Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization

Ryan Compton*, David Jurgens, David Allen

*rfcompton@hrl.com

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00285.

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/NBC, or the U.S. Government.

© 2014 HRL Laboratories, LLC All Rights Reserved

Introduction

- Several research directions rely heavily on geographically-annotated social media (which is extremely rare in public data)

- Twitter feed from April 2012 until April 2014:
  - 37,400,698,296 tweets
  - 359,583,211 users
  - ~77TB of data

Geotag method	Tweets	Users
GPS	584,442,852 (1.6%)	22,413,350 (6.2%)
Profile text (anything, e.g. “beiberland”)	23,236,139,825 (62%)	164,020,169 (45.6%)
Profile text (unambiguous)	3,854,169,186 (10%)	45,284,996 (12.6%)

© 2014 HRL Laboratories, LLC All Rights Reserved

Related Work: Academic

Academics (mostly NLP) are getting interested:

- Eisenstein et al. “A latent variable model for geographic lexical variation” EMNLP 2010
- Cheng et al. “You are where you tweet: a content-based approach to geolocating twitter users” CIKM 2010
- Mahmud et al. “Where is this tweet from? inferring home locations of twitter users” ICWSM 2012
- Rahimi et al. “Exploiting Text and Network Context for Geolocation of Social Media Users” NAACL 2015

1:59 pm ET
Apr 8, 2014 BIG DATA

The Race to Locate Twitter Users

ARTICLE COMMENTS (2)

ESRI GEOLOCATION GNIP IBM RESEARCH TWITTER

Email Print

By ELIZABETH DWOSKIN CONNECT

Few Twitter TWTR +0.41% users broadcast their location. But businesses and researchers are hunting for ways to infer it.

— Getty Images

IBM IBM -0.42% researcher Jalal Mahmud and colleagues created software that can often identify a Twitter user’s home city, based on the user’s 200 most-recent tweets, according to a recent paper. The researchers looked at the times when a user tweets most frequently – which can indicate the user’s time zone – as well as mentions of sports teams and unique place names.

Mahmud says his model can predict a Twitter user’s home city among the 100 largest U.S. cities within a second with 70% accuracy. Outside those cities, the accuracy declines. His team has filed a patent application for the algorithm.

Researchers say fewer than 3% of Twitter users enable a “geo-tagging” feature that allows app developers to see the latitude and longitude of their tweets. About 30% of users list a location in their Twitter bios, but others list false or fictional locations, such as “in Justin Bieber’s heart.”

Twitter itself has more details on where users are. As long as a user activates the location feature on a smartphone, each tweet is marked with the phone’s location. Twitter uses this and other information to offer geo-targeted ads to areas as small as a zip code; it delivers the ad to users without disclosing their identities to the advertiser.

But Twitter doesn’t share this location data with outsiders, leaving businesses and other groups that want to locate Twitter users largely on their own.

“There is a veritable arms race” to locate Twitter users, said Kalev Leetaru, Yahoo Fellow at Georgetown University.

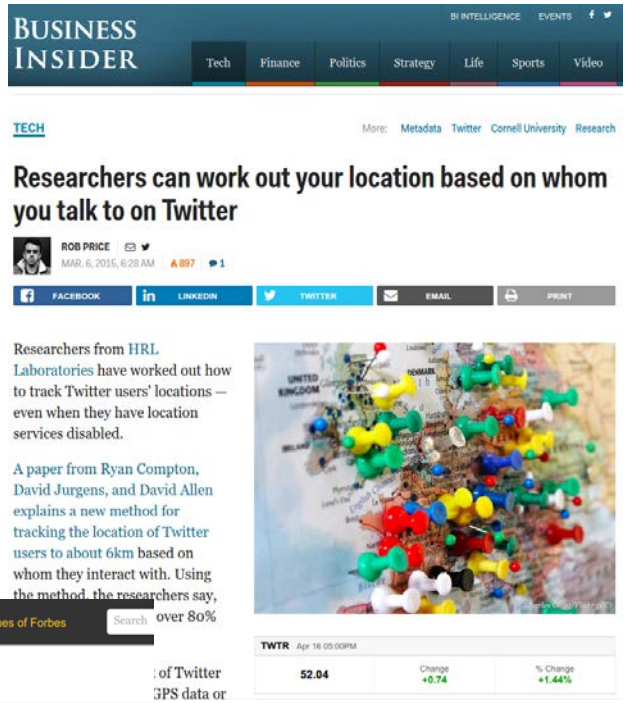
Businesses may want to locate Twitter users to analyze regional differences in sentiment. During a natural disaster, relief organizations have tried to pinpoint flooding to a specific city block.

© 2014 HRL Laboratories, LLC All Rights Reserved

Related Work: Academic

Network science approaches:

- Backstrom et al. “Find me if you can: improving geographical prediction with social and spatial proximity” WWW 2010
- David Jurgens “That's what friends are for: Inferring location in online communities based on social relationships” ICWSM 2013
- **Present work**



© 2014 HRL Laboratories, LLC All Rights Reserved

Related Work: Commercial

- Twitter geocoding has become a commercial product

- *“For years now at GNIP, the most requested feature for our existing data has been `more geodata’ ”*

- Competitors exist:
  - Datasift offering geolocation
  - Tweepsmap.com

Get More Twitter Geodata From Gnip With Our New Profile Geo Enrichment

Posted on August 22, 2013 by Ian Cairns, Product 1 Comment

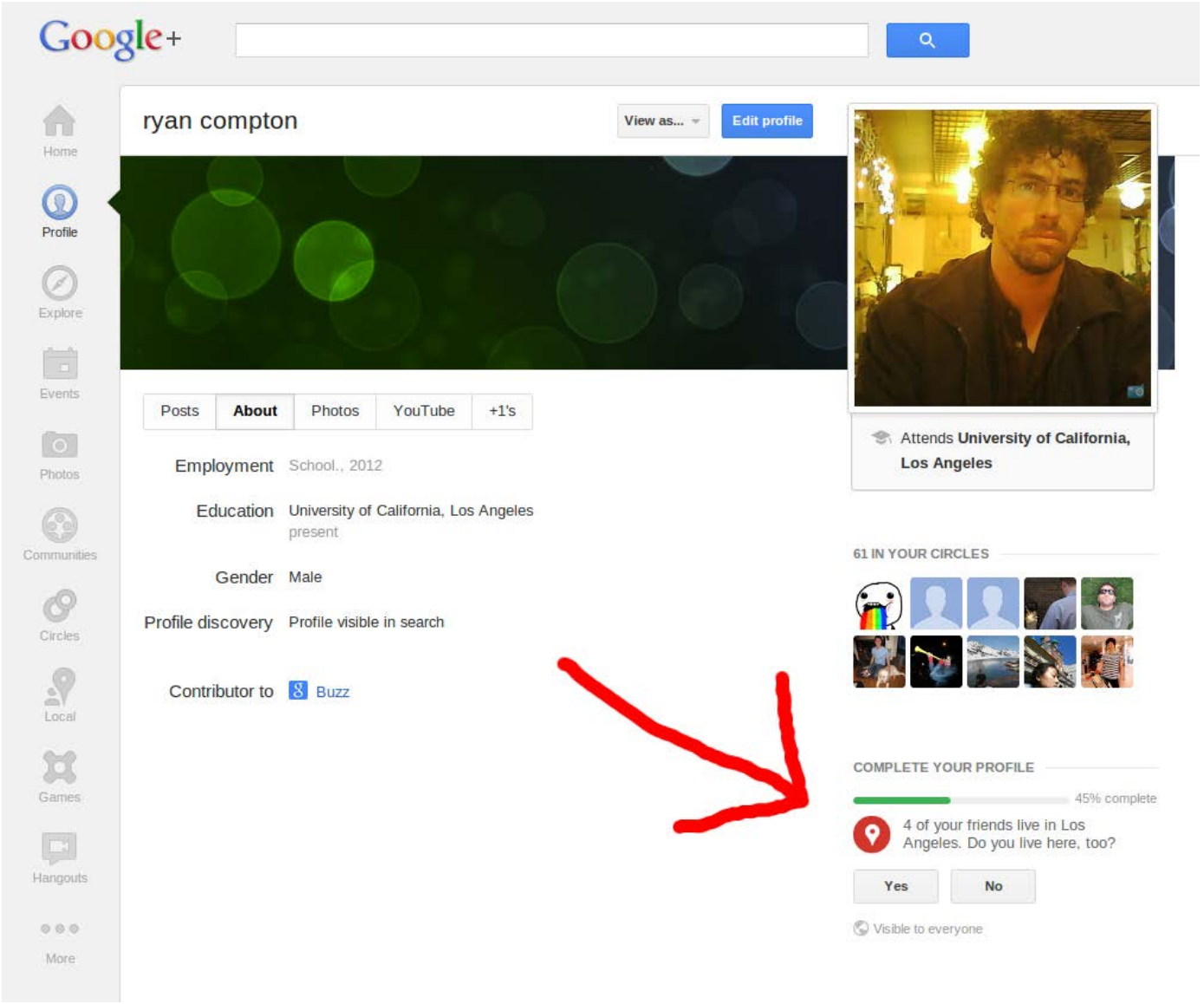
Tweet 86 Like 5 +1 3 Share 27 submit

When it comes to analyzing social data, “where” matters. After the topics of conversations, perhaps the strongest connection between social conversations online and the offline world is location. Location is an implicit part of what we do, who we know, what we need, etc. For years now at Gnip, the most requested feature for our existing data products has been “more geodata” to help our customers understand the offline locations that are relevant to online conversations. Today we’re pleased to announce a major step toward meeting that demand: the public beta launch of our new Profile Geo enrichment.

