How People Describe Themselves on Twitter

Evaggelia Pitoura,
University of Ioannina, Greece

Joint work with

Konstantinos Semertzidis
University of Ioannina, Greece

Panayiotis Tsaparas
University of Ioannina, Greece

Unleashing the Power of Social Networking for Enhancing Regional SMEs

What is Twitter?

An online social networking website and microblogging service that allows users to post and read text-based messages of up to 140 characters, known as "tweets".

- Evolved from a social network to a News media

As of September 2013, the company's data showed that **200 million users** send over **400 million tweets daily**, with nearly *60% of tweets* sent from *mobile devices* (The Guardian Sept 13, 2013)

What is Twitter bio?

Twitter provides a bio box, where users can give some information about themselves in fewer than 160 characters

Example: how twitter account describes twitter account

Example: how Times Square describes Times Square

Goal of this study

Analysis of user profile (or bios) on twitter to:

- ❖ See what twitter users expose on their bio
- ❖ Investigate whether possible to use bio information for tasks such as link prediction

✓ One of the less studied aspects

Recent work in terms of *forming expertise models*

- to assist expertise judgments by humans [Wagner et al 2012], or
- see if they to enhance the credibility of a tweet [Morris et al, 2012].

Outline

Present results

- Of our *analysis* of the content of the bios of twitter users
- Regarding the *similarity* between the information in the bio of connected users as well of the similarity of the profile of theirs friends - Investigate *whether "you friends say a lot about you"*
- of using bios for *link prediction*

