Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, Derek Ruths

jurgens@cs.mcgill.ca, {tyler.finethy,james.mccorriston,yi.t.xu}@mail.mcgill.ca, derek.ruths@mcgill.ca

network dynamics
@mcgill

Introduction and Motivation

Geolocated social media data provides a powerful source of information about place and regional human behavior. Because little social media data is geolocation-annotated, geoinference techniques serve an essential role for increasing the volume of annotated data by predicting its origin location. One major class of inference approaches has relied on the social network of Twitter, where the locations of a user's friends serve as evidence for that user's location. While many such inference techniques have been recently proposed, we actually know little about their relative performance, with methods differing in the evaluation metrics, testing setups, and amount of data. We conduct a critical evaluation of state of the art by testing nine geolocation inference techniques on identical data using three newly-proposed comprehensive evaluation metrics.

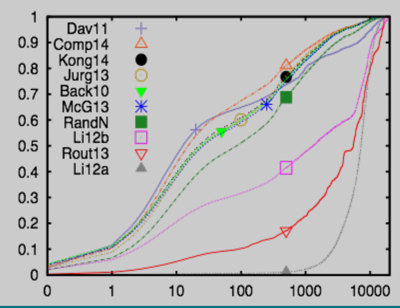
Method	Users	Edges	Ground Truth	% Labeled
Davis Jr. et al (2011)	4.7K	n/a	GPS, GeoIP, Self-reported	40.3%
Li et al. (2012)	139.1K	4.1M	Self-Reported	100%
Li, Wang, and Chang (2012)	139.1K	4.1M	Self-Reported	100%
Rout et al (2013)	206.2K	9.8M	Self-Reported	100%
McGee, Caverlee, and Cheng (2013)	249.6K	81.2M	GPS, Self-Reported	100%
Kong, Liu, and Huang (2014)	660.0K	19.4M	GPS	22.5%
Backstrom, Sun, and Marlow (2010)	2.9M	30.6M	Self-Reported	25.0%
Jurgens (2013)	47.8M	254M	GPS	5.34%
Compton, Jurgens, and Allen (2014)	110.9M	1.03B	GPS	11.1%

The nine evaluated methods and the conditions in which each method was originally tested

Evaluation Metrics

Area Under the Curve (AUC)

Many methods have been evaluated using a Cumulative Distribution Function (CDF) that shows the probability of the distance error when inferring the location of a post. While visually illustrating, these curves are not comparable across works. Therefore, we propose using a form of AUC calculated from the CDF to quantify prediction performance.