The Profile Geo enrichment is simple. Location data is provided publicly by millions of users in their profiles

© 2014 HRL Laboratories, LLC All Rights Reserved

Unpublished Results Likely Exist



© 2014 HRL Laboratories, LLC All Rights Reserved

Method:
Social network analysis for static location inference

© 2014 HRL Laboratories, LLC All Rights Reserved

Network Data

- We build a social network from @mentions
- 10% Twitter data from April 2012 through April 2014
- 25,312,399,718 @mentions (any type)
- Twitter @mention network
  - 8,593,341,111 edges in weighted unidirected network
  - **1,034,362,407 edges** in weighted bidirected network
  - **110,893,747 users** in weighted bidirected network
- Reciprocated @mentions indicate social ties (i.e. the “bidirected network”)

© 2014 HRL Laboratories, LLC All Rights Reserved

GPS Ground-truth user locations

- 13,899,315 users have tweeted with GPS at least three times
- Remove users with dispersion > 30km

$$\underset{x \in \mathcal{G}}{\text{median}} \left(d(x, \text{median}(\mathcal{G})) \right)$$

- Reduces to 12,435,622 users
- Timestamps reveal 86,243 users exceeded the flight airspeed record of 3529.6 km/h (bots, GPS malfunctions)*
- We further remove any user who traveled faster than 1000 km/h
- Leaves us with **12,297,785 GPS-known users**
- **Use l1-multivariate median as "home"**

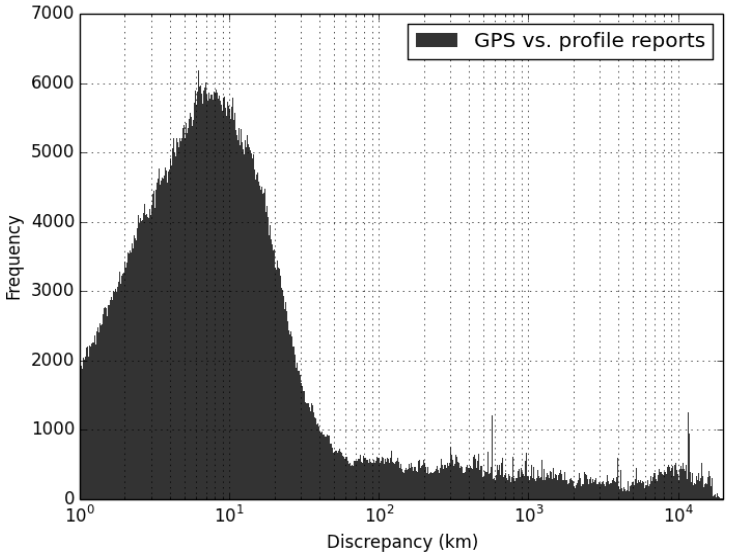
$$\text{median}(\mathcal{G}) = \underset{x}{\text{argmin}} \sum_{y \in \mathcal{G}} d(x, y)$$

*Note: the maximum speed attained by any Twitter user in our data was 67,587,505.24 km/h, over 30x the escape velocity from the Sun

© 2014 HRL Laboratories, LLC All Rights Reserved

Self-reported profile locations

- We extract self-reported home locations by searching profiles for an exact match against a list of 51,483 unambiguous location names
- List obtained via geonames.org, filtered against GPS-known users who have also self-reported profile locations
- 7.10 km median discrepancy with GPS



- 15,360,494 self-report users most recent 90 days

GPS-known or self-reporting users: 24,545,425

© 2014 HRL Laboratories, LLC All Rights Reserved

Geolocation as an Optimization Problem

- **We seek a network such that the sum over all geographic distances between connected users is as small as possible**

$$\min_{\mathbf{f}} |\nabla \mathbf{f}| \text{ subject to } f_i = l_i \text{ for } i \in L$$

$$|\nabla \mathbf{f}| = \sum_{ij} w_{ij} d(f_i, f_j)$$

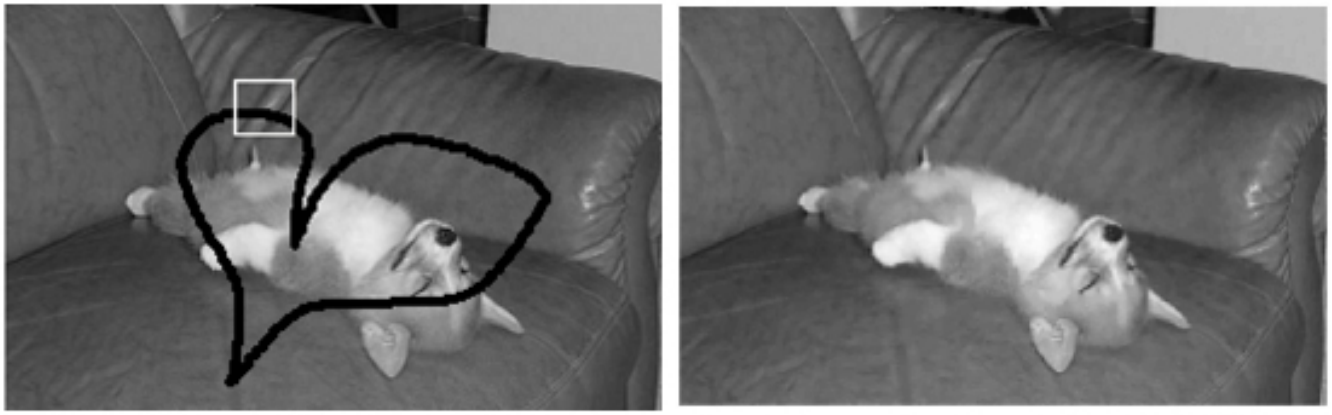
- **f** encodes a location estimate for each user
- L is the set of ground-truth user locations
- *w_ij* is the minimum number of reciprocated @mentions between users i and j
- *d* measures geodetic distance via Vincenty's formulae

© 2014 HRL Laboratories, LLC All Rights Reserved

Remark

$$|\nabla \mathbf{f}| = \sum_{ij} w_{ij} d(f_i, f_j)$$

- **Our objective function is often referred to as *total variation***
- **Used in image processing since 1992**
- **Starting to find use in machine learning**



https://plus.maths.org/content/restoring-profanity

Rudin, Leonid I., Stanley Osher, and Emad Fatemi. "Nonlinear total variation based noise removal algorithms." *Physica D: Nonlinear Phenomena* 60.1 (1992): 259-268.

Bresson, X., Laurent, T., Uminsky, D., & von Brecht, J. (2013). Multiclass total variation clustering. In *Advances in Neural Information Processing Systems* (pp. 1421-1429).

