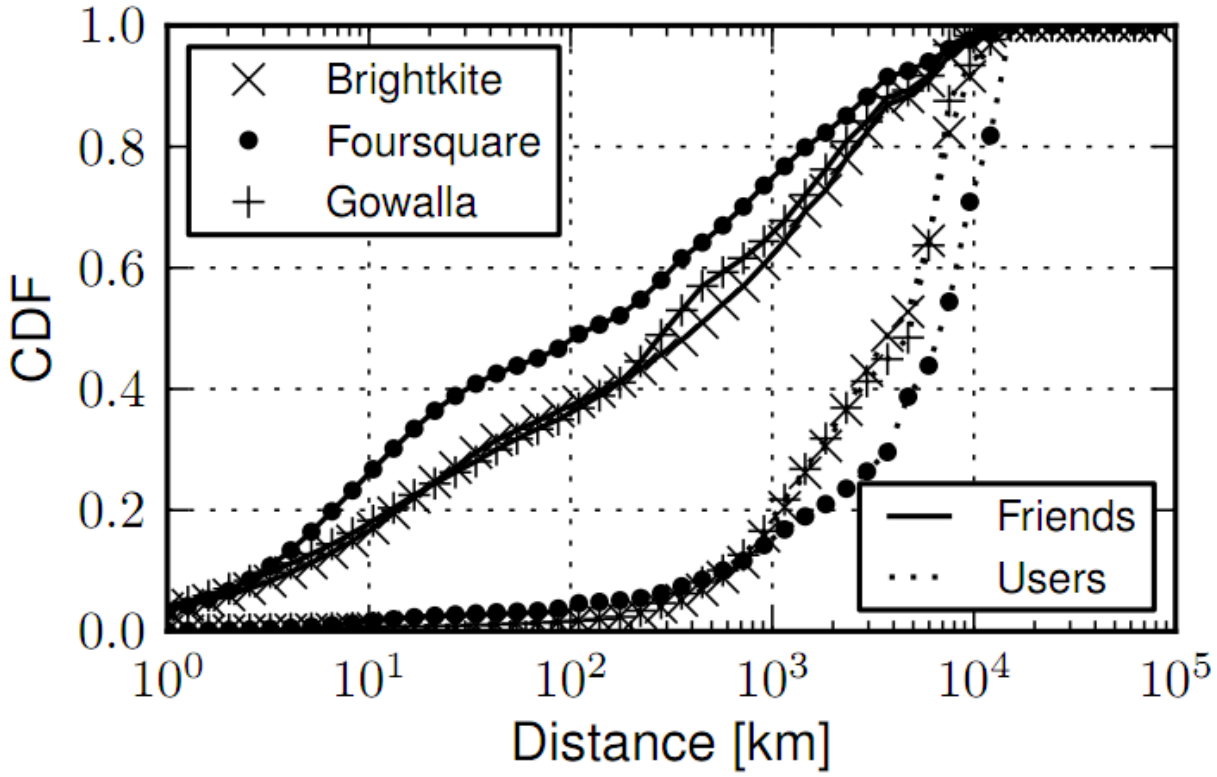
## foursquare & gowalla

| Service | brightkite | foursquare | Gowalla |
|---|---|---|---|
| Nodes | 54,190 | 258,706 | 122,414 |
| Social links | 213,668 | 2,854,957 | 580,446 |
| Average degree | 7.88 | 22.07 | 9.48 |
| Average clustering coefficient | 0.181 | 0.191 | 0.254 |
| Average distance between friends [km] | 2,041 | 1,442 | 1,792 |
| Average distance between users [km] | 5,651 | 8,494 | 5,663 |

from [Scellato et al., 2011]