Analysis of bio content

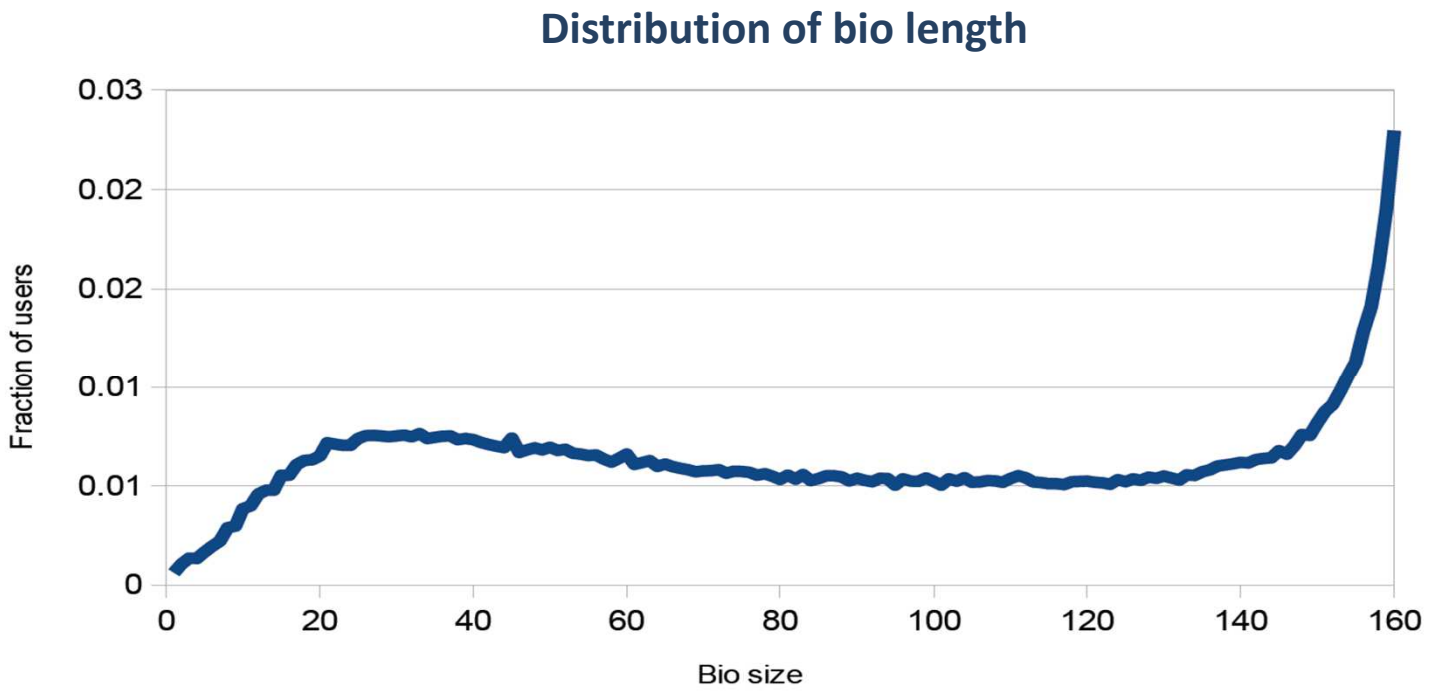
Data Collection

- ✓ Started with a sample of 10K users (from [Cha et al, 2010]), crawled their followings using the twitter api.