Median-Max

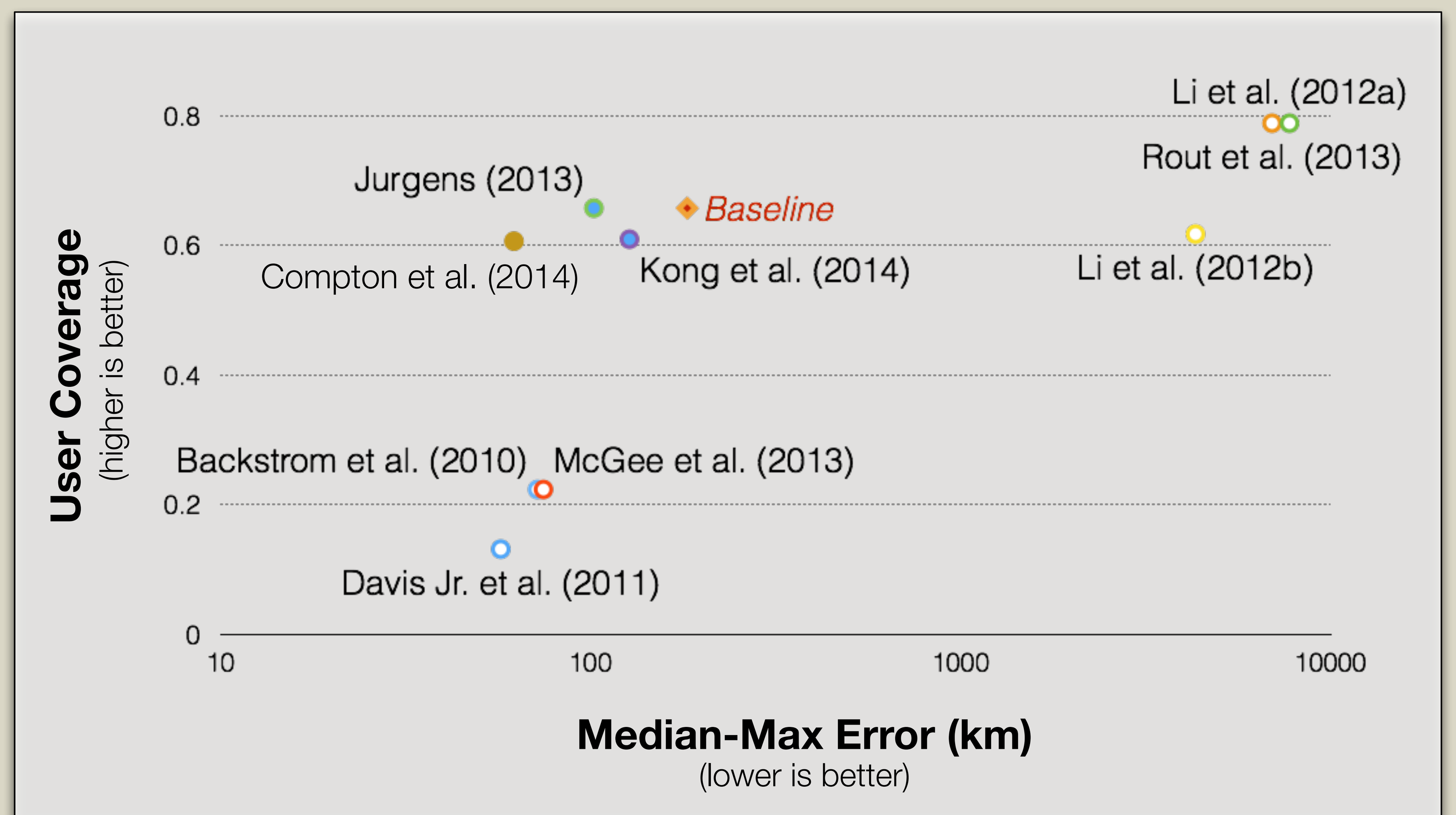
Some analyses rely on having *users* located, rather than posts. To quantify how accurate a method is at the user-level, we compute the maximum post-prediction error per user and report the median of these errors. This metric has an intuitive interpretation: half of the users have a maximum error of at most this distance

Coverage

Methods vary in how much data for which they are able to make a prediction. Therefore, we include a third metric, Coverage, that measures the percentage of data able to be located by the geoinference method, where data may be users or posts.

Q1: How do the methods compare?

Setup - All methods were tested using five-fold cross validation on a dataset built from a one-month sample of Twitter (15.2M users, 26M edges). The baseline comparison method infers locations by simply picking a random neighbor's location to use a user's location. Below we show results when the ground-truth is derived from GPS-annotated data.



