That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships

David Jurgens✢
Sapienza University of Rome

✢Work done while at HRL Laboratories, LLC

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI / NBC) contract number D12PC00285. The IARPA research focuses solely on Latin America. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBE, or the U.S. Government.

Location matters

Regional collapse
or local occurrence?

Location matters

Regional collapse
or local occurrence?

Budding epidemic
or just a case of the flu?

Location matters

Kathy Yakaitis @kathyyak 12m
@flyboy_1776 @customjewel I **lost my job** in Oct. & still haven't found another. It's bad here. He really made a mess of things...
View conversation Reply Retweet Favorite More

Regional collapse
or local occurrence?

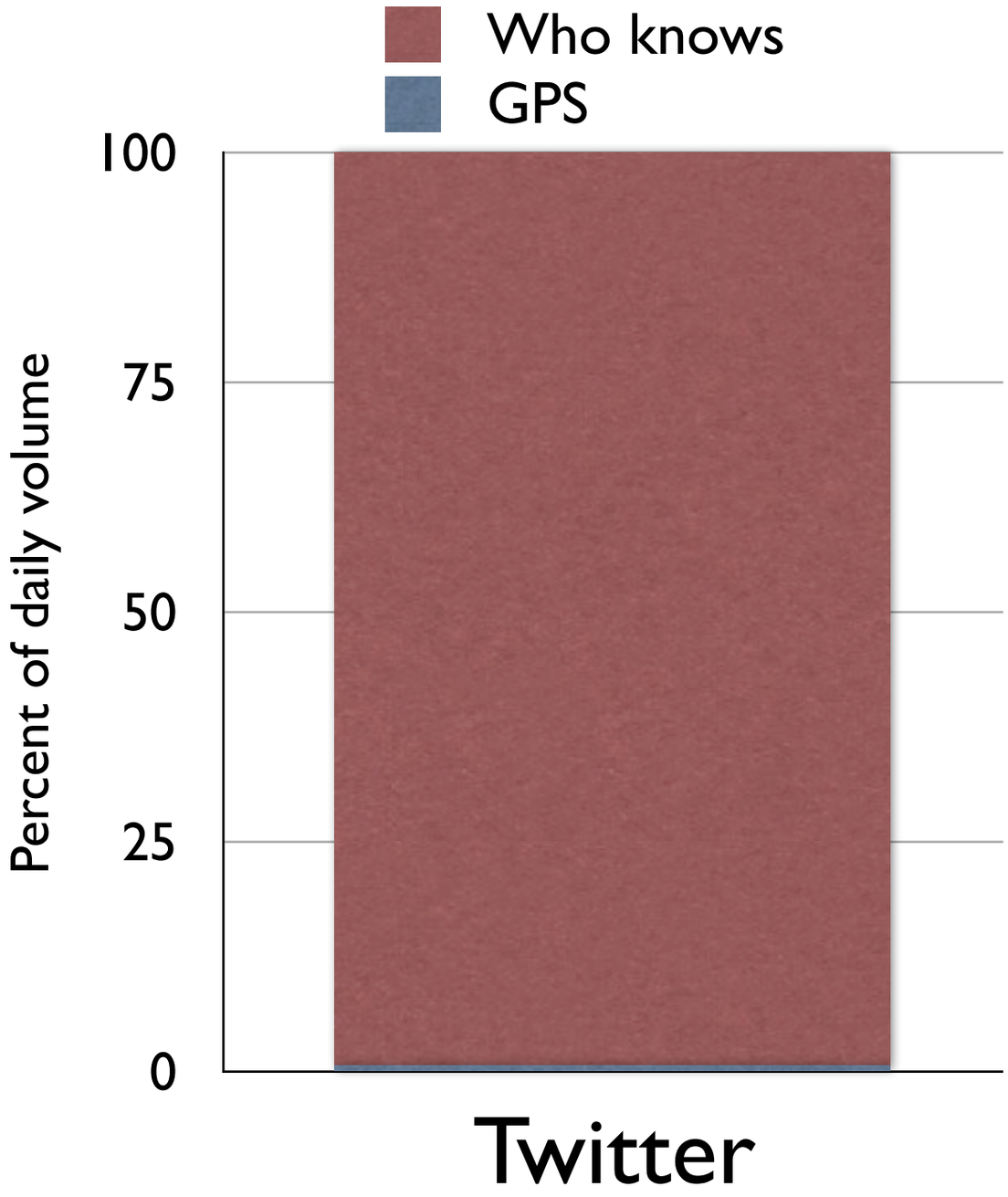
Mark Currie @MarkCurrieNZ 1h
So now not only are my bones and muscles still sore from fitness test, now I've **got the flu** and some sort of stomach bug. Ridiculous.
Expand Reply Retweet Favorite More

Budding epidemic
or just a case of the flu?

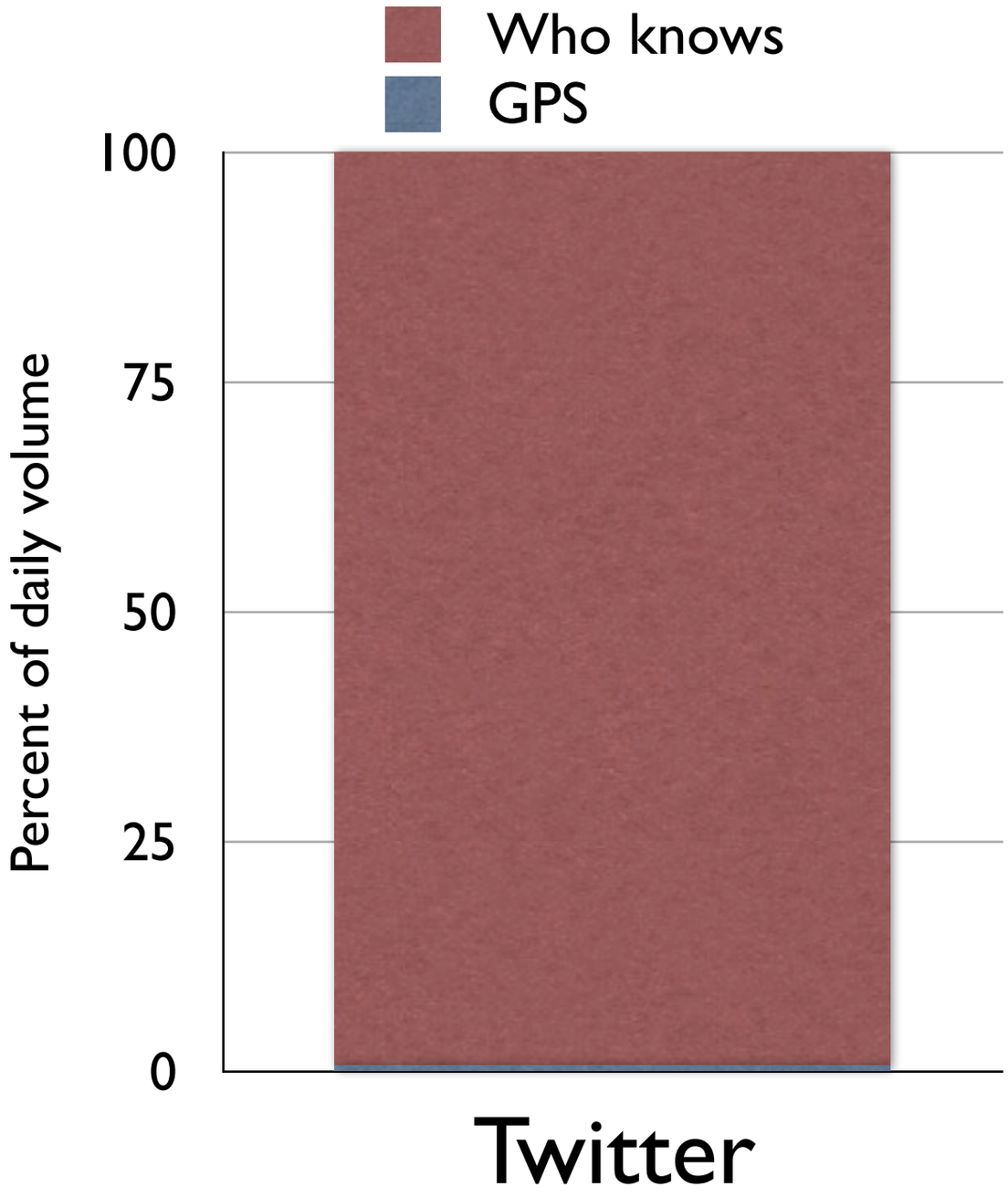
Nathzão @_iamtwisted 1m
Gente, deu uma saudade do dia do **protesto** do PeBa, as tretas por causa das mudanças, a rinite atacada com a poeira.. ah, saudades!
Expand Reply Retweet Favorite More

The start of a mass riot
or just an unhappy person?

But good location data is relatively sparse



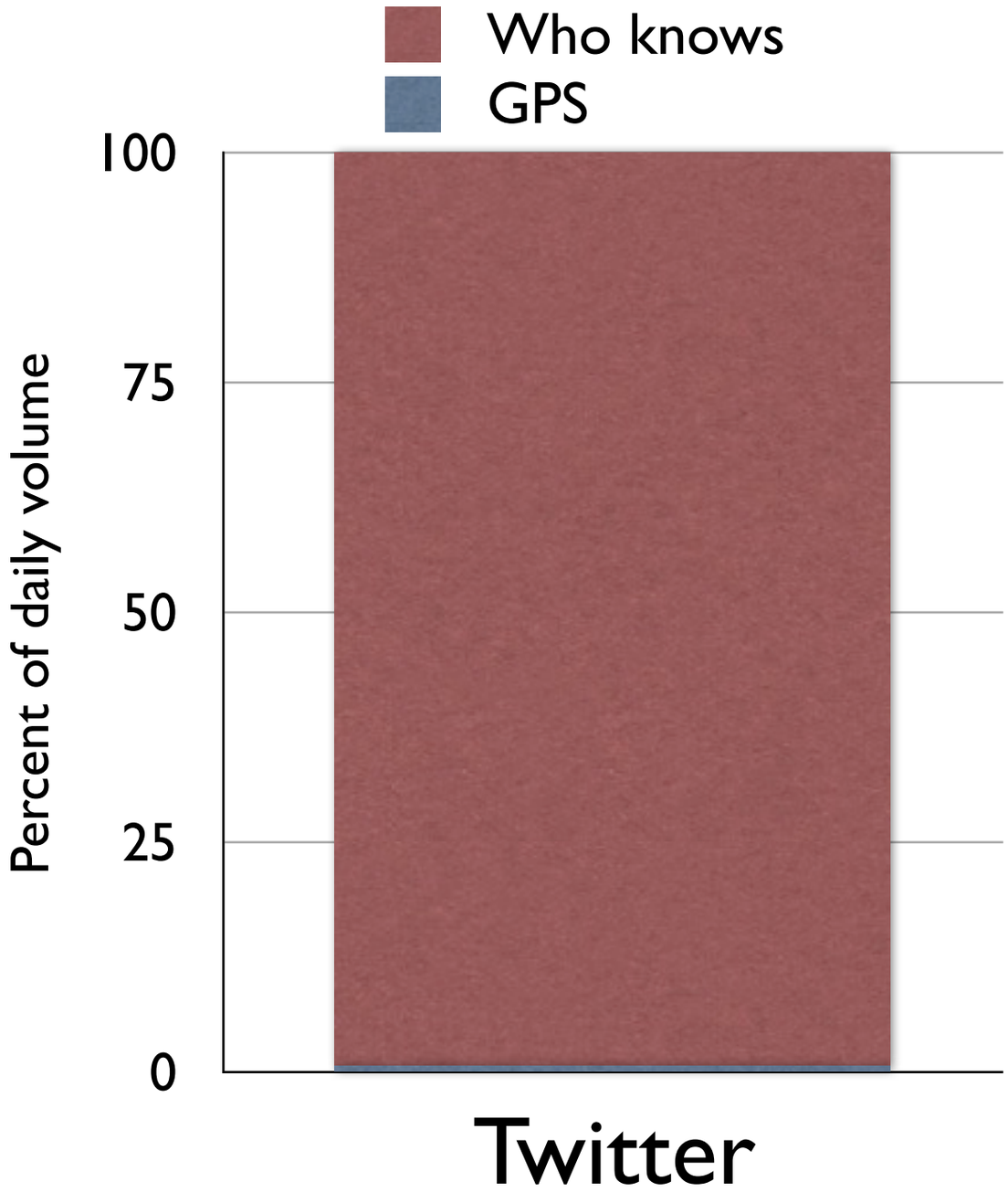
But good location data is relatively sparse