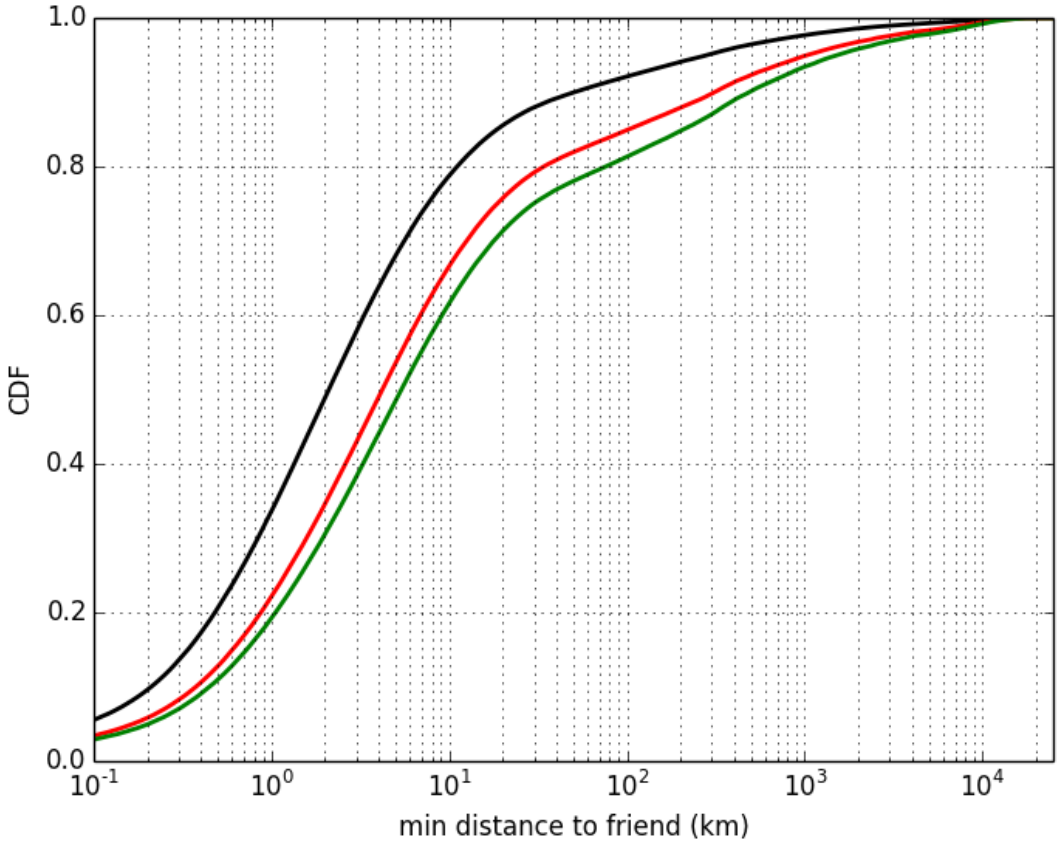
© 2014 HRL Laboratories, LLC All Rights Reserved

Assumption: Online Social Ties Correspond to Proximity

- **Twitter:** Yuri Takhteyev et al. “Geography of twitter networks.” Social Networks 2012.
- **Facebook :** Backstrom et al. “Find me if you can: improving geographical prediction with social and spatial proximity” WWW 2010
- **Wikipedia**: Lieberman and Lin. "You Are Where You Edit: Locating Wikipedia Contributors Through Edit Histories" *ICWSM* 2009

© 2014 HRL Laboratories, LLC All Rights Reserved

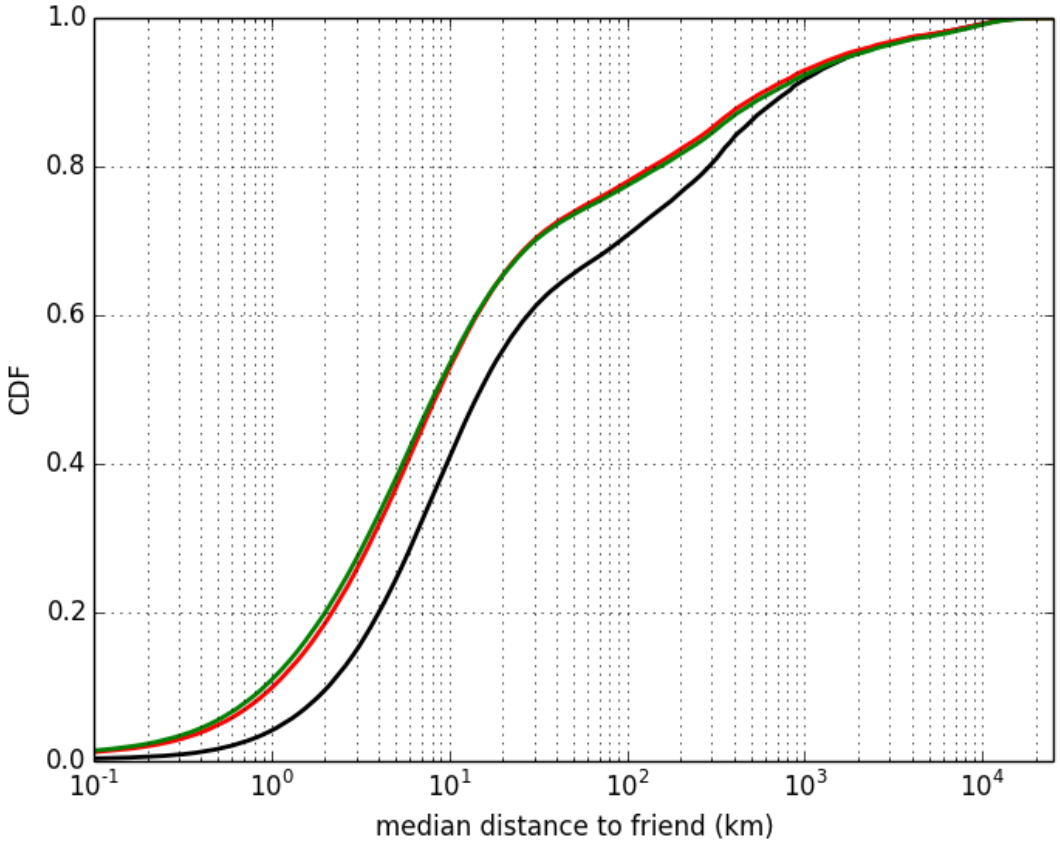
Proximity of Social Ties in our Data



Restrict to 953,557 edges which occur between GPS-known users in three networks: unidirected (black), bidirected (red), triadically-closed subgraph of the bidirected network (green)

© 2014 HRL Laboratories, LLC All Rights Reserved

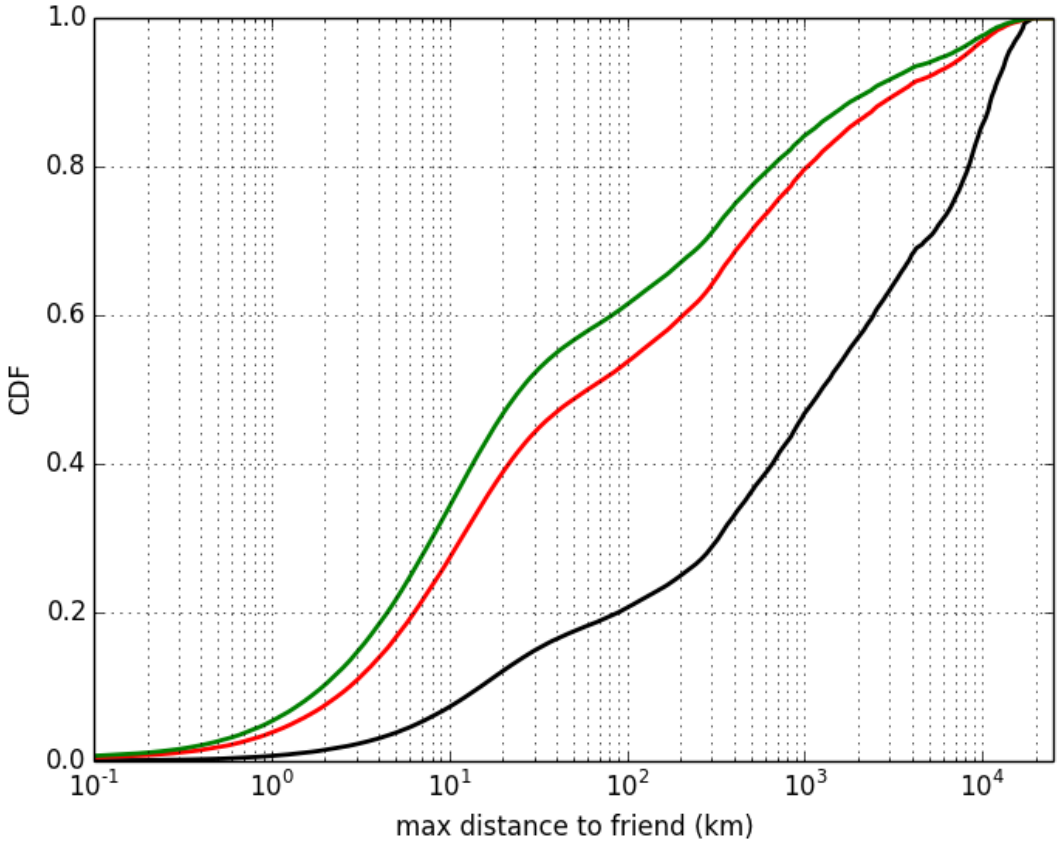
Proximity of Social Ties in our Data



Restrict to 953,557 edges which occur between GPS-known users in three networks: unidirected (black), bidirected (red), triadically-closed subgraph of the bidirected network (green)

© 2014 HRL Laboratories, LLC All Rights Reserved

Proximity of Social Ties in our Data



Restrict to 953,557 edges which occur between GPS-known users in three networks: unidirected (black), bidirected (red), triadically-closed subgraph of the bidirected network (green)