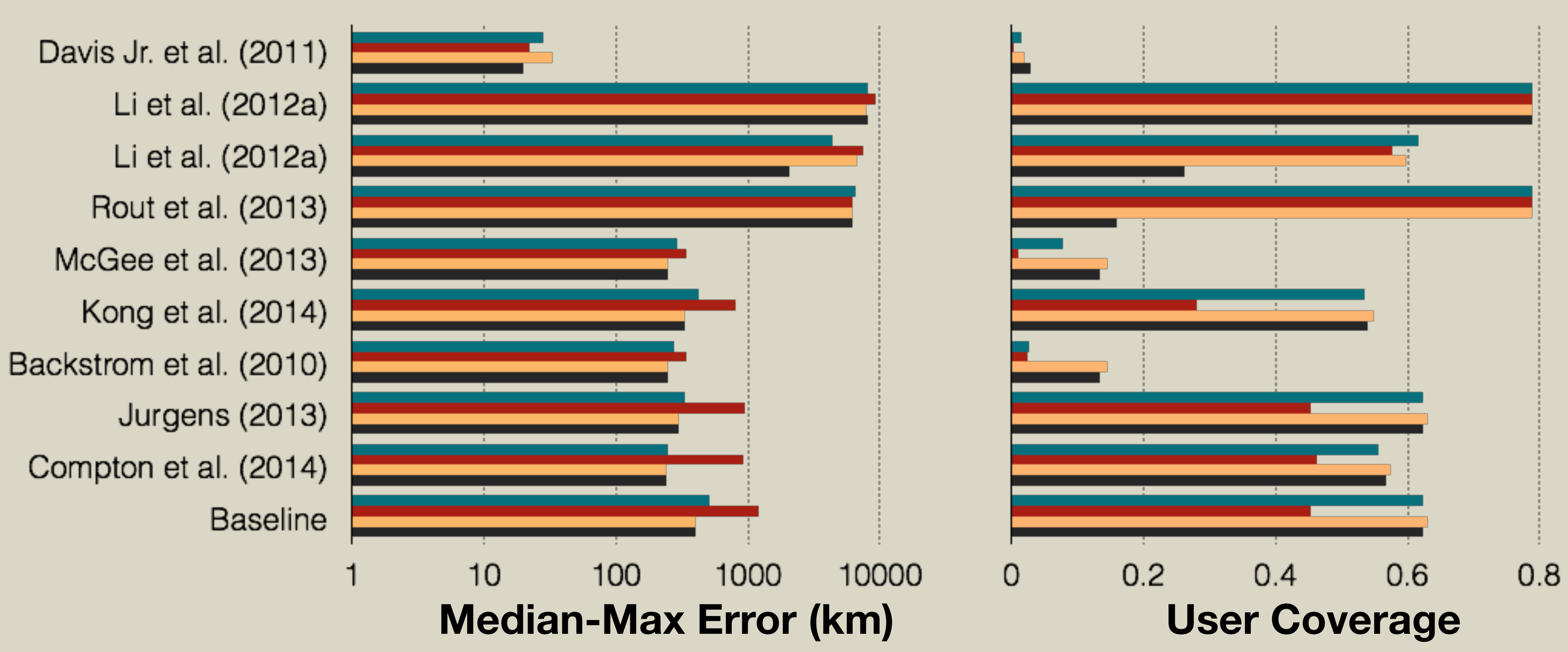
Incorporating multiple passes through the data can provide significantly higher coverage without much loss in precision

The best performance was seen for methods originally tested in conditions that mirrored real-world. Four methods tested only on smaller data had to be modified for scalability.

Six methods were able to outperform the baseline in prediction accuracy. However, the inclusion of coverage demonstrates significant differences in the methods' abilities to label content.

Q2: Are self-reported locations more beneficial than GPS locations?

Setup - The text from users' profile location fields were extracted and matched with the location names in one of four gazetteers: (1) GeoNames, (2) DBPedia, (3) GeoLite, and (4) a gazetteer built from queries to Google's reverse geocoder service. The methods were then tested using the same cross-validation setup as when using GPS-derived locations.