- ✓ This amounted to a set of 553,690 users.

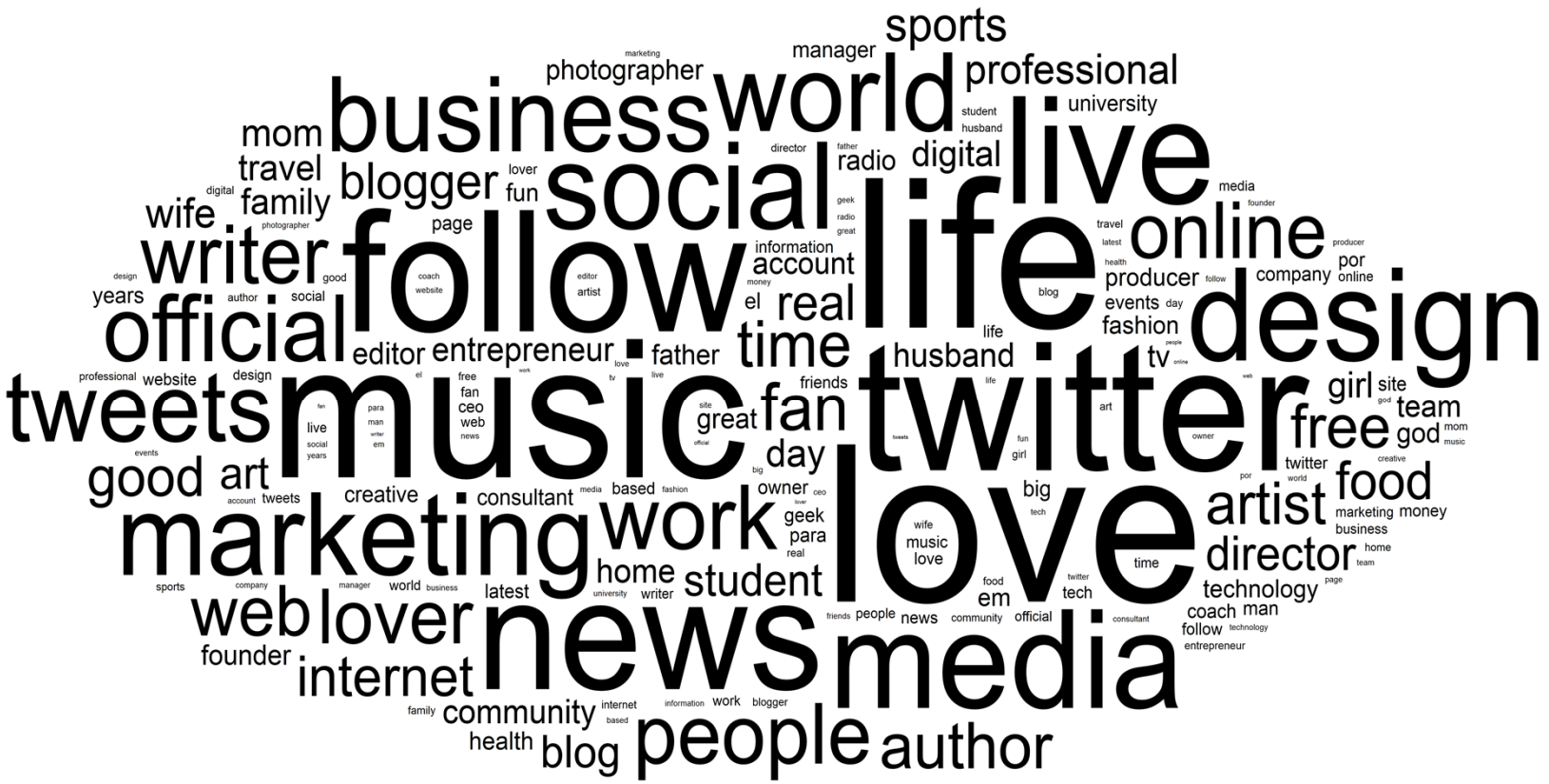
	Number of users	Percentage
Without bio	105,041	18.97%
Unreadable bio	3,027	0.55%
Readable bio	**445,622**	80,48%