User-provided locations

Hecht et al. (2011), Pontes et al. (2012)

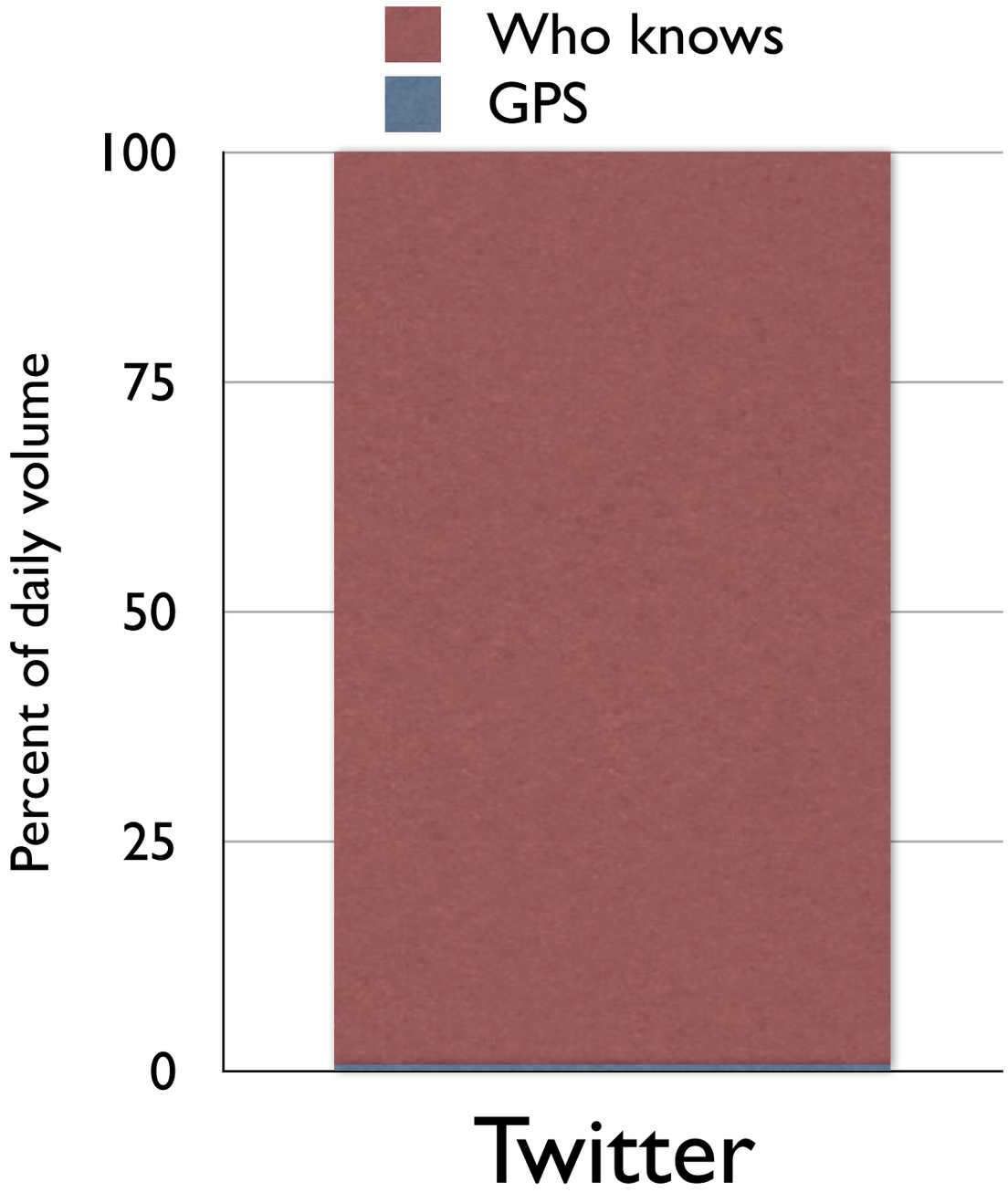
But good location data is relatively sparse



User-provided locations
Hecht et al. (2011), Pontes et al. (2012)

Message Content
Cheng et al. (2010), Mahmud et al. (2012), Ikawa et al. (2012), Bo et al. (2013)

But good location data is relatively sparse

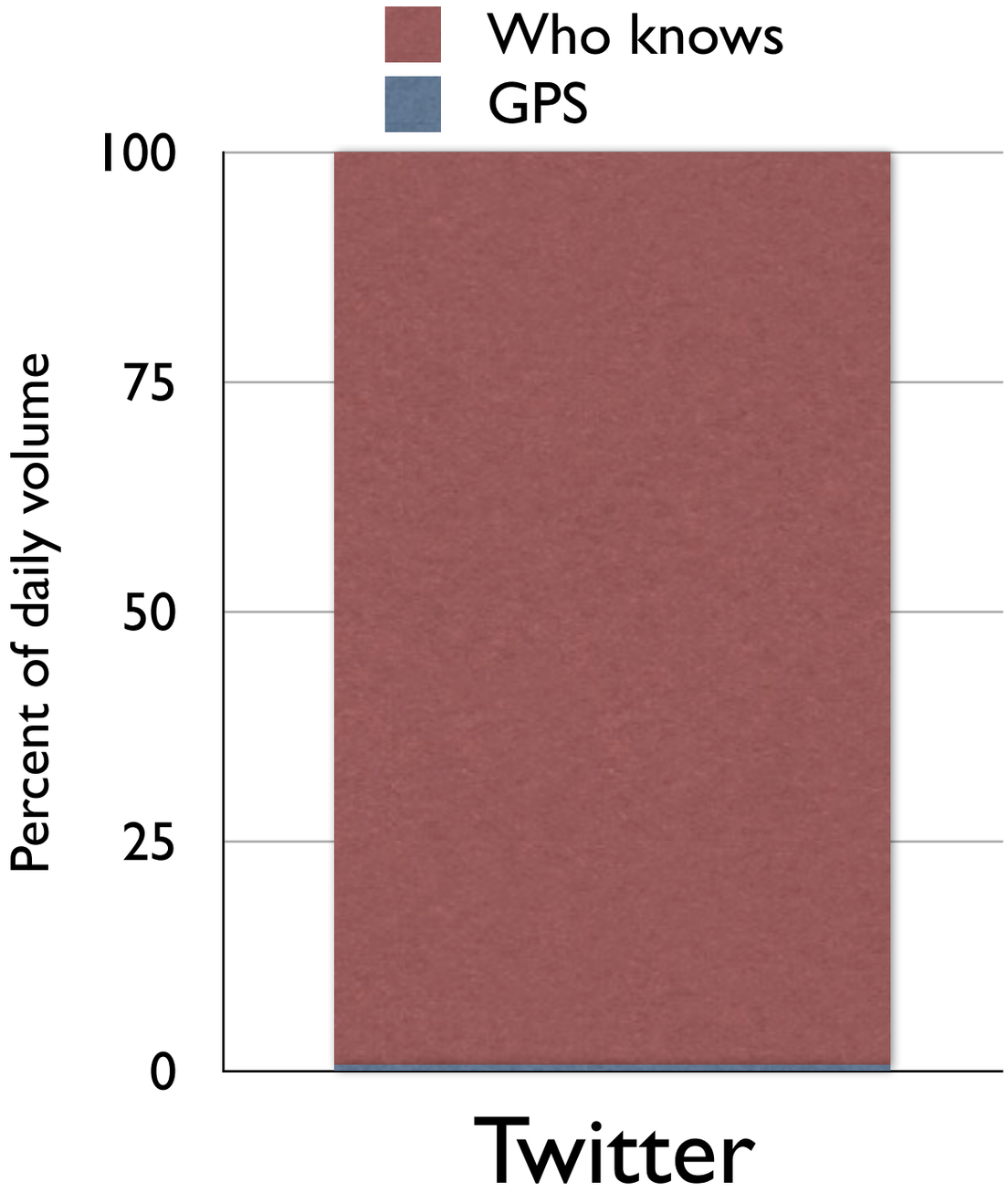


User-provided locations
Hecht et al. (2011), Pontes et al. (2012)

Message Content
Cheng et al. (2010), Mahmud et al. (2012), Ikawa et al. (2012), Bo et al. (2013)

Social Network
Backstrom et al. (2010), Davis Jr. et al. (2011), Salidek et al. (2012)

But good location data is relatively sparse



User-provided locations
Hecht et al. (2011), Pontes et al. (2012)

Message Content
Cheng et al. (2010), Mahmud et al. (2012), Ikawa et al. (2012), Bo et al. (2013)

Social Network
Backstrom et al. (2010), Davis Jr. et al. (2011), Salidek et al. (2012)

Sociological Contribution

Locality is still a dominant factor in the social relationships people have online

Pragmatic Contribution

Geo-tag 77% of all Twitter data

Pragmatic Contribution

Geo-tag 77% of all Twitter data independent of country

Pragmatic Contribution

Geo-tag 77% of all Twitter data

independent of country

independent of language

Pragmatic Contribution

Geo-tag 77% of all Twitter data

independent of country

independent of language

(mostly) independent of ego-network size

Pragmatic Contribution

Geo-tag 77% of all Twitter data

independent of country

independent of language

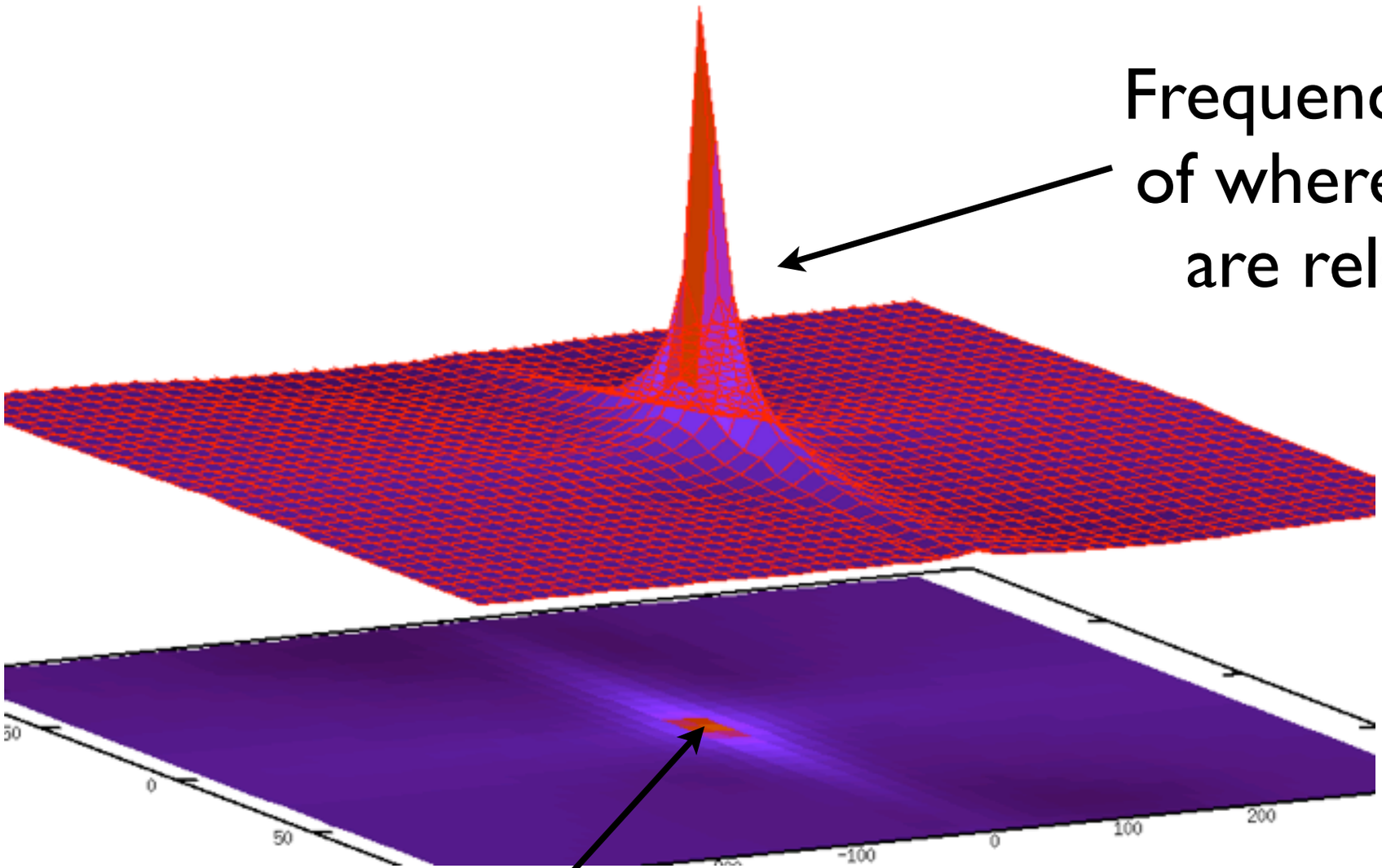
(mostly) independent of ego-network size

Median error ~ 10km

In olden times, your social network was only people nearby

Does location matter if we can be friends with anyone, anywhere?