© 2014 HRL Laboratories, LLC All Rights Reserved

Not Everyone Tweets with their Neighbors

Additional heuristic: Don't use social networks to infer location for users whose friends are dispersed around the globe

Quantify geographic dispersion using median absolute deviation

$$\min_{\mathbf{f}} |\nabla \mathbf{f}| \text{ subject to } f_i = l_i \text{ for } i \in L \text{ and } \tilde{\nabla} f_i \leq \gamma$$

$$\tilde{\nabla} f_i = \text{median}_j \left(w_{ij} d(f_i, f_j) \right)$$

© 2014 HRL Laboratories, LLC All Rights Reserved

Implementation

© 2014 HRL Laboratories, LLC All Rights Reserved

Remarks on Infrastructure

- 35-node cluster
- 32 processors per node (i.e. 1120 cores)
- 363.88TB total hdfs capacity
- 128g memory per node
- Current bidirected network requires over 200g once loaded
- Deployed distributed processing framework:
  - Hadoop file system for storage 
  - Pig for GPS and network extraction 
  - Spark for graph computation 
  - Redis for fast access to results 

© 2014 HRL Laboratories, LLC All Rights Reserved

Parallel Coordinate Descent

- Distributed graph processing is hard
- Richtarik, “Parallel coordinate descent methods for big data optimization.” arXiv:1212.0873
- Lyubich et al. “Subharmonic functions on a directed graph” Siberian Mathematical Journal, 1969 (convergence?)

Algorithm 1: Parallel coordinate descent for constrained TV minimization

Initialize: $f_i = l_i$ for $i \in L$
for $k = 1 \ldots N$ **do**
  parfor i **:**
    if $i \in L$ **then**
      $f_i^{k+1} = l_i$
    else
      $f_i^{k+1} = \underset{f}{\operatorname{argmin}} |\nabla_i(\mathbf{f}^k, f)|$
    end
  end
  $\mathbf{f}^k = \mathbf{f}^{k+1}$
end

© 2014 HRL Laboratories, LLC All Rights Reserved

Parallel Coordinate Descent

- Include the dispersion constraint in the simplest way possible

Algorithm 2: Parallel coordinate descent for dispersion-constrained TV minimization.

Initialize: $f_i = l_i$ for $i \in L$ and parameter γ
for $k = 1 \ldots N$ **do**
  parfor i **:**
    if $i \in L$ **then**
      $f_i^{k+1} = l_i$
    else
      if $\nabla \tilde{f}_i \leq \gamma$ **then**
        $f_i^{k+1} = \underset{f}{\operatorname{argmin}} |\nabla_i(\mathbf{f}^k, f)|$
      else
        no update on f_i
      end
    end
  end
  $\mathbf{f}^k = \mathbf{f}^{k+1}$
end

© 2014 HRL Laboratories, LLC All Rights Reserved

Parallel Coordinate Descent

- Implementation in is straightforward in Spark
- Advanced distributed graph-processing frameworks, such as GraphX, GraphLab, or some other implementation of Google's Pregel model are possible

Listing 1: Distributed implementation of alg. 2 using the Spark framework [36]

```
1  val edgeList = loadEdgeList()
2  var userLocations = loadInitialLocations()
3  for (k <- 1 to N) {
4    val adjListWithLocations = edgeList.join(userLocations) ↙
         .keyBy(x => x._2).groupByKey()
5    val updatedLocs = adjListWithLocations.map(x => (x._1, ↙
         l1Median(x._2), dispersion(x._2)))
6    userLocations = updatedLocs.filter(x => x._3 < GAMMA)
7  }
8  return userLocations
```

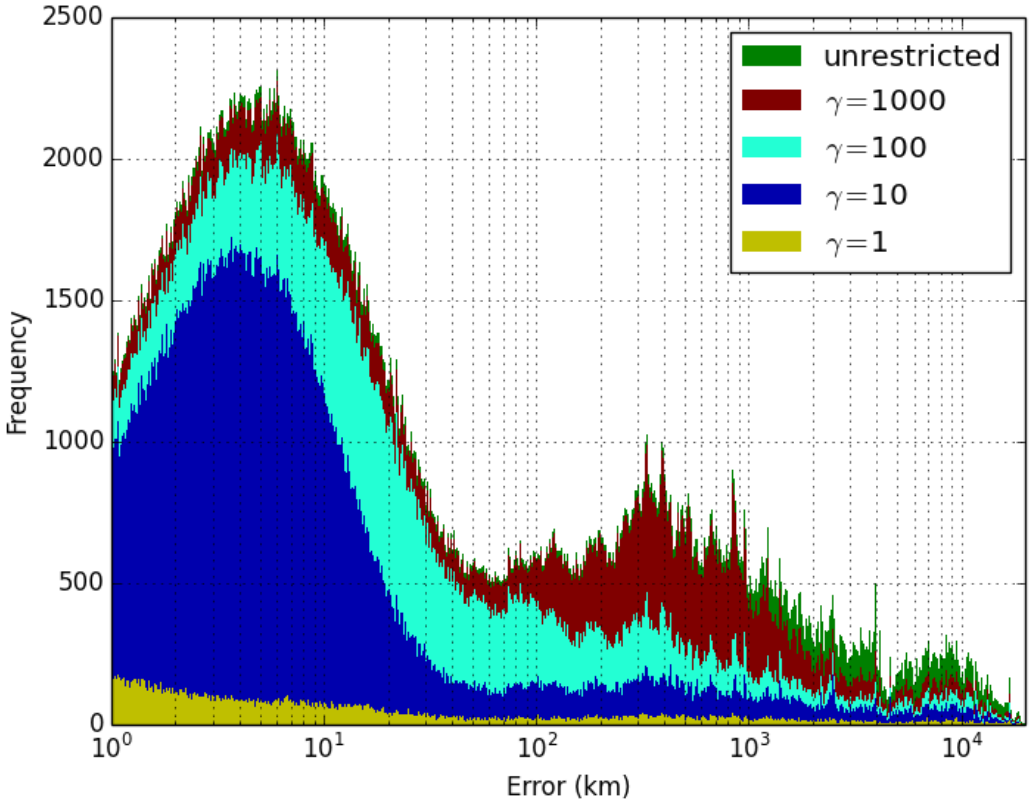
© 2014 HRL Laboratories, LLC All Rights Reserved

Results

© 2014 HRL Laboratories, LLC All Rights Reserved

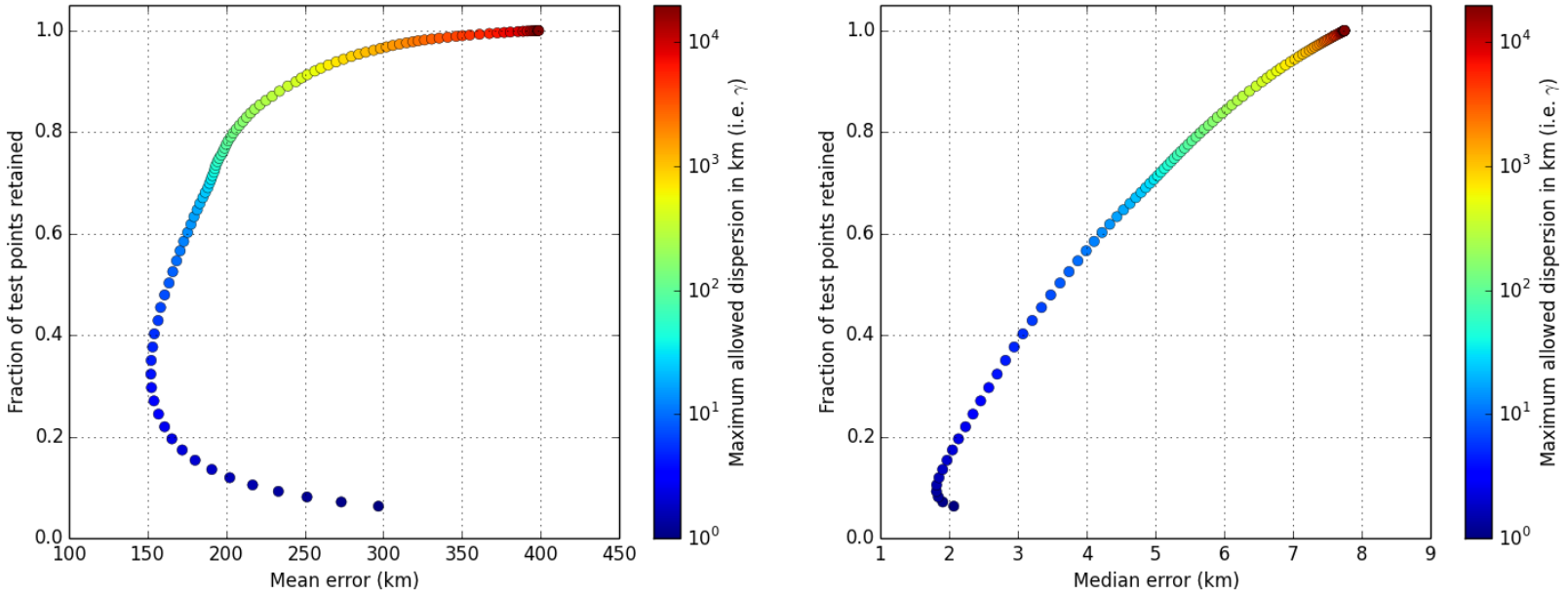
Leave-many-out validation

- Hold out 10% of GPS-known users
- Run 5 iterations
- Report discrepancy between GPS and inferred locations
- Algorithm 1:
  - **115,410,410** users
  - **8.27 km** median error
  - **430.56** km mean error
- Algorithm 2,100km dispersion:
  - **101,846,236** users
  - **6.33 km** median error
  - **291.5 km** mean error