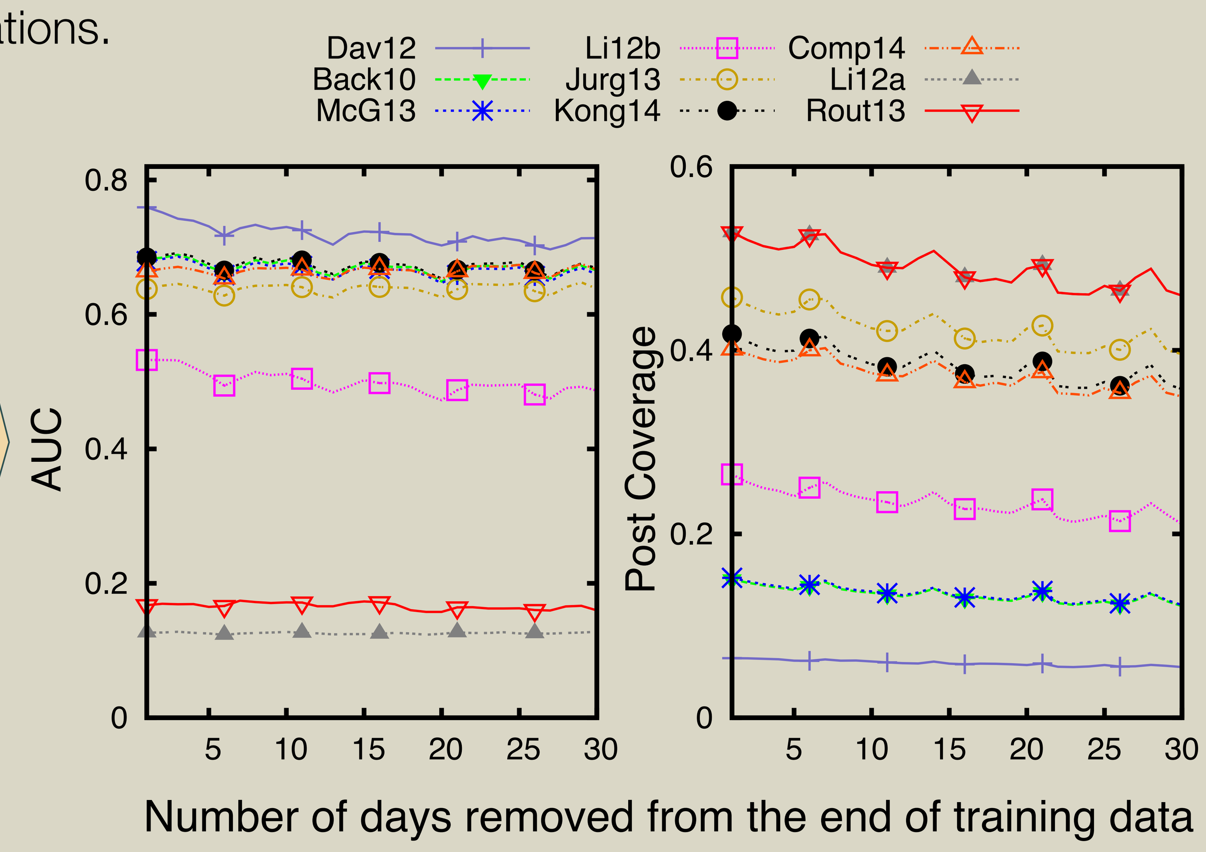
Self-reported locations resulted in universally-worse performance for all gazetteers when used as ground truth data instead of GPS-derived locations, even though they provide roughly 50% more ground truth from which methods can learn.

Users' location fields matched far fewer gazetteer names (3.4%) than reported in prior work. This lower rate may be due to our study's analysis of global users who write in a variety of languages or due to shifting user behaviors from increased privacy concerns.

Q3: How stable is performance over time?

Setup - All methods were trained on a full month of data and then asked to predict the locations of posts for each day in the following month. Results are shown when using GPS-derived locations.

Prediction accuracy remained mostly stable when inferring locations for data in the future



Post coverage drops 15-26% after only one month!

Get the code: https://github.com/networkdynamics/geoinference

Want to see how well your method does? **Try out our platform and API for geoinference on the *same datasets* in this paper: http://networkdynamics.org/resources/geoinference**

For more details, see our paper in the ICWSM Workshop on Social Media Standards and Practices "FREESR: a Framework for Reproducible Evaluation of Experiments with Sensitive Resources"!

References

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW*, 61–70. ACM.

Davis Jr, C.; Pappa, G.; de Oliveira, D.; and de L Arcanjo, F. 2011. Inferring the location of twitter messages based on user relationships. *Transactions in GIS* 15(6):735–751.

Compton, R.; Jurgens, D.; and Allen, D. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *Proeceedings of the IEEE International Conference on Big Data*.

Jurgens, D. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of ICWSM*.

Kong, L.; Liu, Z.; and Huang, Y. 2014. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment* 7(13).

Li, R.; Wang, S.; and Chang, K. C.-C. 2012. Multiple location profiling for users and relationships from social net- work and content. *Proceedings of the VLDB Endowment* 5(11):1603–1614.

Li, R.; Wang, S.; Deng, H.; Wang, R.; and Chang, K. C.- C. 2012. Towards social user profiling: unified and dis criminative influence model for inferring home locations. In *Proceedings of KDD*, 1023–1031. ACM.

McGee, J.; Caverlee, J. A.; and Cheng, Z. 2013. Location prediction in social media based on tie strength. In *Proceedings of CIKM*, 459–468. ACM.

Rout, D.; Bontcheva, K.; Preotiuc-Pietro, D.; and Cohn, T. 2013. Where's @Wally?: a Classification Approach to Geolocating Users Based on Their Social Ties. In *Proceedings of HT.* ACM