Location is still alive in online social networks



Based on 20.5M relationships in Twitter

Online Social Networks under focus

- Twitter
  - Bi-Directional @mentions
  - Bi-Directional followers
- Foursquare
  - Explicit friendships

All have location data

Twitter Social Networks

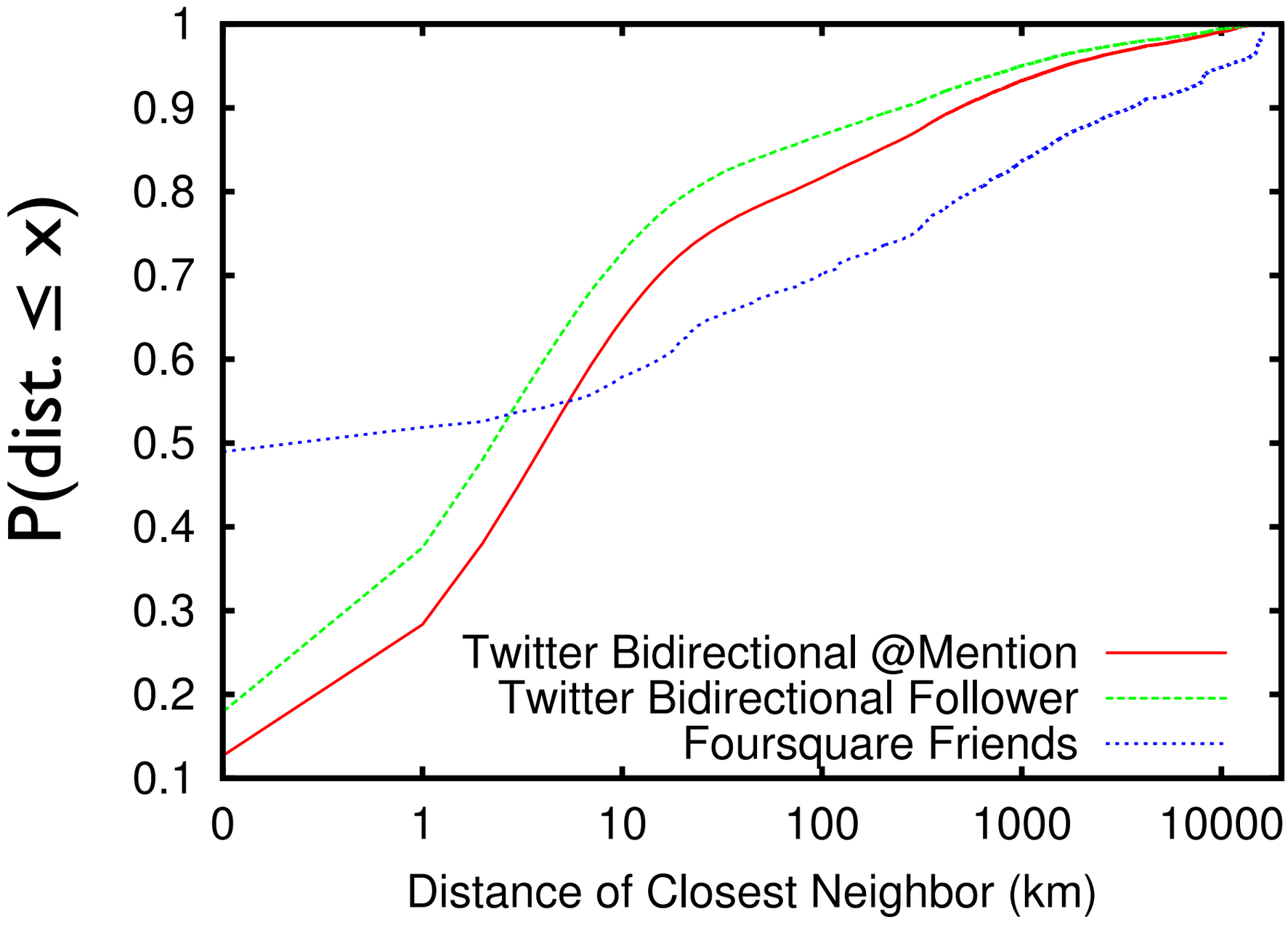
- Bi-directional followers (crawled)
  - ~96K individuals and 16.6M relationships
- Bi-directional mentions
  - from a 10% sample of Twitter over 7 months
  - 47.7M individuals and 254M relationships
  - 5.3% tagged with user-level location

Foursquare overview

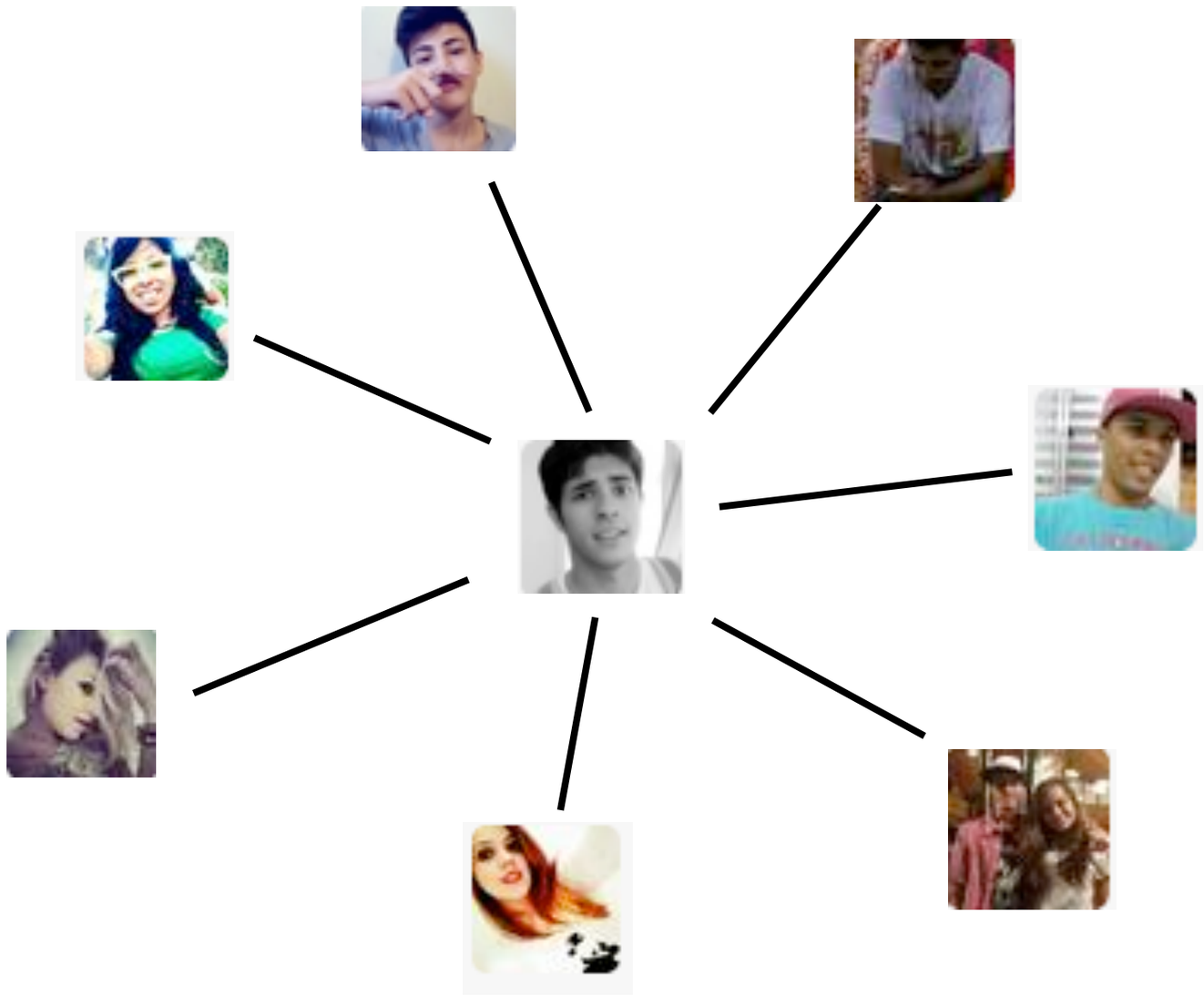
- Built from a crawl over 3 months
- ~4M individuals and17.6M relationships
  - 1.6M also had linked Twitter accounts
    - 52.8% of Foursquare relationships for Twitter-linked accounts also had bi-directional mentions in Twitter
- Self-reported location was highly accurate, so we mapped 68.8% of users to a location

How close is the closest friend?

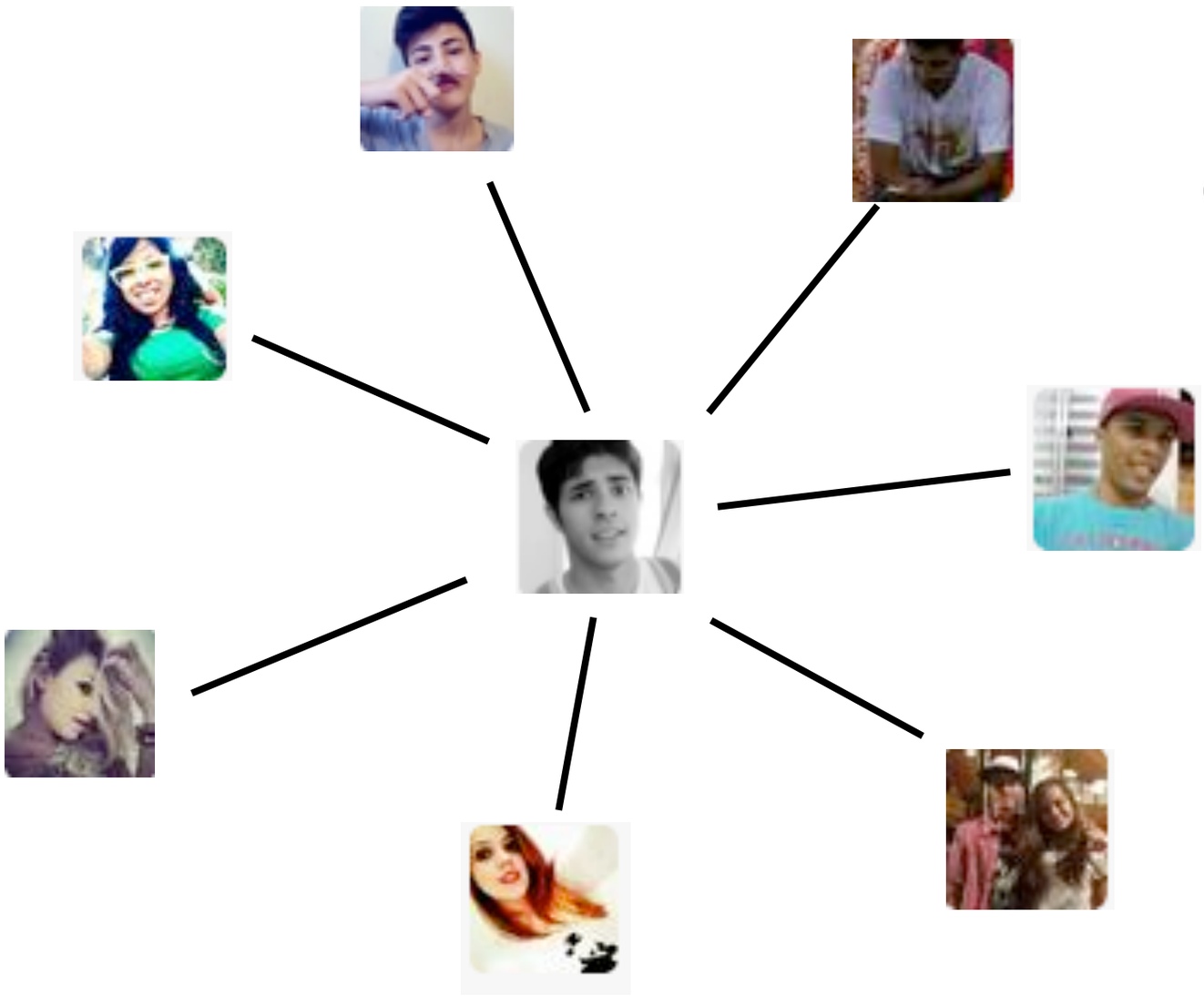
How close is the closest friend?