Bio length

- ✓ Most users used all 160 characters - Average length 87



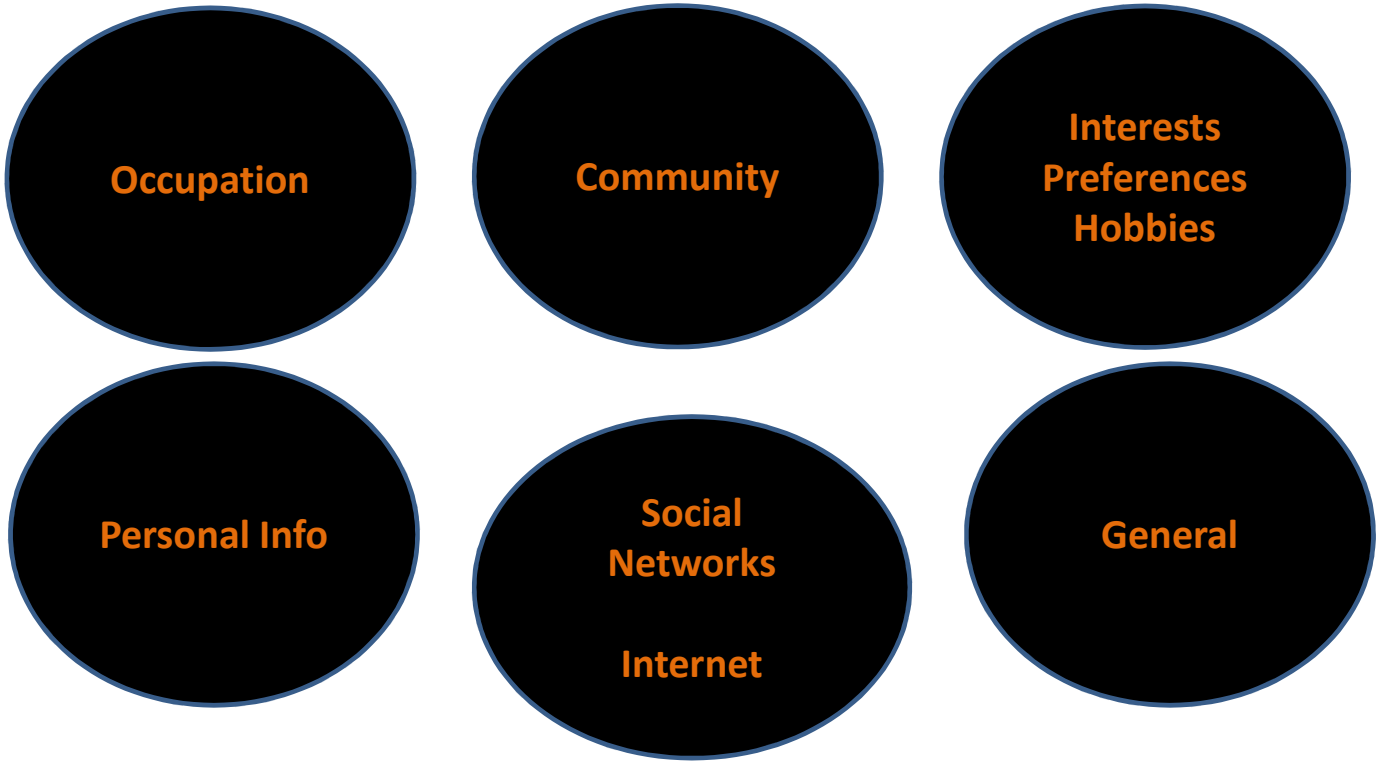
Frequent words



Some example bios

- I love all things sporting and sports related, like my cervelo bike!
- love good music, great movies, going on vacation, good food, reading inspiring books, spring time, earning
- linux journal editor. writer. geek.
- self employed silicon jockey. married a good wife, have a good dog, what more can u ask 4? chk my blog when you get time for more info.
- proud new england native. always interested in solid story ideas. football obsessed.

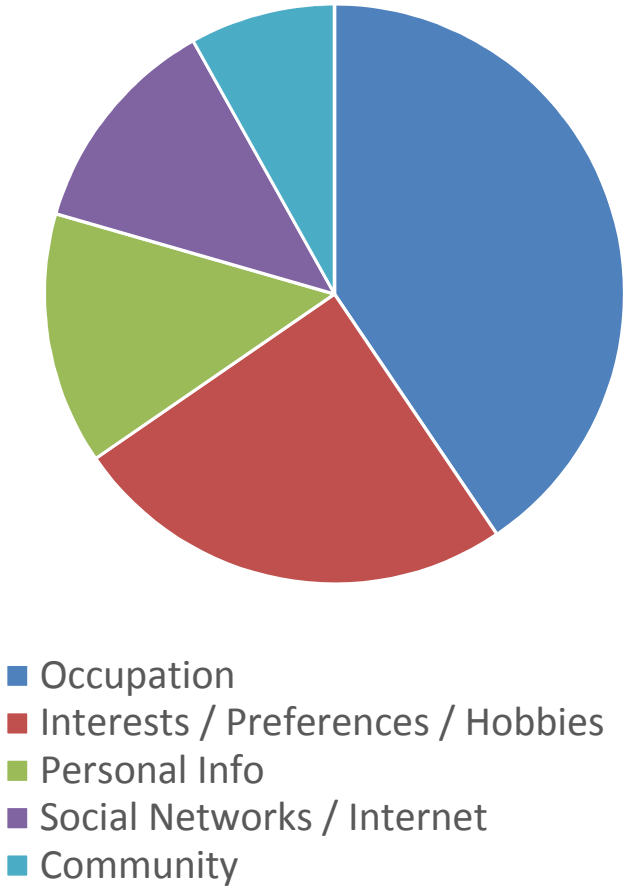
Categories of the top-200 words



Categories - examples

Occupation	marketing, business, writer, photographer, author, artist, designer	love, music, writer, lover, fan, games, blog, football	**Interests Preferences Hobbies**
Personal Info	wife, mom, family, girl, husband, father, god, christian	twitter, follow, official, people, tweets, account, facebook, network	**Social Networks Internet**
Community	news, media, community, events, join, club, up-dates, member	show, latest, making, public, international, high, call, estate, share	**General**

Categories – popularity



Users more likely talk about their *occupation*, then their *interests*, and then *personal information*.