© 2014 HRL Laboratories, LLC All Rights Reserved

Error Control via Dispersion Constraint



Scatter plots describing the coverage/accuracy trade off obtained when modifying gamma.

© 2014 HRL Laboratories, LLC All Rights Reserved

Error vs Iteration

Iteration	Test users	New test users	Median error (km)	Median new error (km)
1	771,321	771,321	5.34	5.34
2	926,019	15,468	6.02	12.31
3	956,705	30,686	6.24	45.50
4	966,515	9,810	6.32	150.60
5	971,731	5,216	6.38	232.92

© 2014 HRL Laboratories, LLC All Rights Reserved

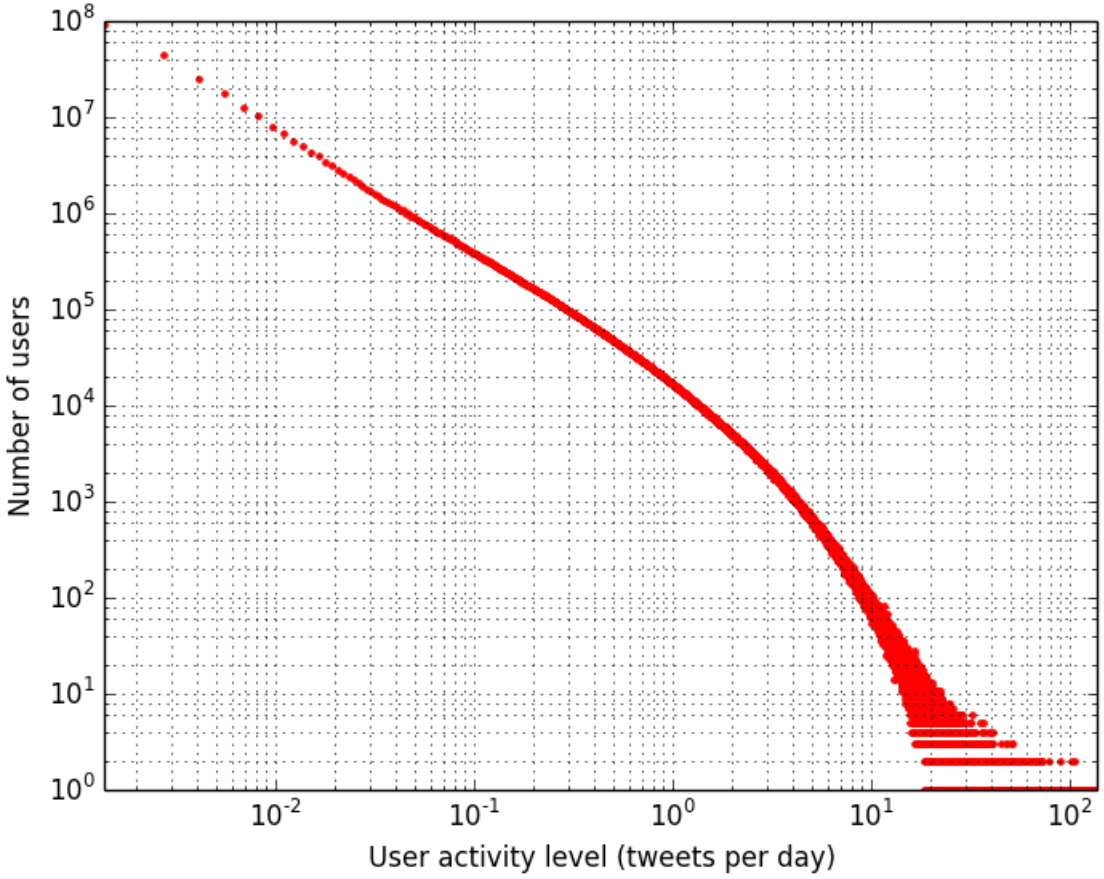
Coverage

- Study full Twitter dataset
- 37,400,698,296 tweets generated by 359,583,211 users

Geotag method	Tweets	Users
GPS	584,442,852 (1.6%)	22,413,350 (6.2%)
Profile text (anything, e.g. “beiberland”)	23,236,139,825 (62%)	164,020,169 (45.6%)
Profile text (unambiguous)	3,854,169,186 (10%)	45,284,996 (12.6%)
Total Variation	30,617,806,498 (81.9%)	101,846,236 (28.3%)

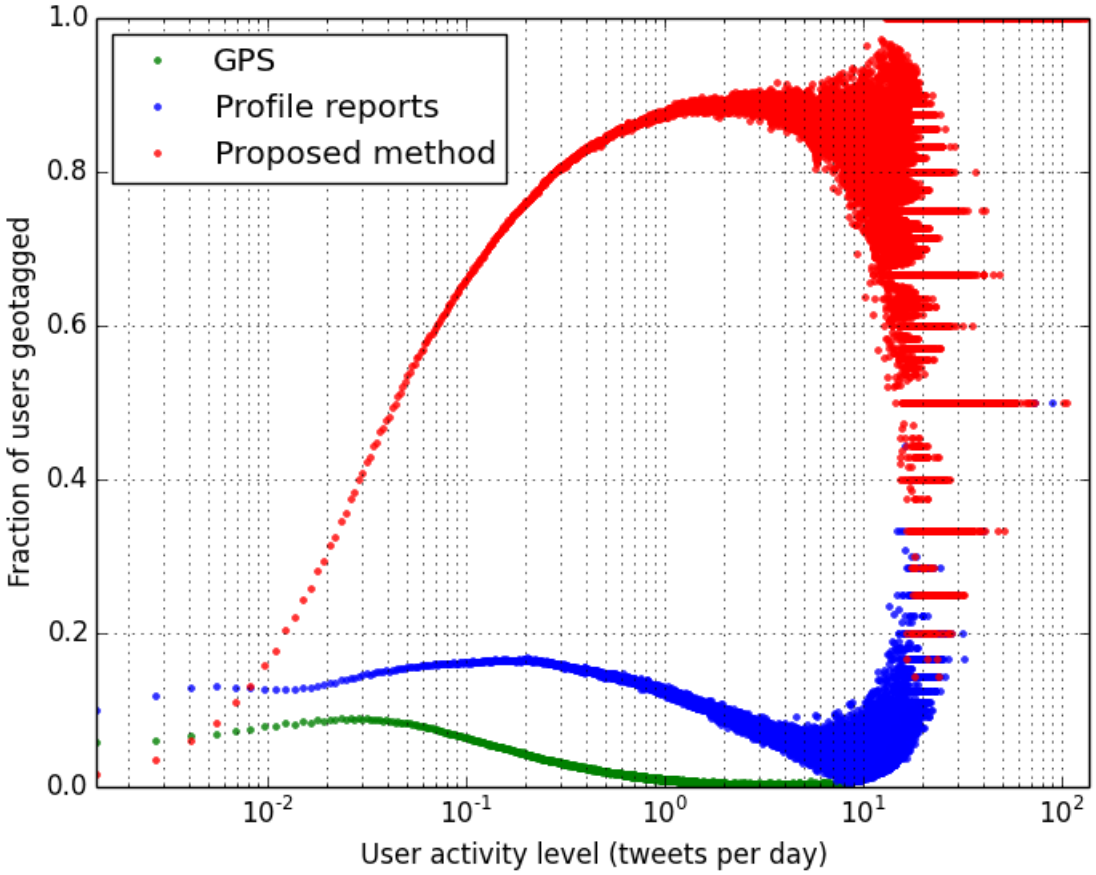
© 2014 HRL Laboratories, LLC All Rights Reserved