High-level Algorithm:
Your location is a function of your friends' locations



High-level Algorithm:
Your location is a function of your friends' locations

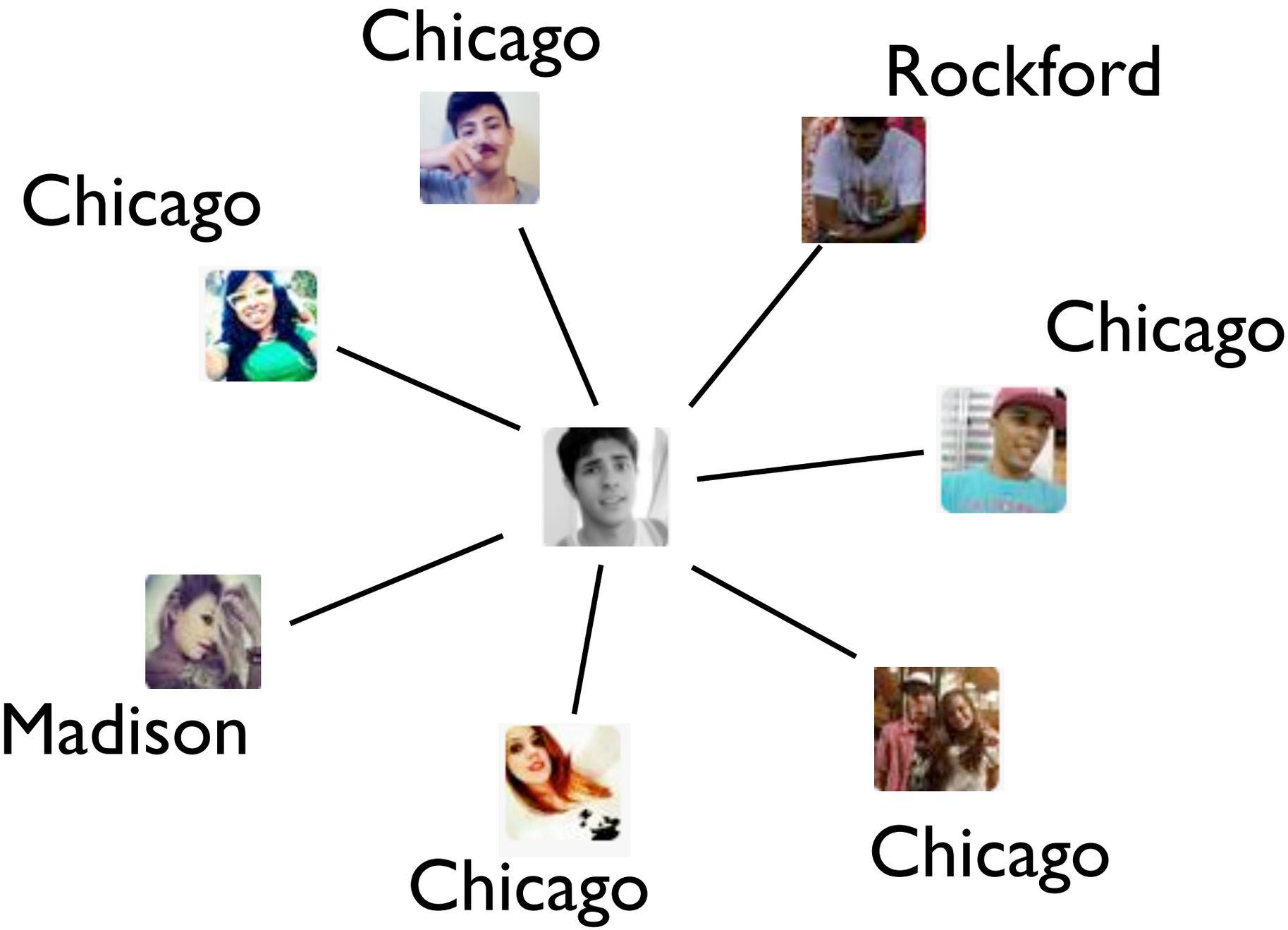


do this for a while:

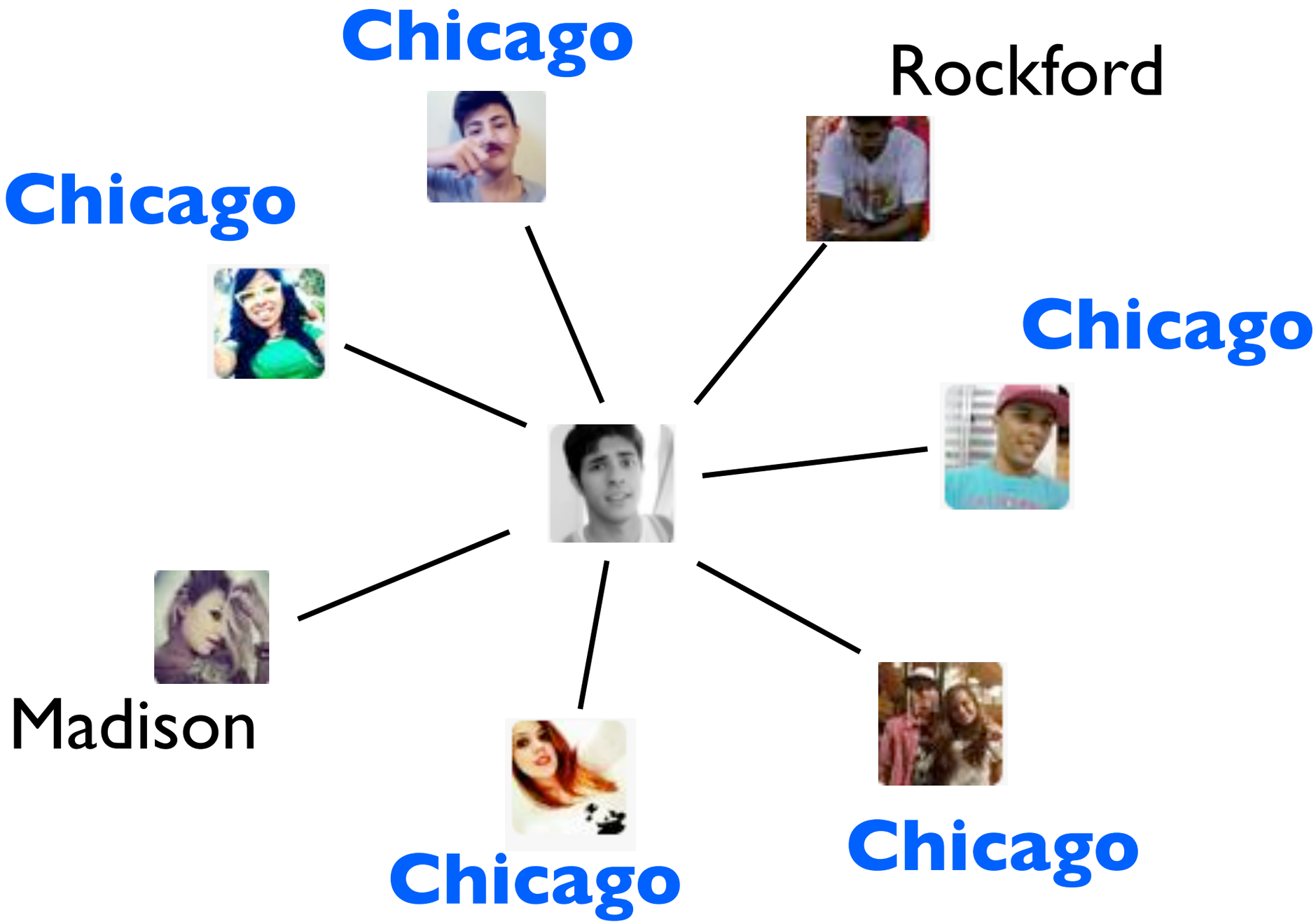
for everyone in the network:

1. Get their friends' locations
2. Pick one of them (smartly) as the user's location

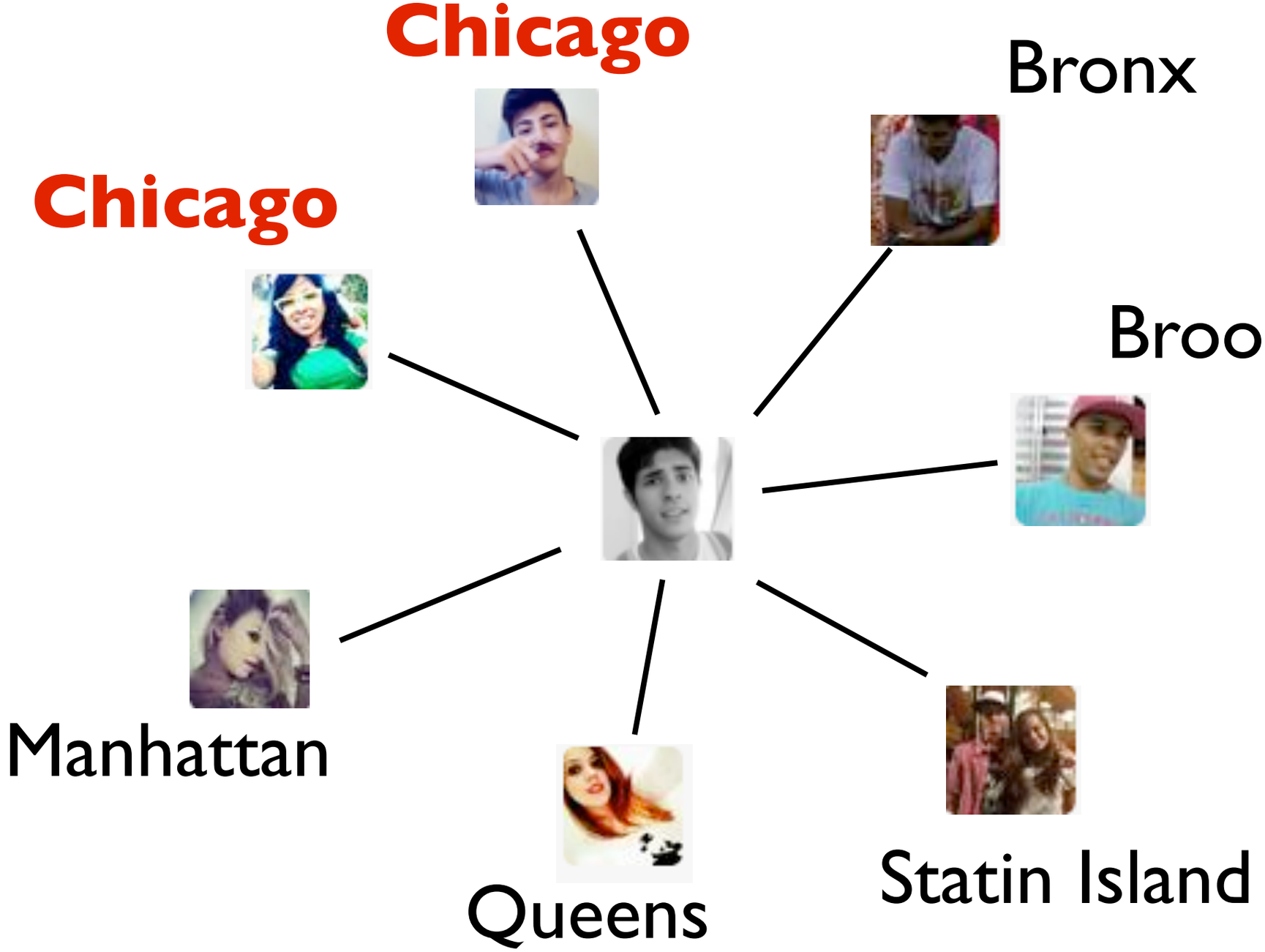
Label Propagation



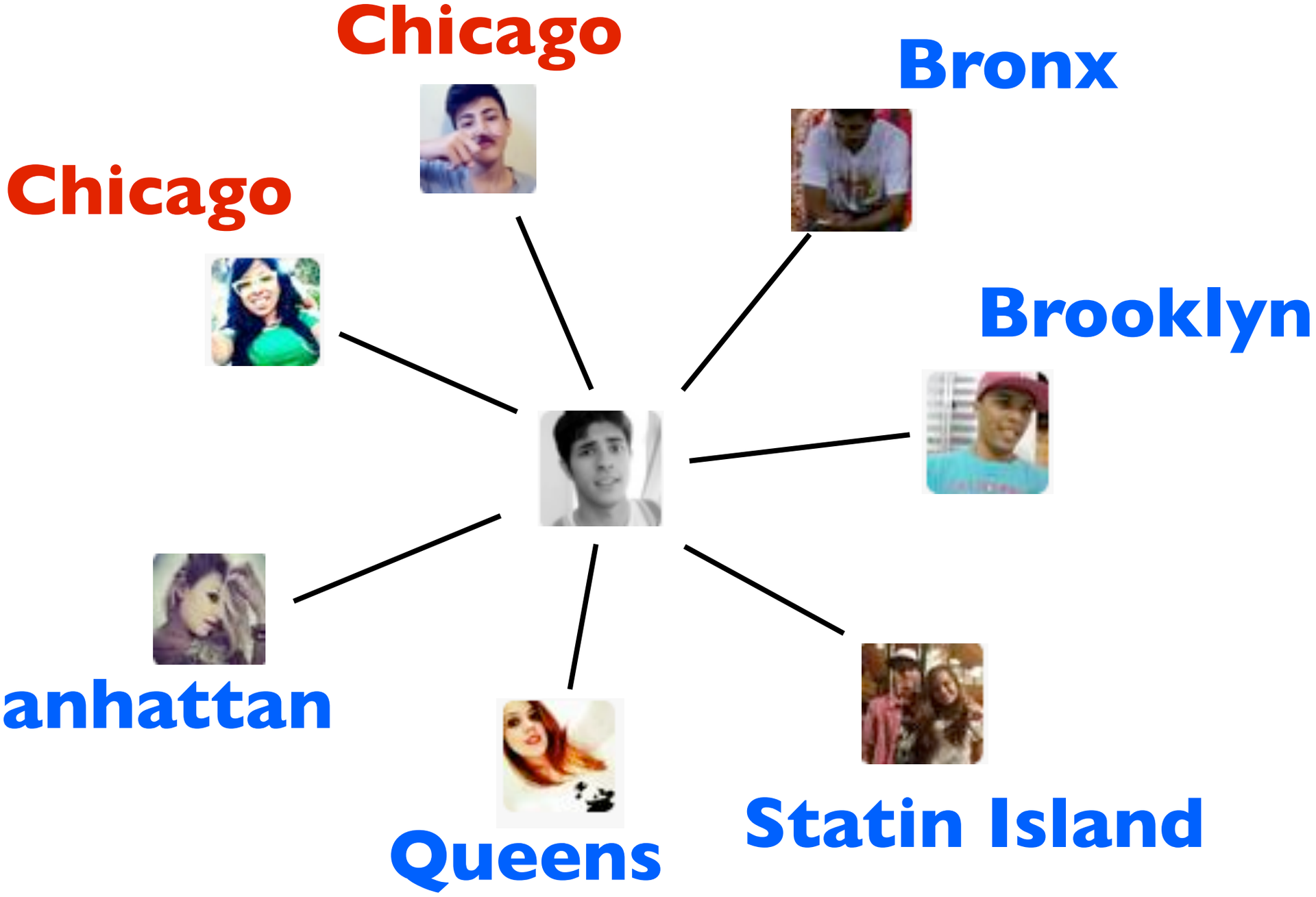
Label Propagation



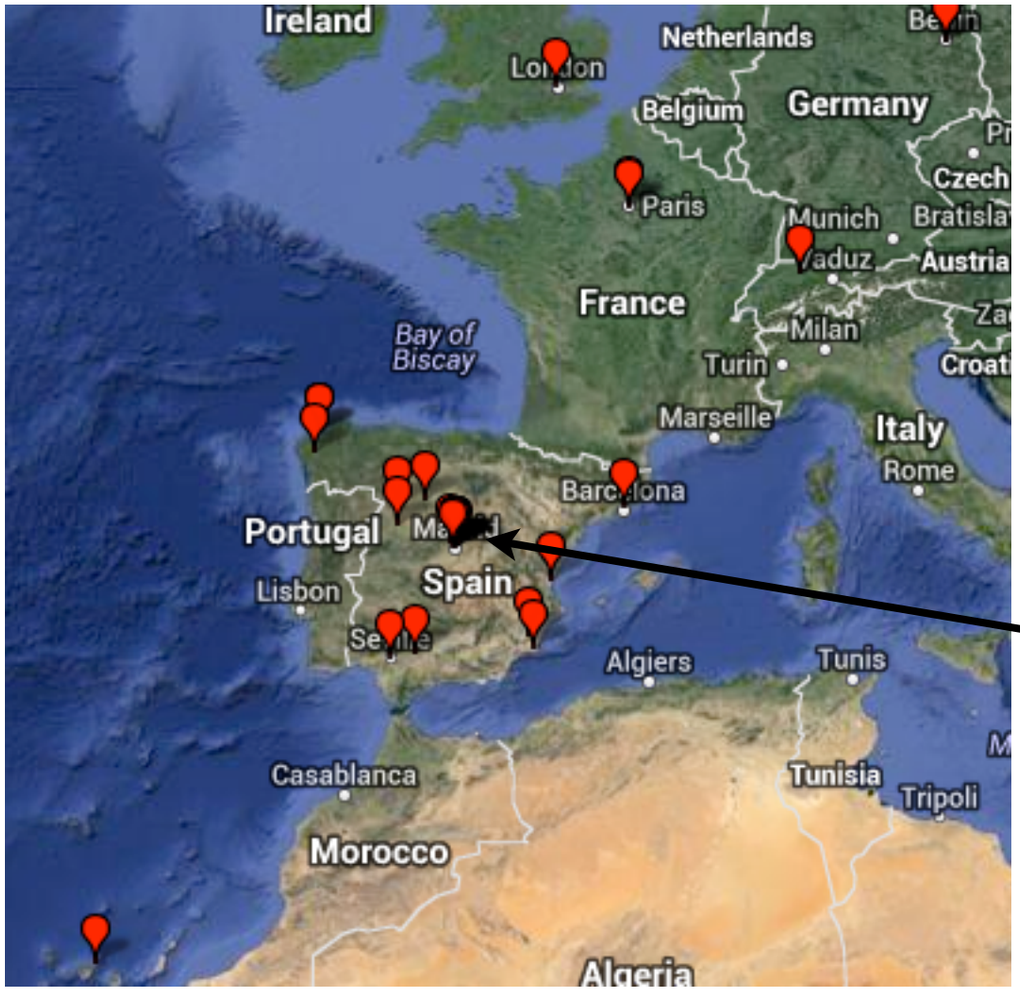
The slight problem with Label Propagation



The slight problem with Label Propagation



Spatial Label Propagation



Location data is actually **latitude** and **longitude**

Pick the **geometric median** of the friends' locations

Comparisons

do this for a while:

for everyone in the network:

1. Get their friends' locations
2. **Pick one of them (~~smartly~~) as the user's location**

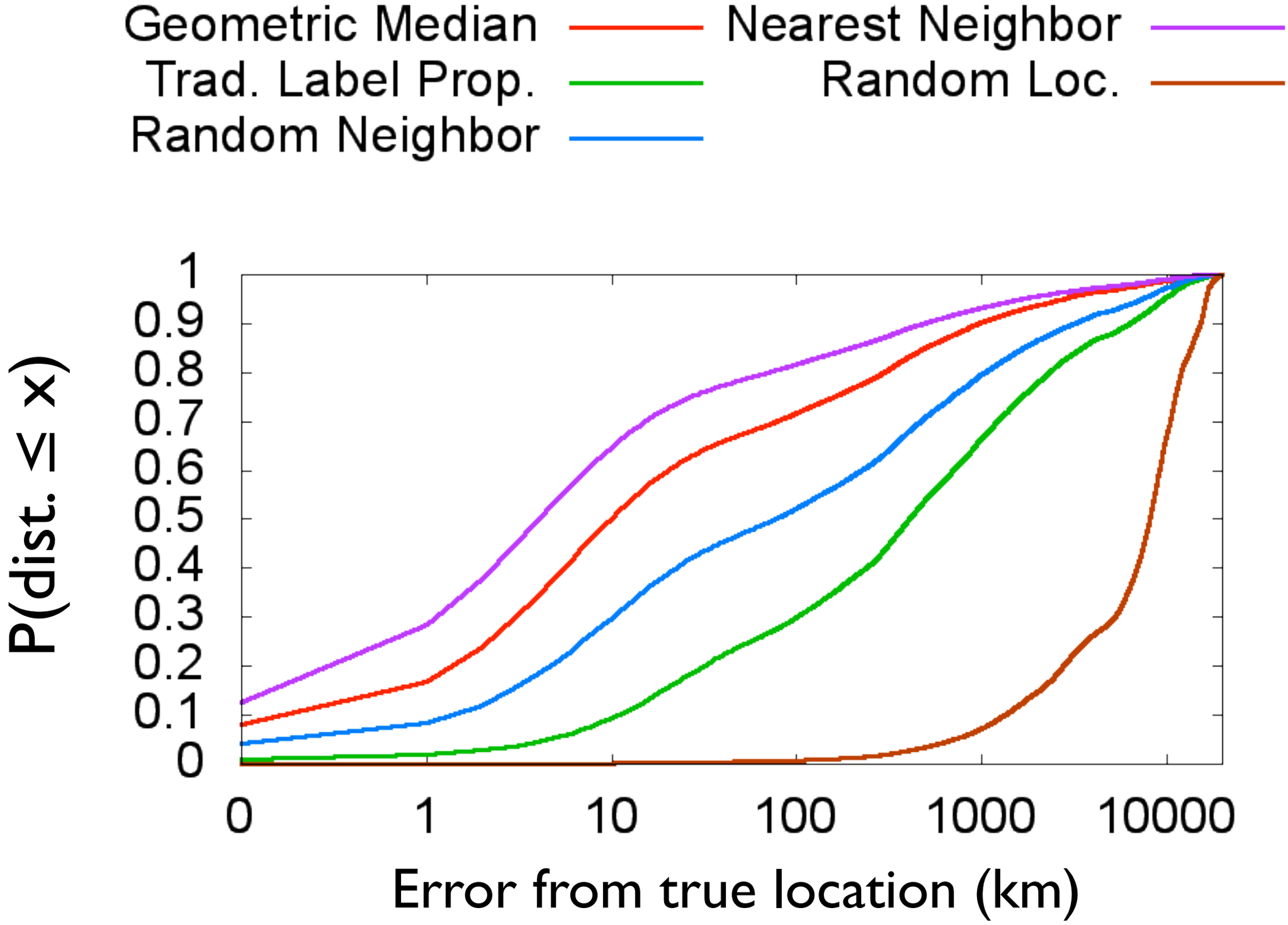
1. Pick any random user's location
2. Pick a random *friend's* location
3. Pick the most frequent location name among friends'

(assumes coordinates have been converted to names)