- ✓ Not surprising since, many *users use Twitter for professional purposes* (e.g., journalists, politicians, athletes).
- ✓ Also, common for people to *use their interests to define themselves online*.
- ✓ *When people use Twitter more as a social network* (to keep in touch with their friends), likely to include personal information, such as their marital status, their religion, or the community they belong to.

Bio similarity between friends

Bio similarity

Are connected users more likely to have similar bios than random users?

Are people "more similar" with their friends than with others?

Two types of connections:

- The simple follow relationship, where user *u* follows user *v*
- The mutual follow relationship, where user *u* follows user v and user *v* follows *u* back (more indicative of a true social connection between two users [Cheng et al, 2011])

Bio similarity

Crawled sample of 300K users, following and followers

Keep users:

- ✓ Have non-empty bio
- ✓ No more than 5,000 followers
- ✓ No more than 5,000 followings

595,573 simple follow relationships
245,828 mutual follow relationships (besides simple follow)

Bio similarity

For each user u, *bio vector B(u)*

- length equal to the total number of words in all profiles.
- 0/I vector, 1 if the corresponding word appears in the profile and 0, otherwise.

Bio similarity between users u and v:

$$S_{b(u,u)} = \frac{B(u) \cdot B(v)}{\|B(u)\| \|B(v)\|}$$

Average Bio similarity

	Average Bio Similarity
Non-connected Pairs	0.00514
Simple Follow Pairs	0.01825
Mutual Follow Pairs	0.02278

For non-connected pairs: average similarity over 595,578 randomly sampled non-connected pairs.

- Connected users have more similar bios compared to non-connected users.
- Similarity increases as the strength of the relationship between pair of users increases (mutual followers)

✓ a *t-test of significance* that showed that differences in similarity are ***statistically significant***

Consistent with homophily observed in social networks: users are likely to follow other users who describe themselves in a similar way

Followings Bio

Followings bio (f-bio) fB of user u: the bio vectors of all the users who u follows

$$fB(u) = \sum_{u \in N(u)} B(u)$$

- Contrary to B(u), fB(u) may count a *word multiple times*.
  Why? multiple occurrences of a keyword in the f-bio means that the user follows many users that are described by this word, and thus stronger affiliation with this word

- Again, measure similarity between two f-bios using the cosine similarity.

Average f-Bio similarity

	Average f-Bio Similarity	Average Bio Similarity
Non-connected Pairs	0.44871	0.00574
Simple Follow Pairs	0.78697	0.01898
Mutual Follow Pairs	0.83202	0.01912

A subset of 1,126 users, with at least 10 reciprocal connections
Statistically significant

Link prediction using bio

Link prediction

We consider the following link prediction tasks

- Follower prediction: given a pair of users u and v, predict if there is a follower relationship between u and v (in either, or both directions).
- Follow-back prediction: given a pair of users where u follows v, predict if user v will follow user u back.
- Mutual-follow prediction: given a pair of users u and v, predict if user u will follow v and user v will follow u

Link prediction

✓ *APPROACH: For each prediction task, we predict that the users will connect, if their similarity is above a threshold.*

Three different similarity measures and corresponding prediction techniques:

1. Bio similarity
2. Following bio similarity
3. Neighborhood similarity: measure of common followings

$$S_n(u,v) = \frac{|\,\mathrm{N(u)} \cap N(v)\,|}{|\,N(u) \cup N(v)\,|}$$

Link prediction: evaluation methodology

For a given similarity measure *S*, and a threshold value θ,

- Let $\boldsymbol{Ts(\theta)}$ be the set of pairs that have similarity at least θ (predicted to be connected)

and C the set of the connected pairs

Precision $P_s(\theta) = \frac{|\,\mathrm{T}_s(\theta) \cap C\,|}{|\,\mathrm{T}_s(\theta)\,|}$