Coverage



© 2014 HRL Laboratories, LLC All Rights Reserved

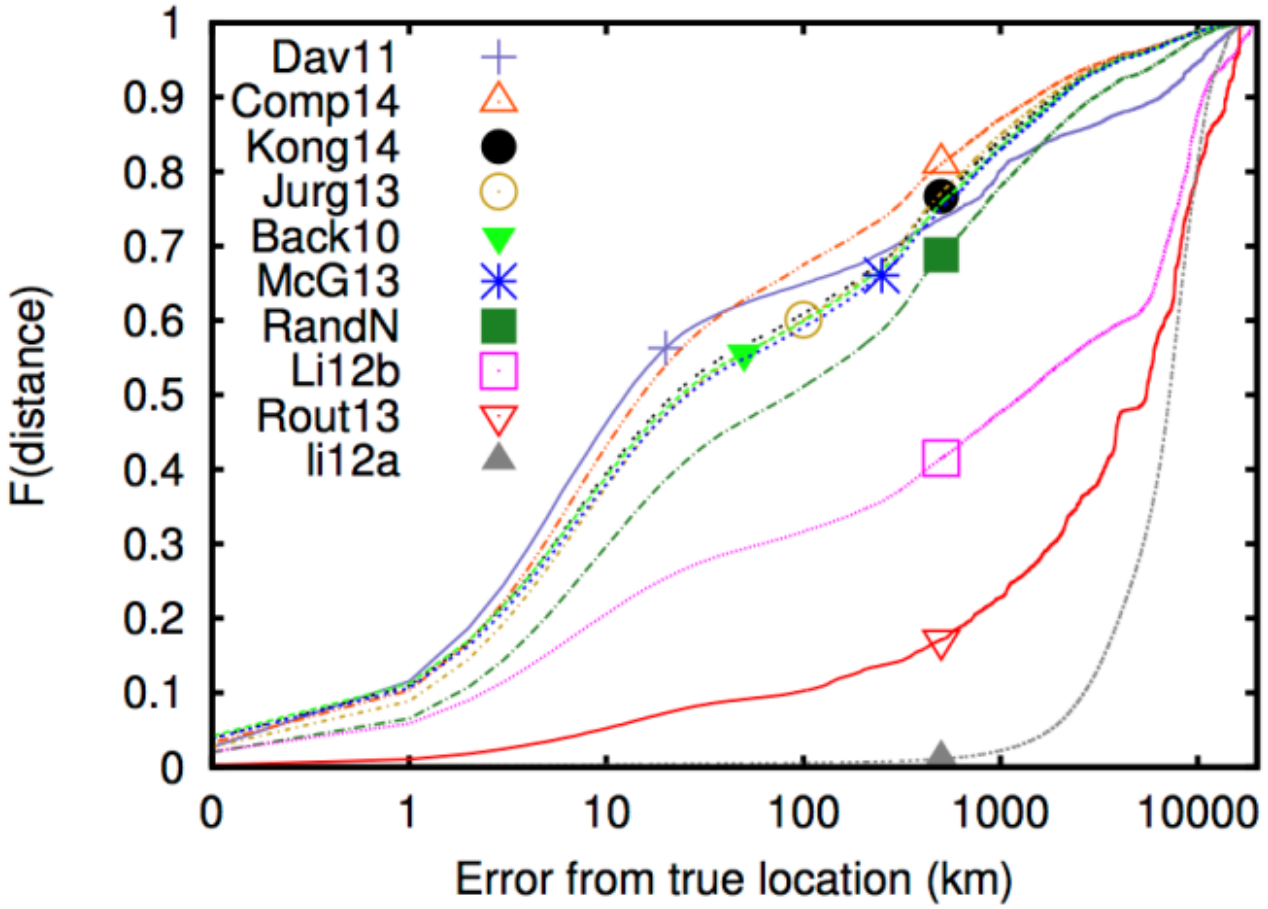
Coverage



© 2014 HRL Laboratories, LLC All Rights Reserved

Upcoming paper

"testing nine geolocation inference techniques, all published recently in top-tier conferences"



Jurgens et al. "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice" ICWSM 2015

© 2014 HRL Laboratories, LLC All Rights Reserved

Applications

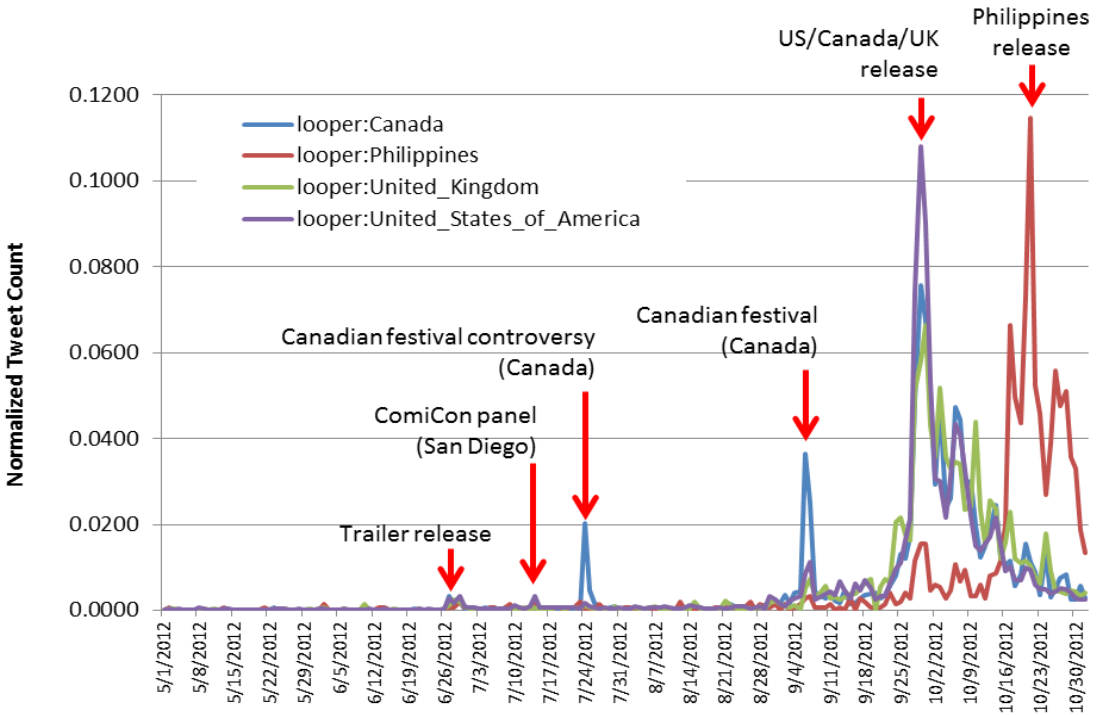
© 2014 HRL Laboratories, LLC All Rights Reserved

Movies

There is interest in using Twitter to forecast opening weekend box office revenue

When topic modeling is not an issue there are clear signals in Twitter data

Remark: We had better success using Wikipedia page view statistics



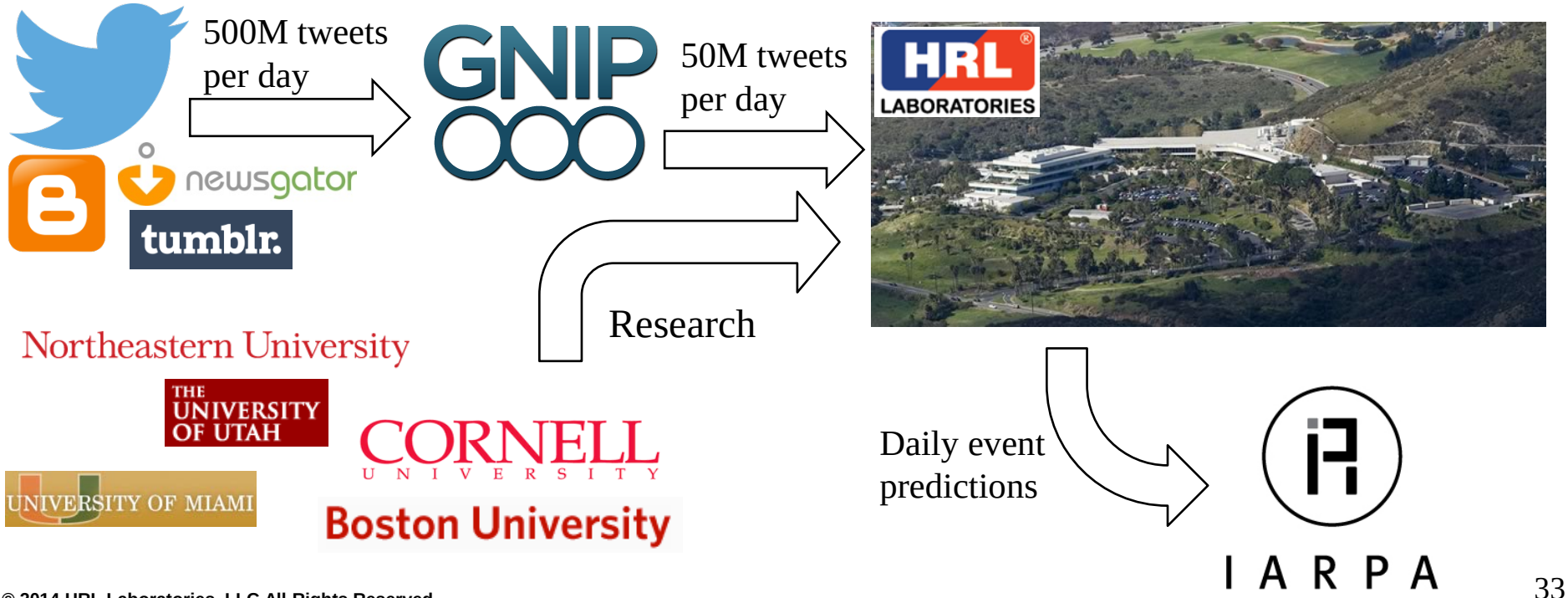
Geolocated tweets per day mentioning “Looper”.
(annotations by Laurent Giovangrandi)