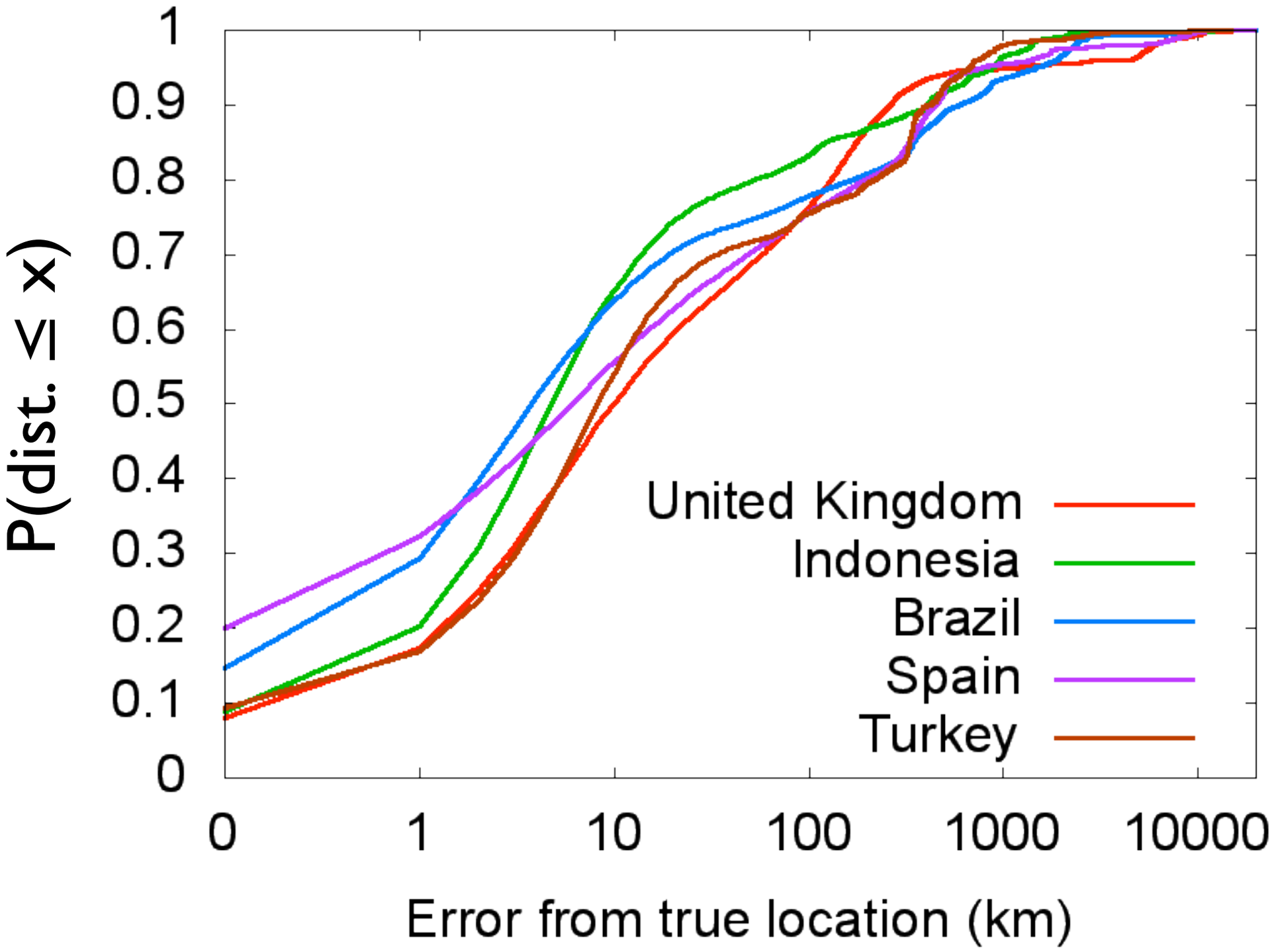
Evaluation Methodology

- Partition users with known locations into five sets
- Hold out one set, run method on complete graph using other four as seed locations (~2M seeds; 4% of network labeled)
- Measure error on held out set (0.5M test)

Results

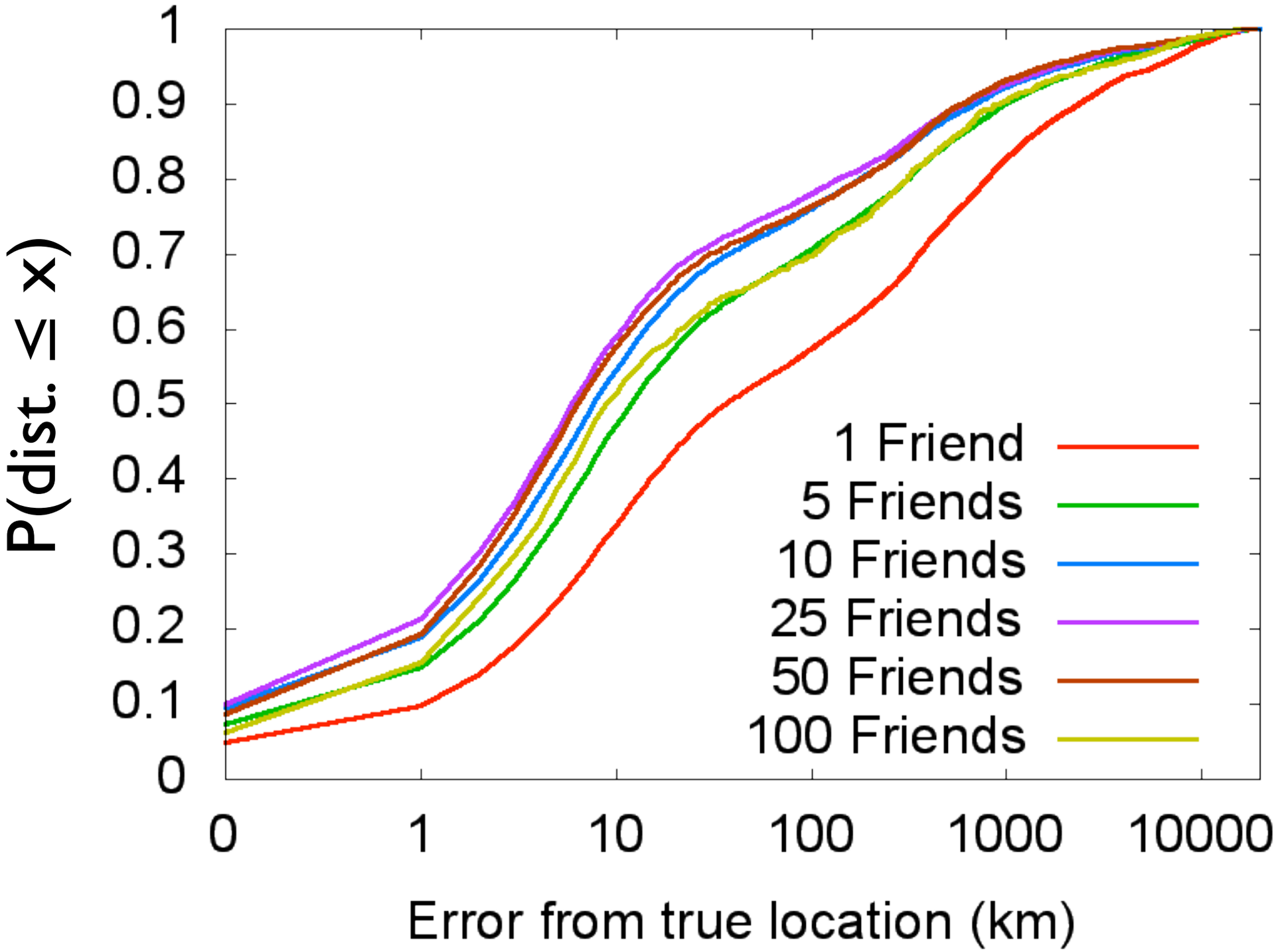

Geometric Median
Trad. Label Prop.
Random Neighbor
Nearest Neighbor
Random Loc.
P(dist. ≤ x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
10
100
1000
10000
Error from true location (km)

Country-level Results



Using geometric median only

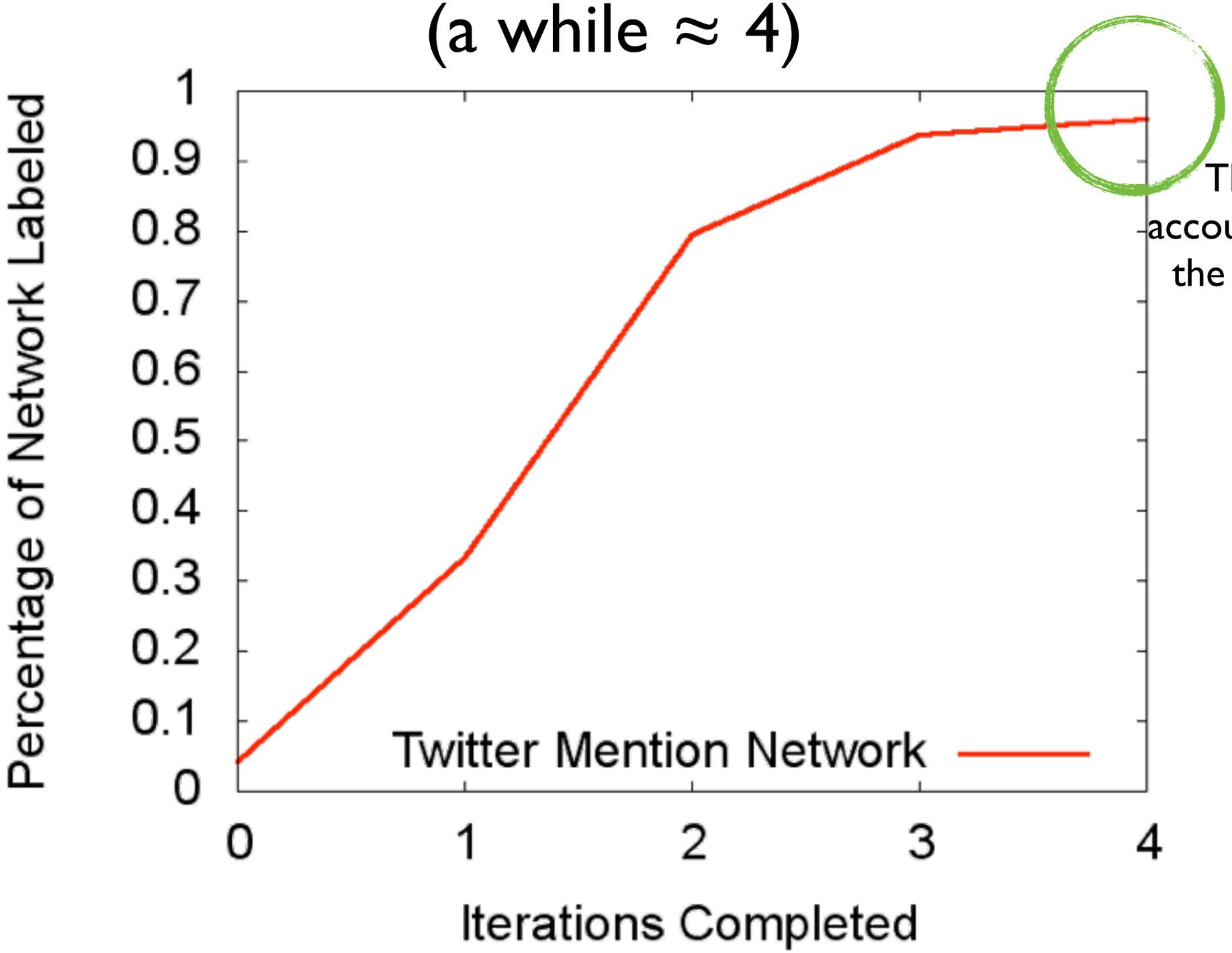
Results per ego-network size



Using geometric median only

Convergence is quick

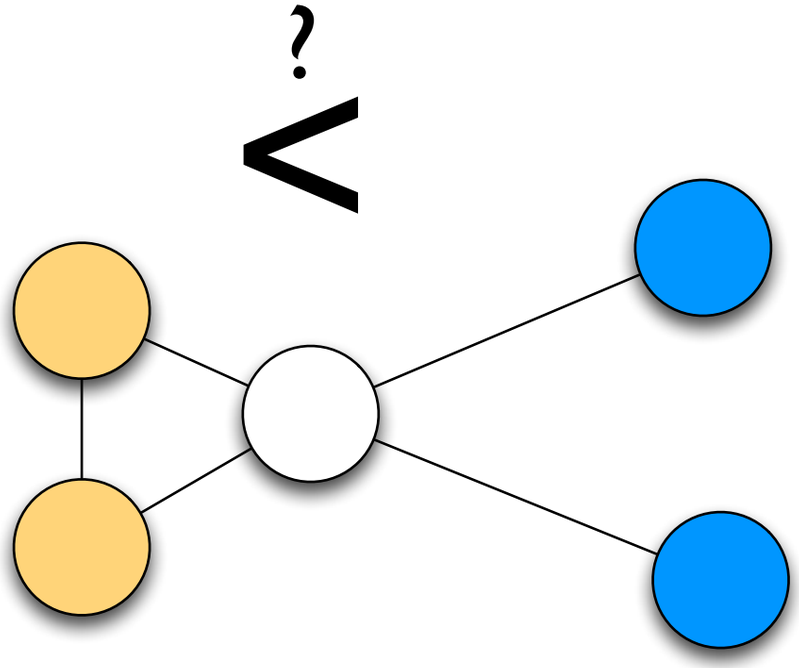
(a while ≈ 4)