Recall $\mathrm{R}_s(\theta) = \frac{|\,\mathrm{T}_s(\theta) \cap C\,|}{|\,C|}$

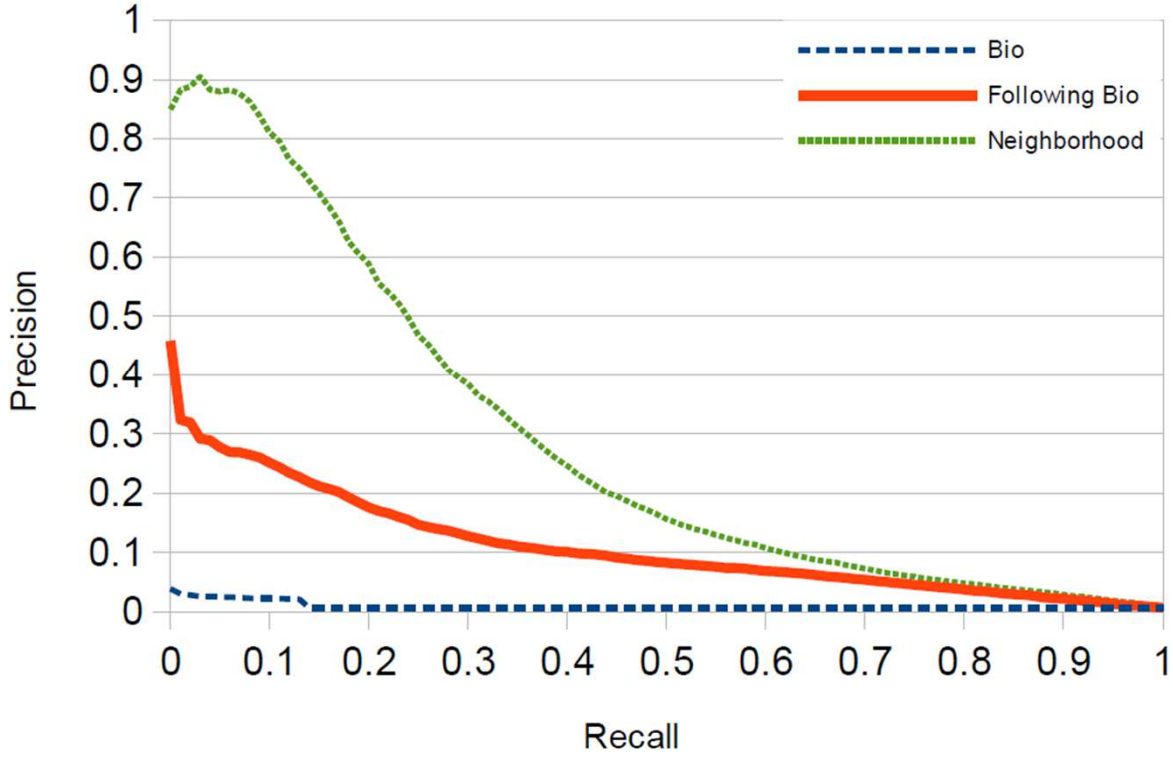
Link prediction: evaluation methodology

We use the following methodology to evaluate the different approaches.