de Silva, Brian, and Ryan Compton. "Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity." *arXiv preprint arXiv:1405.5924* (2014).

© 2014 HRL Laboratories, LLC All Rights Reserved

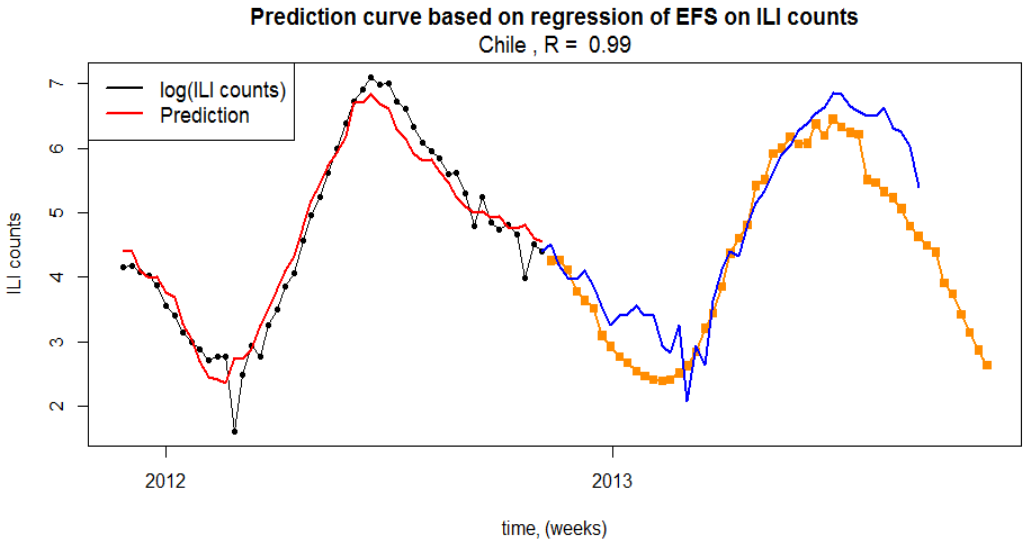
Introduction

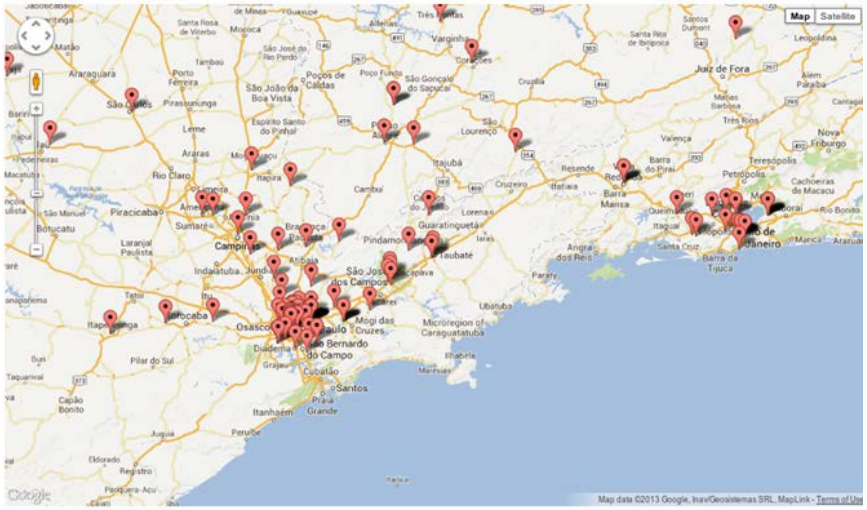
- Since April 2012 HRL has been collecting a 10% of publicly-visible Twitter data via a commercial feed from GNIP
- Used for IARPA-OSI program. Goal: Generate (in real time) predictions for:
  - disease outbreaks
  - elections
  - strikes/protests
  - financial catastrophes



© 2014 HRL Laboratories, LLC All Rights Reserved

Trend Forecasting vs. Discrete Event Prediction





Twitter:

Compton et al. “Detecting future social unrest in unprocessed twitter data” IEEE-ISI, 2013 (best paper nomination)
Compton et al. “Using publicly-visible social media to build detailed forecasts of civil unrest” Springer Security Informatics, 2014 (invited publication)
Sofia Apreleva, Craig Lee, Tsai-Ching Lu “Robust tracking of morbidity trends using amplified signals extracted from Google Trends and Twitter” (to be submitted)

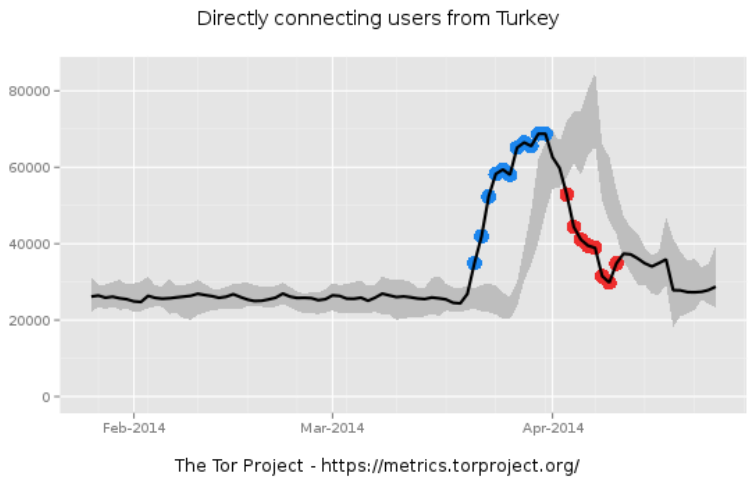
© 2014 HRL Laboratories, LLC All Rights Reserved

Censorship Monitoring

Turkey:
Censorship of Twitter March 20 – April 3

Is it visible in public data?

http://www.bbc.com/news/world-europe-26677134



BBC News Sport Weather Capital Future Shop

NEWS EUROPE

Home | US & Canada | Latin America | UK | Africa | Asia | Europe | Mid-East | Business | Health | Sci/Environ

20 March 2014 Last updated at 20:10 ET

Share

Twitter website 'blocked' in Turkey

The BBC's James Reynolds tries to access the site

Twitter users in Turkey report that the social media site has been blocked in the country.

Some users trying to open the twitter.com website are apparently being redirected to a statement by Turkey's telecommunications regulator.

It cites a court order to apply "protection measures" on the website.

This comes after PM Recep Tayyip Erdogan vowed to "wipe out Twitter" following damaging allegations of corruption in his inner circle.

The BBC's James Reynolds in Istanbul reports that he is unable to access Twitter.

"I don't care what the international community says at all. Everyone will see the power of the Turkish Republic," Mr Erdogan said earlier on Thursday.

He spoke after some users had posted documents reportedly showing evidence of corruption relating to the prime minister - a claim he denies.

His office said that Twitter had not responded to Turkey's court rulings to remove some links, forcing Ankara to act.