These users account for 77% of the daily Twitter volume

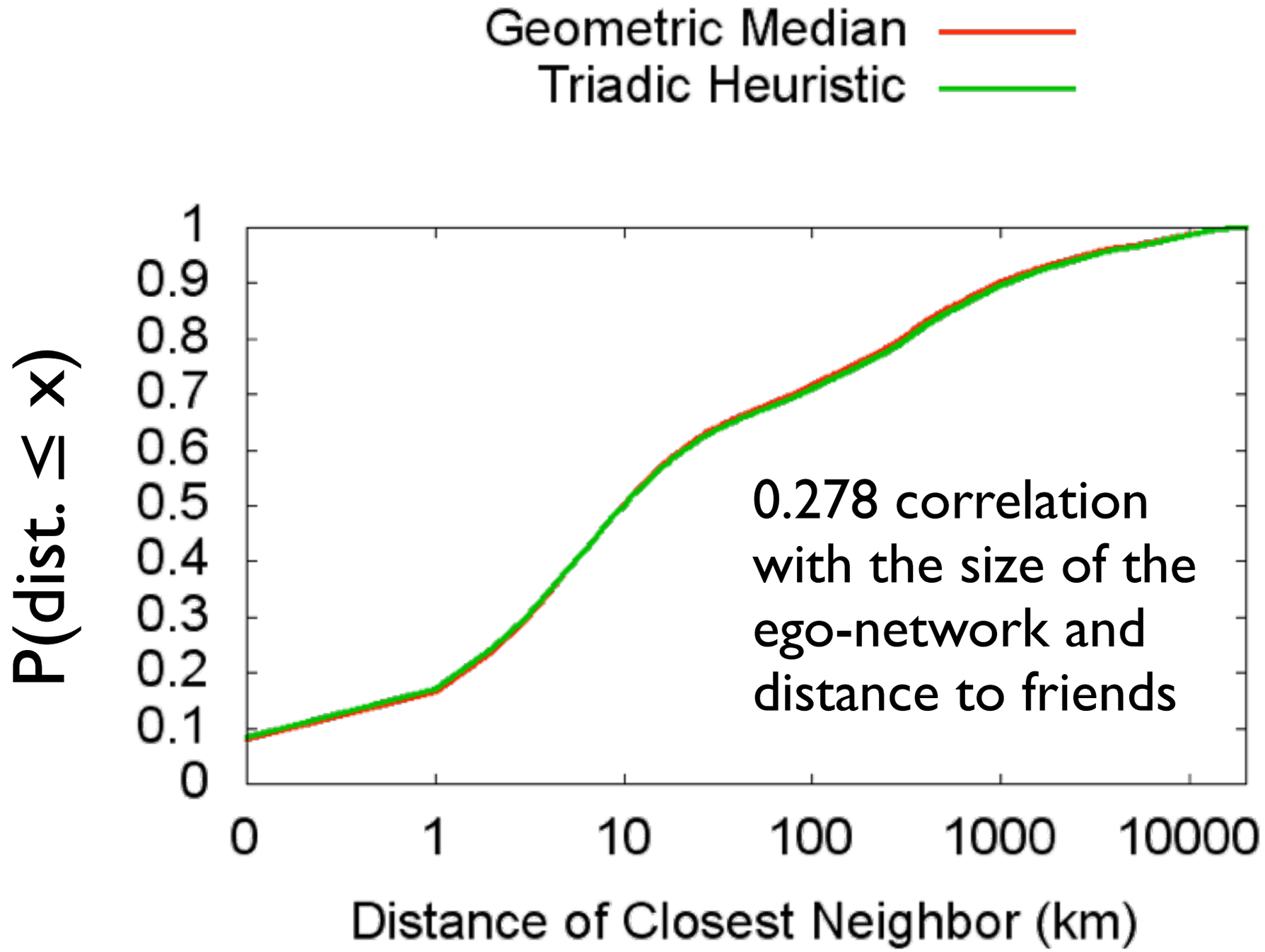
Can we do better?

RQ1: Does triadic structure predict locality?

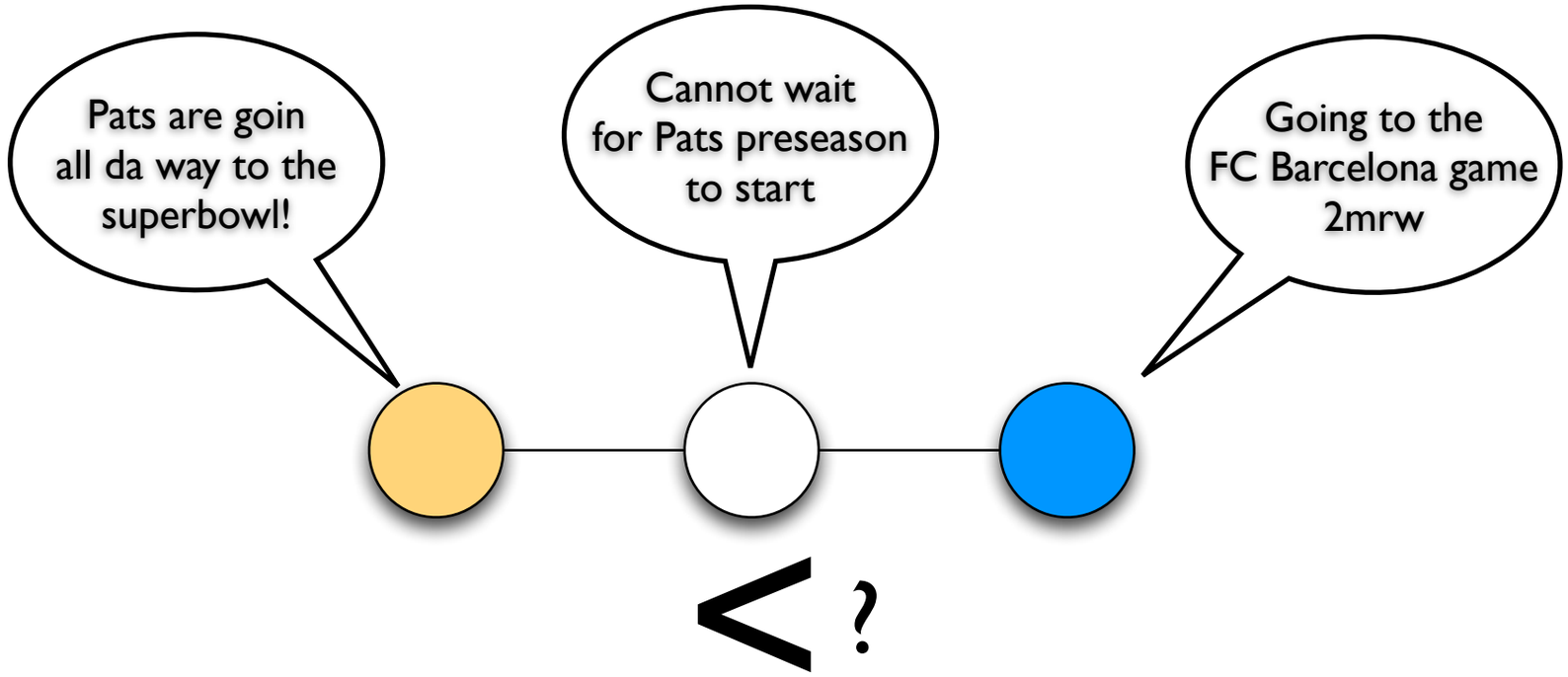


Pick the geometric median among the locations for closed triads in the ego-network

RQ1: Does triadic structure predict locality? No



RQ2: Does linguistic similarity predict geographic closeness?



- Two representations of all of a user's tweets
  - A unigram language model
  - A vector-space based model
- Correlate the similarity of two users' representations with their distances

RQ2: Does linguistic similarity predict geographic closeness? No*

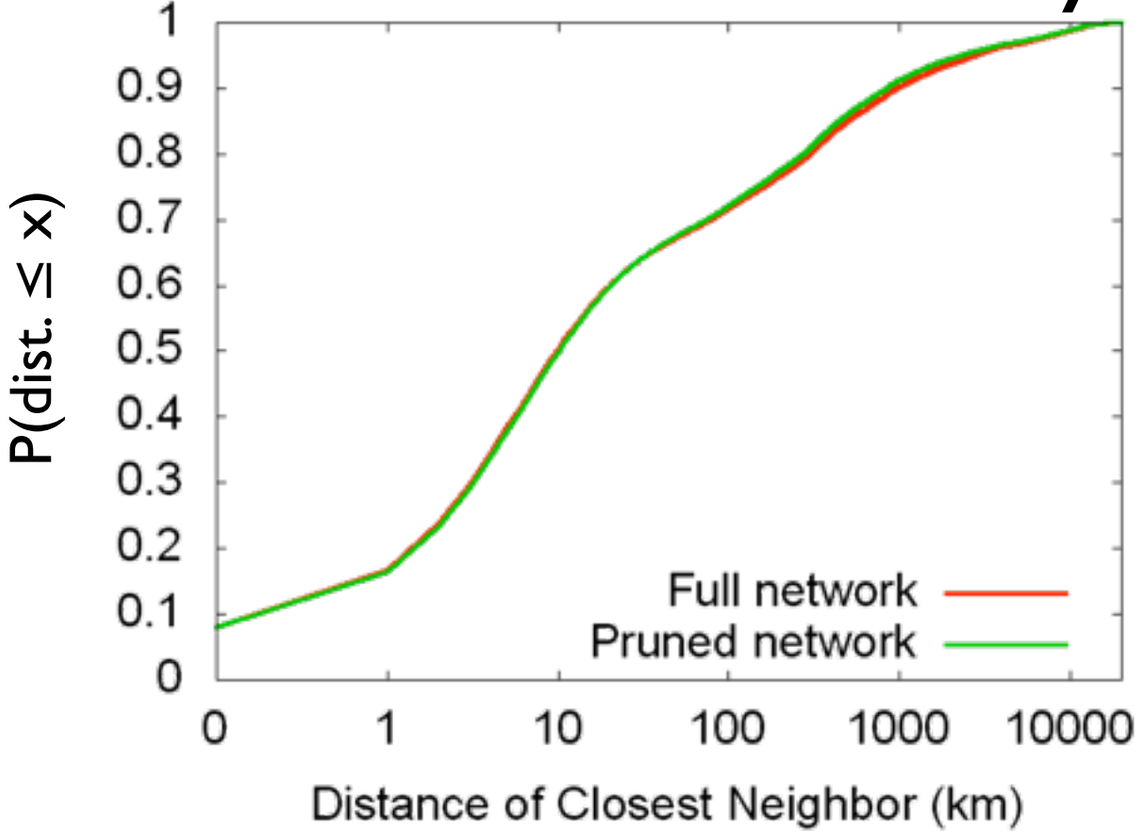
- 0.030 Spearman's correlation for language model
- 0.011 Spearman's correlation for vector space
- Correlation was consistent across country and ego-network size

RQ3: Can we improve using platform metadata?

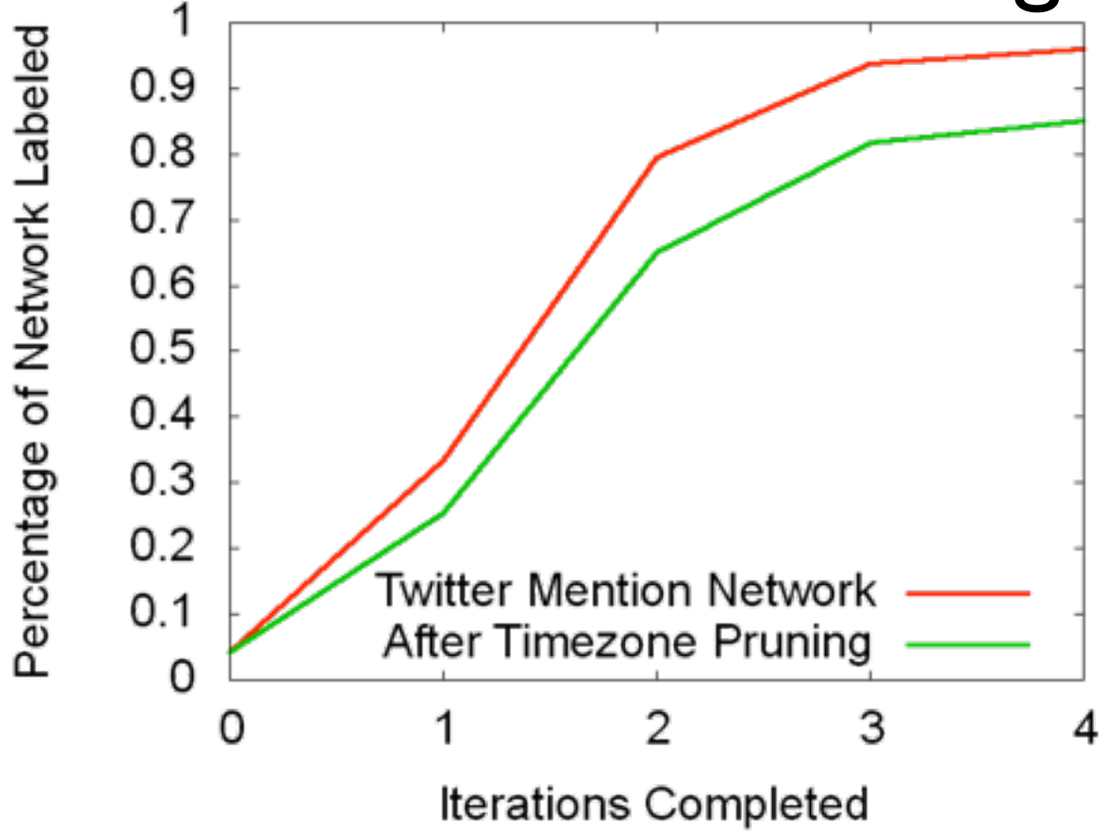
- Leverage self-reported Time Zone data
- Remove a relationship between two users if their set of time zones is disjoint
  - But only if they self-report
- Pruned 96.7M edges from network (38%)