For each technique,

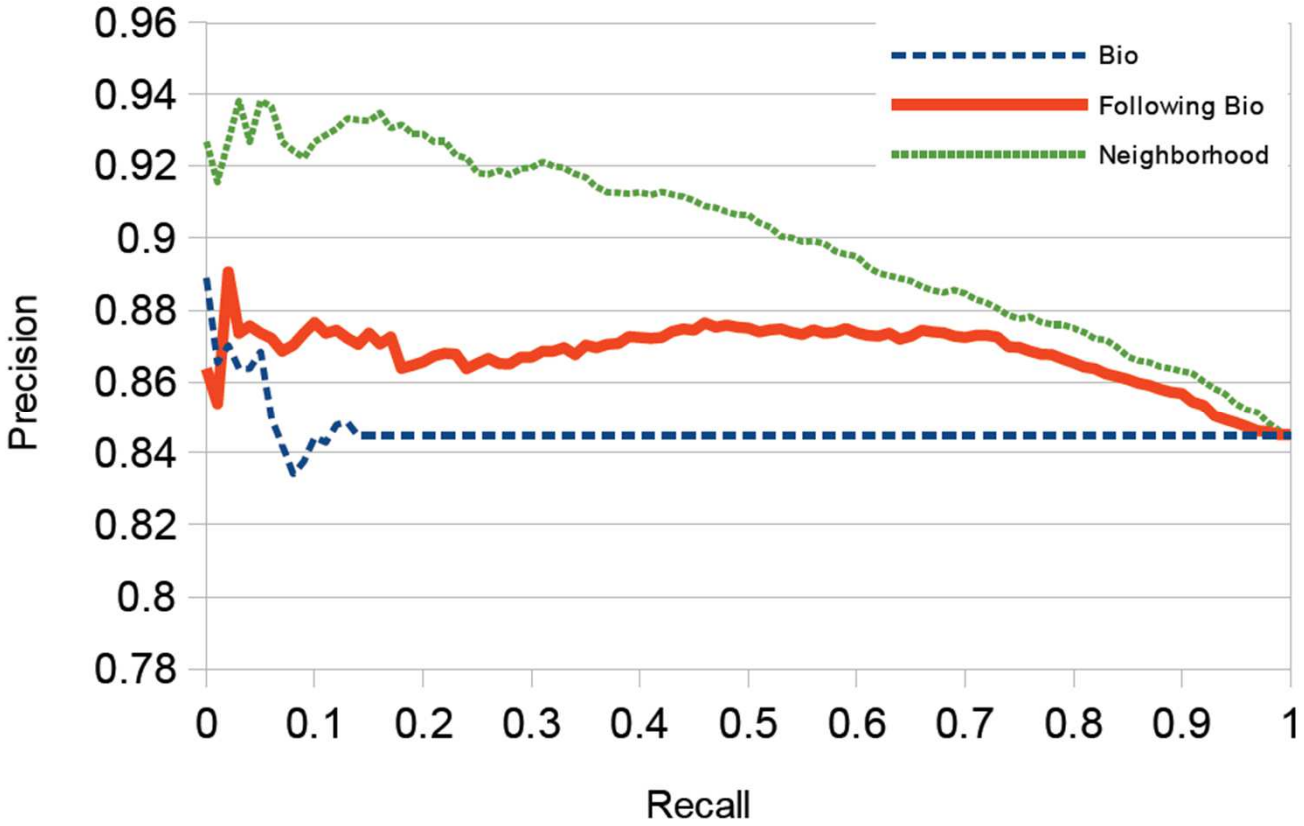
1. we consider *different values of recall* and compute the *threshold* that gives this recall.
2. Using this threshold, we compute the *precision* of the technique and
3. Plot the corresponding *precision-recall curve*