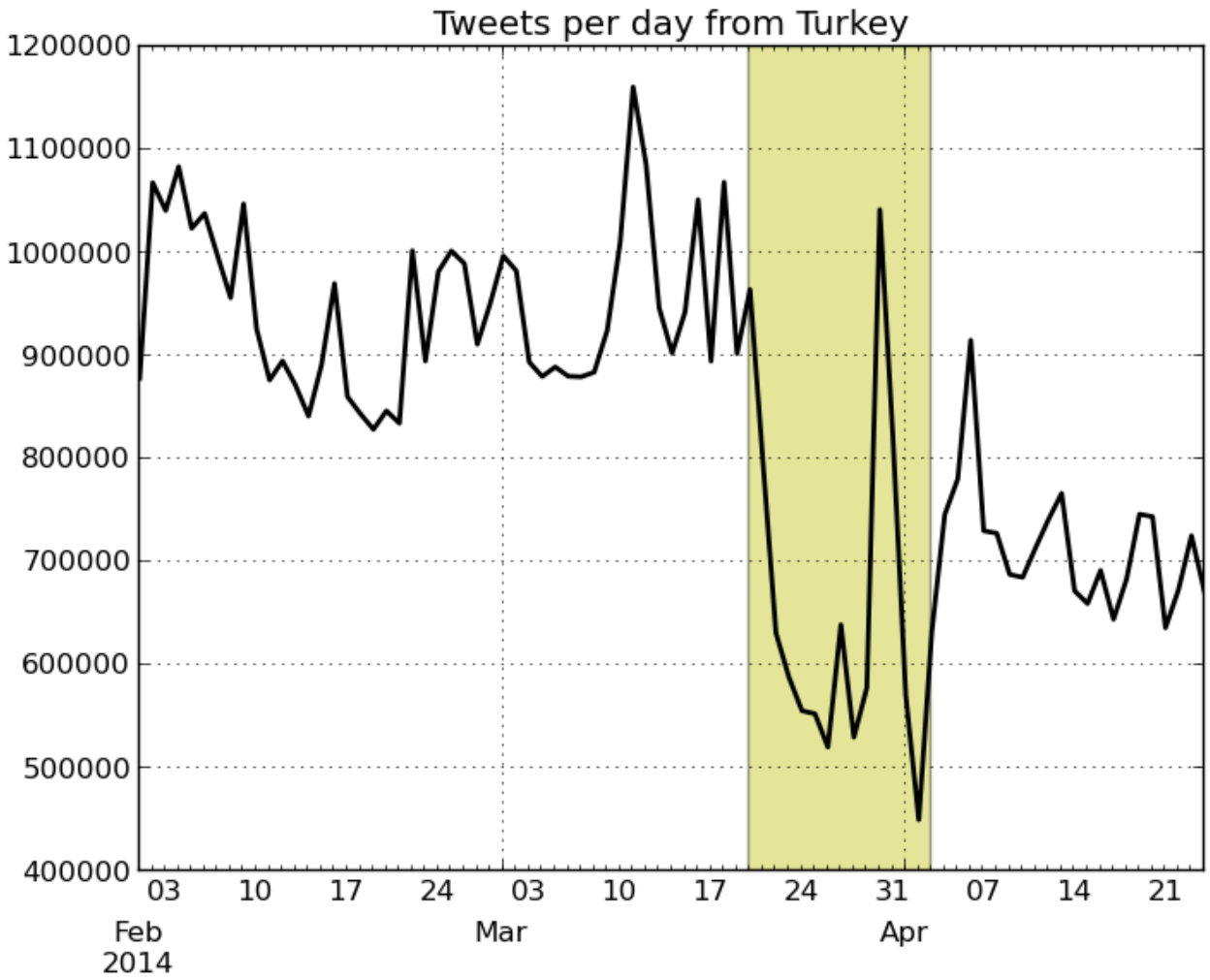
Twitter has so far made no public comment on the issue.

Related Stories

- Clashes at Istanbul internet protest
- Turkey tightens control of internet
- Half a million 'internet censorship' tweets

© 2014 HRL Laboratories, LLC All Rights Reserved

Censorship Monitoring



© 2014 HRL Laboratories, LLC All Rights Reserved

Geotagging Other Digital Media via Twitter Sharing Locations

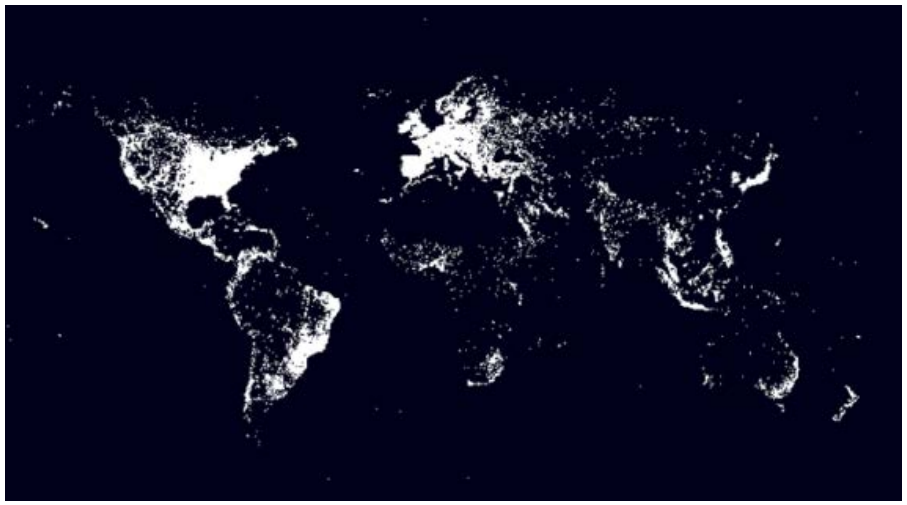
	Test points	Median error (km)	Mean error (km)
YouTube	5022	22.80	1001.58
Flickr	42	371.88	2475.04
GDELT	1580	304.74	2432.81
Manual news	1115	36.66	902.01

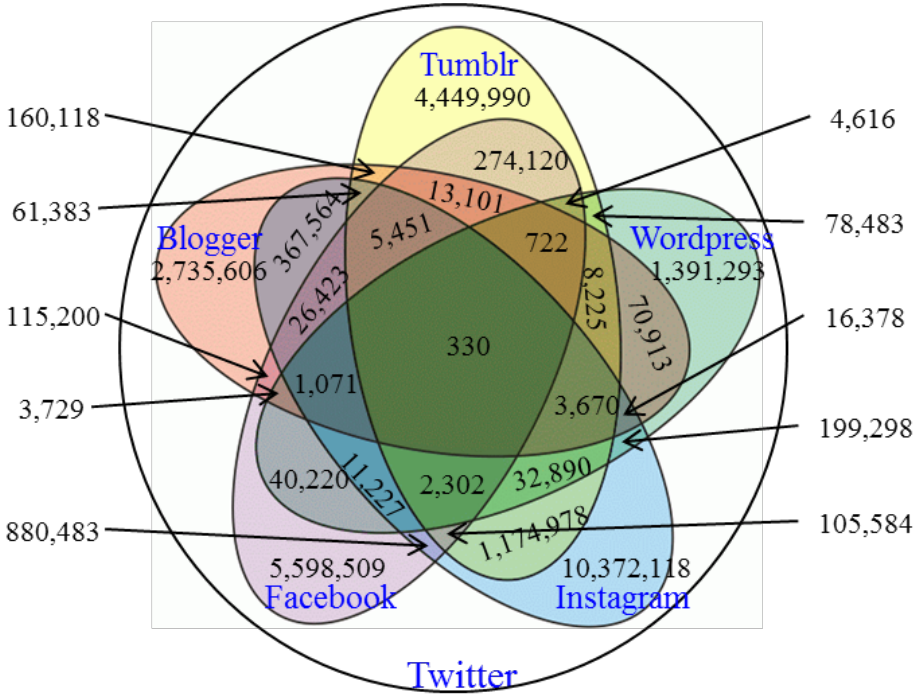
Summary of discrepancy (in km), with a 100km restriction on dispersion, between the median of the locations of the users who share the link and API-based geotagging

Compton, R., Keegan, M. S., & Xu, J. (2014). Inferring the geographic focus of online documents from social media sharing patterns. *arXiv:1406.2392*.

© 2014 HRL Laboratories, LLC All Rights Reserved

Geotagging Other Digital Media via User Alignment





Geospatial distribution of Tumblr-Twitter aligned accounts

Xu, J., Lu, T. C., Compton, R., & Allen, D. (2014, June). Quantifying cross-platform engagement through large-scale user alignment. In *Proceedings of the 2014 ACM conference on Web science* (pp. 281-282). ACM.

Xu, J., Compton, R., Lu, T. C., & Allen, D. (2014, June). Rolling through tumblr: Characterizing behavioral patterns of the microblogging platform. In *Proceedings of the 2014 ACM conference on Web science* (pp. 13-22). ACM.

© 2014 HRL Laboratories, LLC All Rights Reserved

How to opt out?

Some users may want to communicate publicly – but only to humans

- **Private @mentions don't exist**
- **www.captchatweet.com does exist, but has limited functionality**
- **Suggestions?**

© 2014 HRL Laboratories, LLC All Rights Reserved