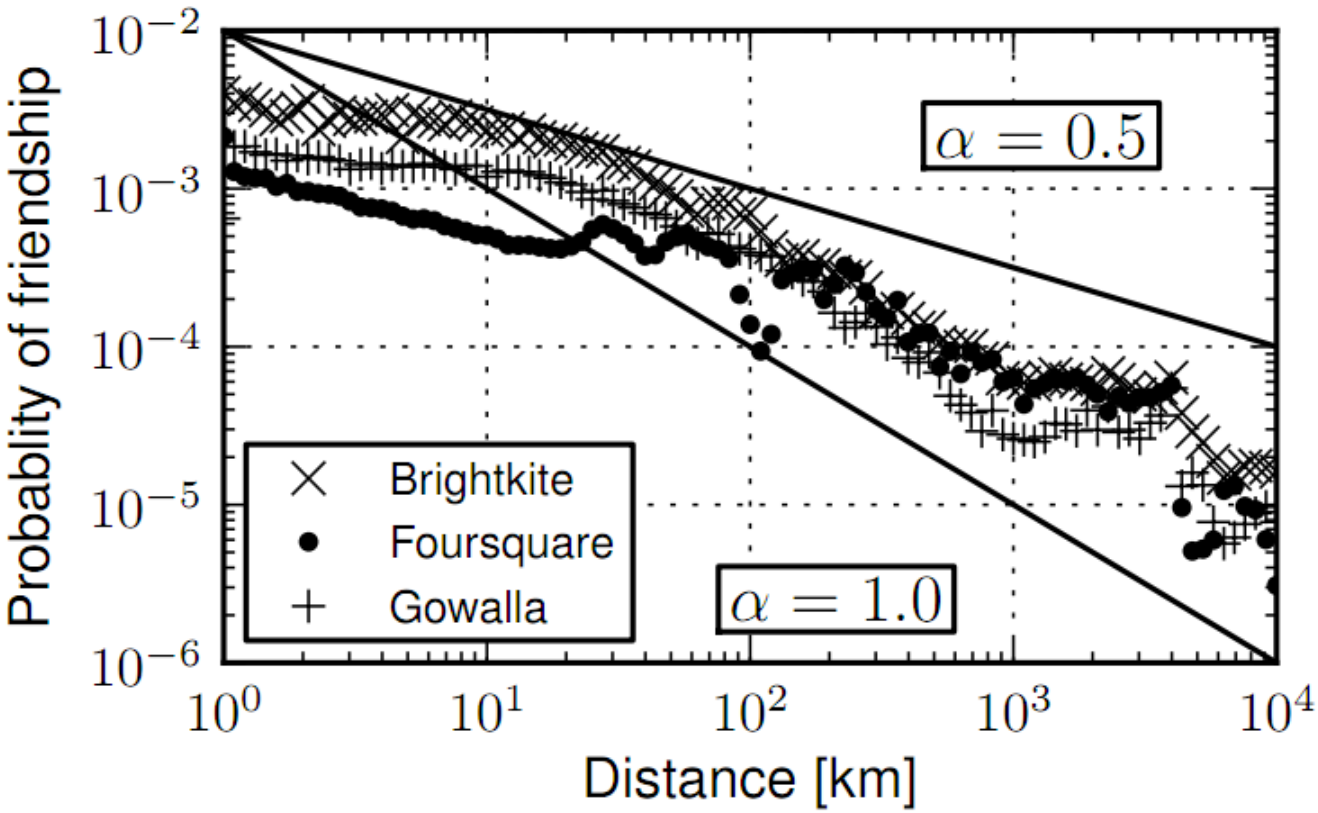
# Location-based social networks

## foursquare & gowalla




Friends tend to be much closer than random users: about 50% of social links span less than 100 km, while about 50% of users are more than 4 000 km apart.

## Probability of friendship vs. geographic distance



# Location-based social networks

## facebook

- user characteristics vs. location sharing and responds [Chang and Sun, 2011]

(a) male

(b) female

# Location-based social networks

## facebook

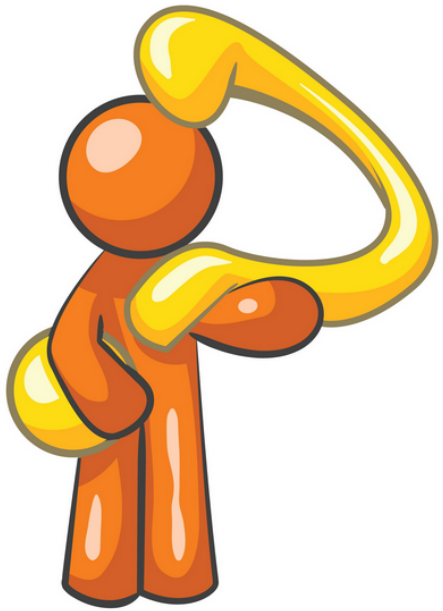
- **to predict next check-in**
  strongest: the number of previous check-ins by the user;
  significant: is the number of check-ins previously made by friends;
  small but significant: the day;
  not significant: the day of week.
- **to predict response**
  significant: the distance between the user (comments) and the actor (checks-in);
  the actor is near the user $\rightarrow$ the likelihood of a comment goes up dramatically.

| likes | comments |
|---|---|
| disneyland | hospital |
| fitness | medical |
| in-n-out | airport |
| disney | center |

Table 4: Top words predictive of likes and comments.

# Questions

## Questions



# Bibliography I


L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

J. Chang and E. Sun. Location 3: How users share and respond to location-based data on social networking sites. *In Proceedings of Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436, 2010.

D. Gayo-Avello and D. Brenes. Overcoming spammers in twitter-a tale of five algorithms, 2010.


RuSSIR
Russian Summer School in Information Retrieval

# Bibliography II

D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33): 11623, 2005.

S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. 2011.

J. Weng, E. P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010.

S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. *First Monday*, 15(1):1–13, 2010.

RuSSIR
Russian Summer School in Information Retrieval