Follower prediction



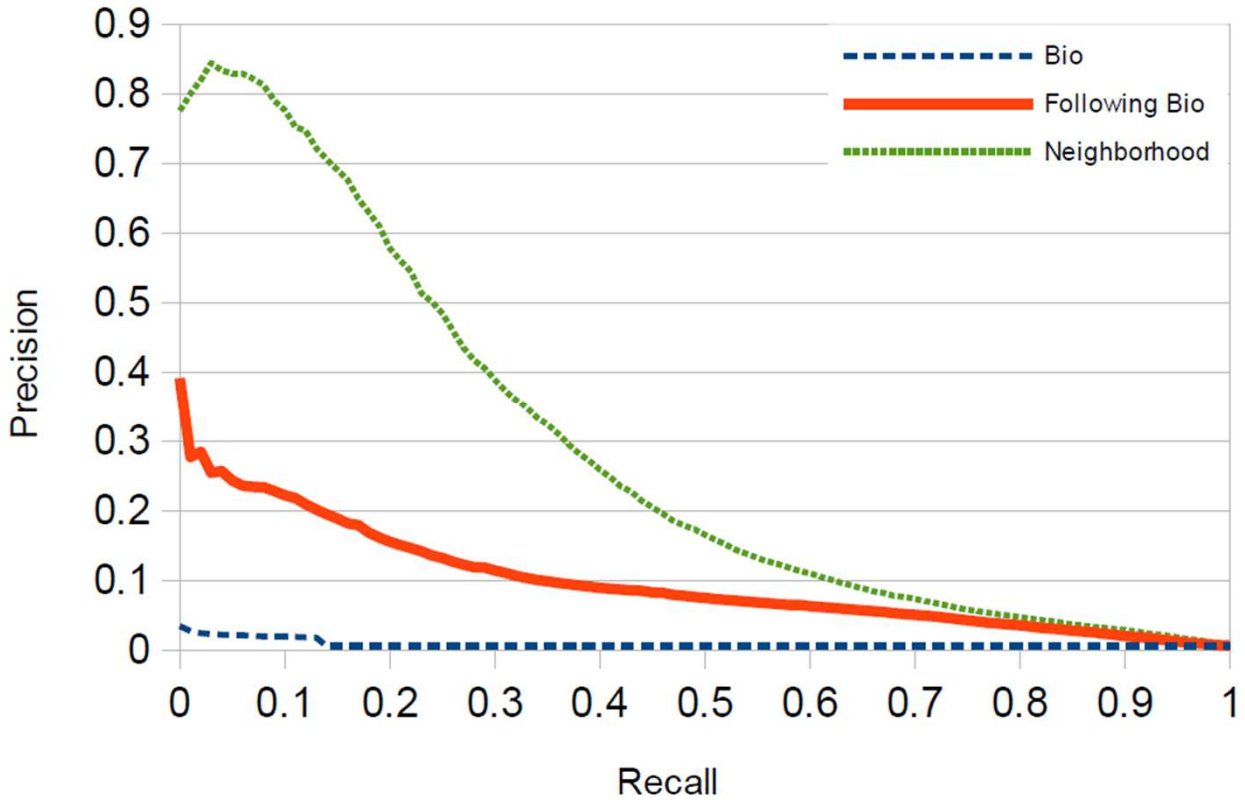
✓ People with common friends are likely to become friends

Follow-back prediction



- 85% of our links are reciprocal, so this task is easier.
- Still we get a non-negligible increase in the prediction accuracy for higher thresholds, using either the f-bio, or the neighborhood similarity

Mutual Follow prediction



Slightly more difficult than follow prediction

Discussion

- Neighborhood similarity performs the best (to some extend expected, e.g., triadic closure)
- Between the two bio-based methods, f-bio similarity better (confirming that the bios of one's followings give a better description of the user than her own bio)

- *Some useful signal in the bio description that can be used towards predicting connections (possibly in conjunction with the neighborhood information).*

Conclusions

- *Studied the user bio descriptions on Twitter*. *To the best of our knowledge, first detailed analysis for bios.*
- Shown that connected users tend to share common keywords, establishing homophily with respect to descriptions.
- Homophily more pronounced when a description for a user is constructed from the bios of the users that she follows (the f-bio).
- Tested bio similarity to predict connections between Twitter users.

Future Work

How the signal from the bio can be combined with other signals for different mining tasks.

- Combine all the similarity measures into a single link prediction algorithm.
- Predict when a tweet will be retweeted, or
- whether a user will express interest in a new hashtag.

Thank you!

Questions?

K. Semertzidis, E. Pitoura, P. Tsaparas, How People Describe Themselves on Twitter. ACM SIGMOD Workshop on Data Bases and Social Networks (DBSocial), 2013. **(Best paper award)**