RQ3: Can we improve using platform metadata? Sort of

No loss in accuracy



Some loss in coverage



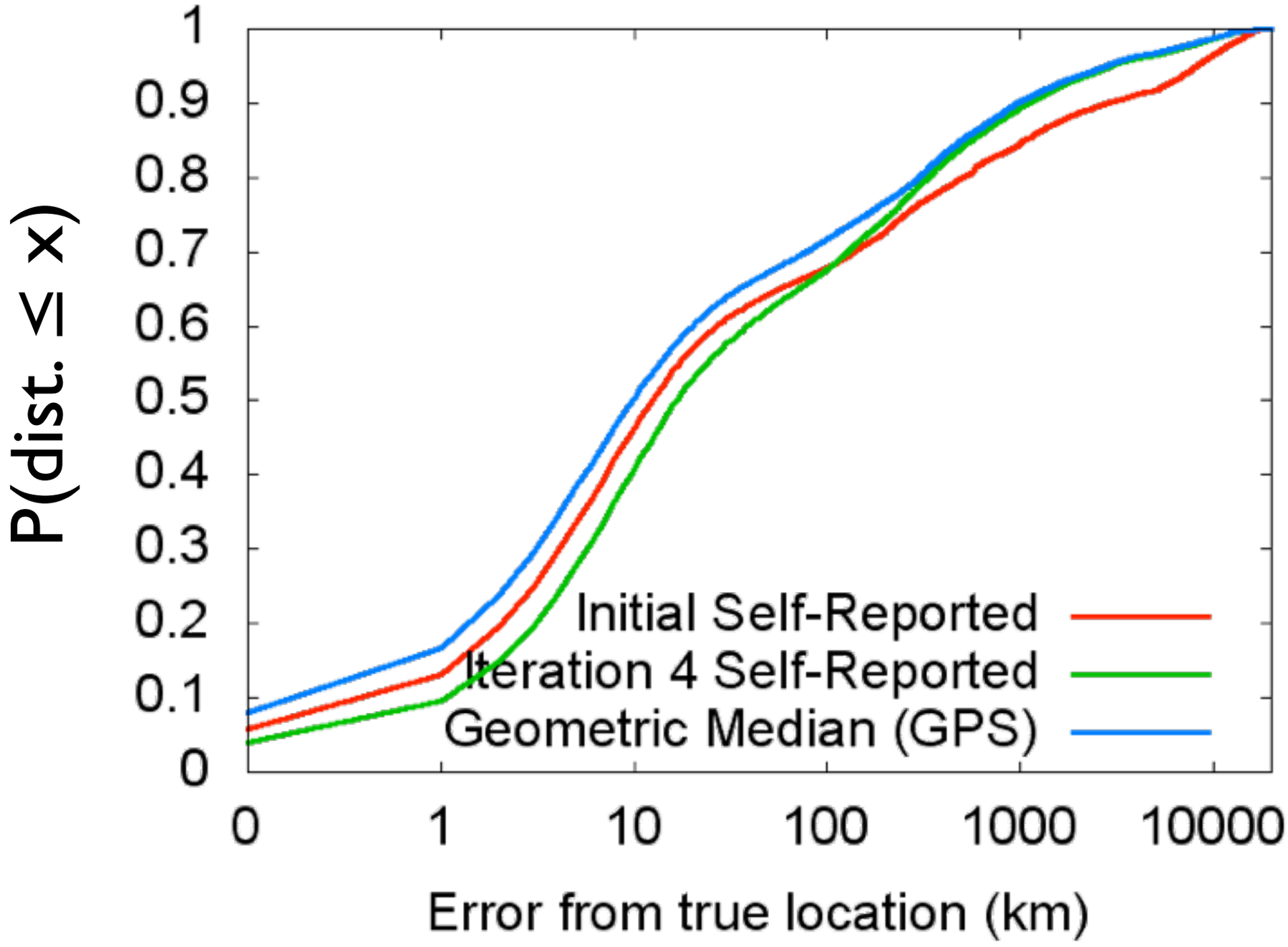
3X performance improvement

What if we had no ground truth?

Option 1: Use whatever users provide

- Conservatively map self-reported location names to coordinates
  - 11.3M users tagged (23.7%)
- Run using only self-reported data and test against held-out GPS data

Option 1: Use whatever users provide

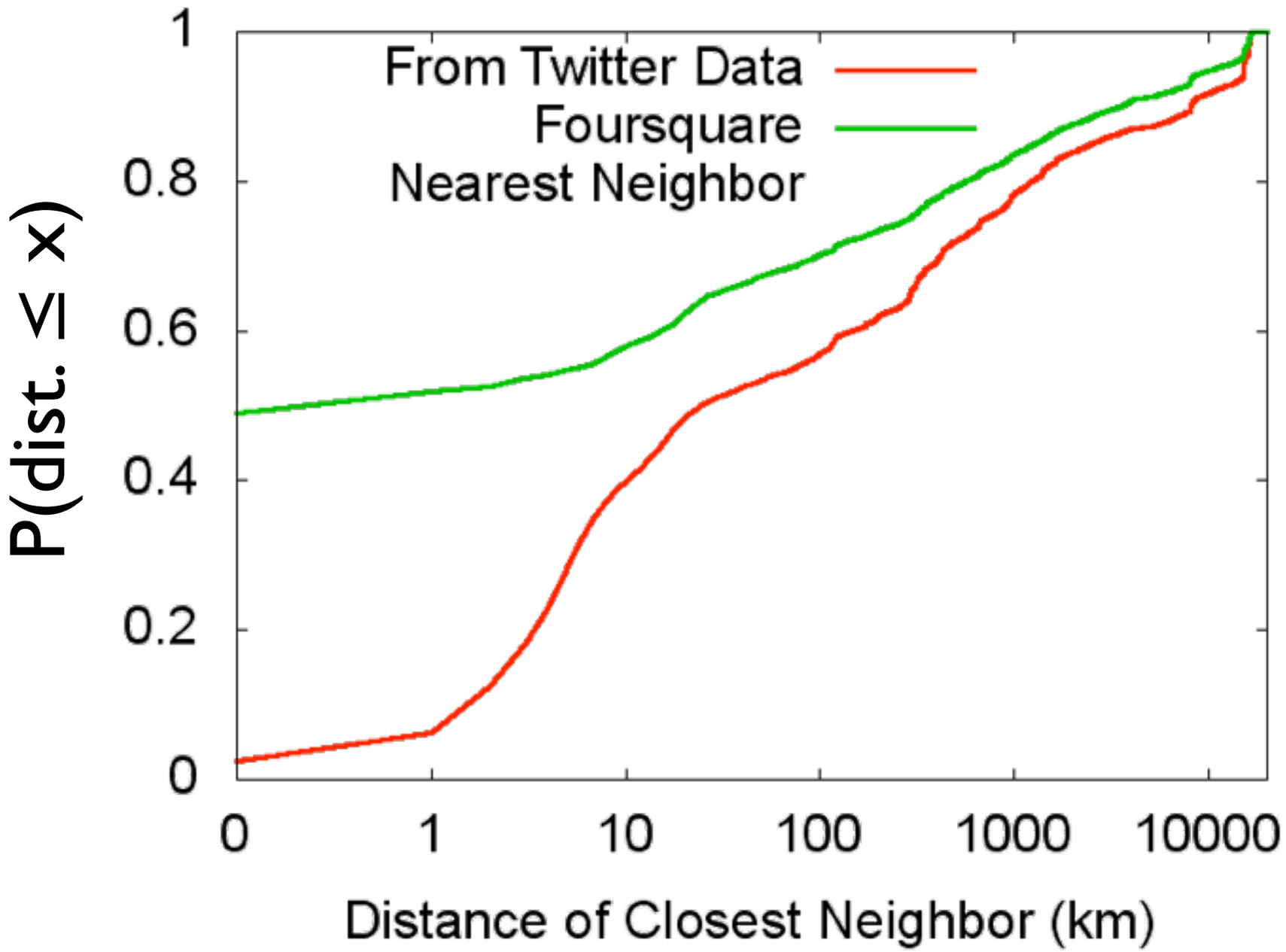


Option 2: Get the locations from *another* online social network

Option 2: Get the locations from *another* online social network

- **Goal**: Predict locations of Foursquare users using *only* location data from Twitter
- Merge the networks using the 1.6M of the 4M Foursquare users who have identities in both platforms
- Test on Foursquare-only users

Option 2: Get the locations from *another* online social network



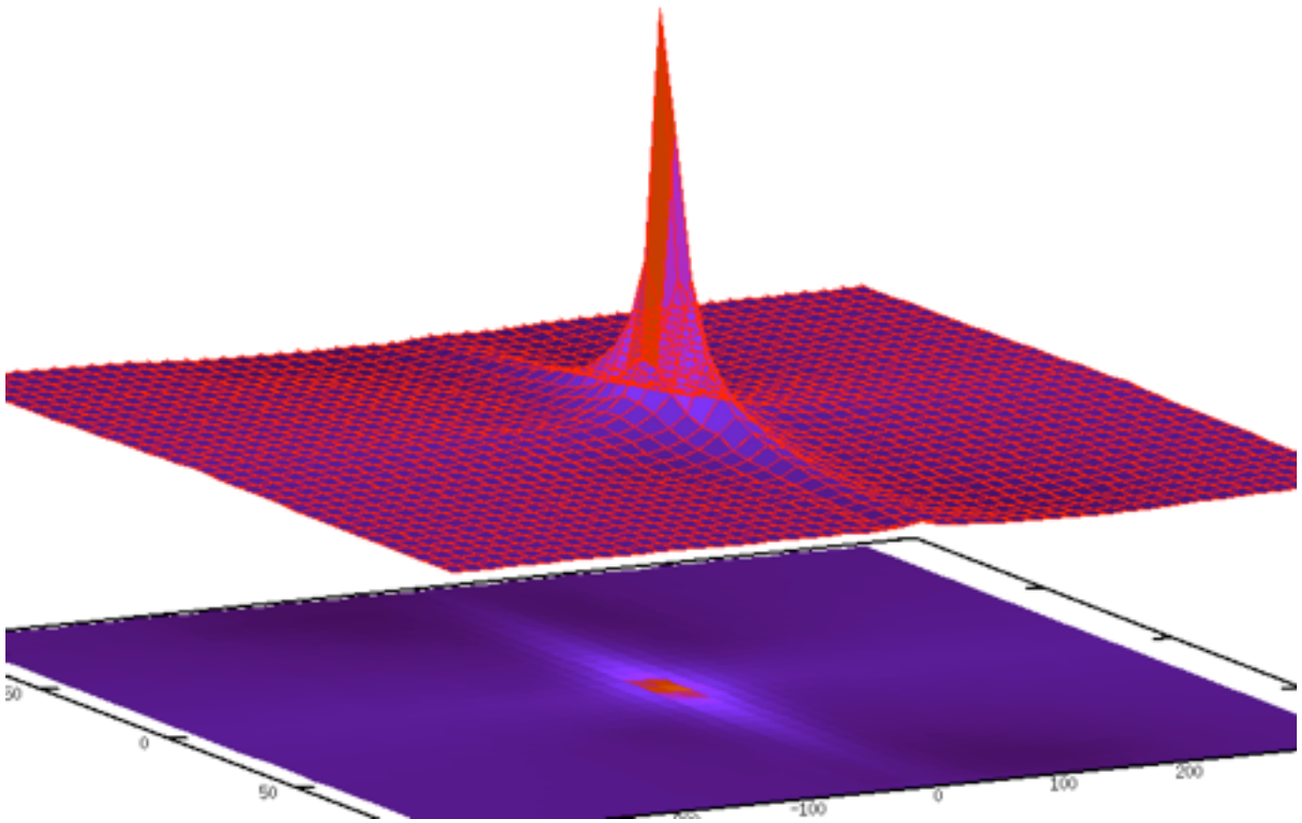
Insights

- Social networks provide a huge source of location information
- A little bit of good location data goes a long way, but even bad data is okay
- Multi-platform identities enable having new types of geolocated data

Open Questions

- What types of communications do predict locality?
- How does the structure of the ego-network relate to locality?
- What benefit can be seen by applying both network-based and linguistic based geolocation approaches

Thank you



David Jurgens

jurgens@di.uniroma1.it

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI / NBC) contract number D12PC00285. The IARPA research focuses solely on Latin America. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBE, or the U.S. Government.