Proceedings of the Ninth International AAAI Conference on Web and Social Media

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, derek.ruths@mcgill.ca,
{tyler.finethy,james.mccorriston,yi.t.xu}@mail.mcgill.ca
School of Computer Science
McGill University

Abstract

Geolocated social media data provides a powerful source of information about place and regional human behavior. Because little social media data is geolocation-annotated, inference techniques serve an essential role for increasing the volume of annotated data. 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 the amount of ground truth data varying between 5% and 100% of the network, the size of the social network varying by four orders of magnitude, and little standardization in evaluation metrics. We conduct a systematic comparative analysis of nine state-of-the-art network-based methods for performing geolocation inference at the global scale, controlling for the source of ground truth data, dataset size, and temporal recency in test data. Furthermore, we identify a comprehensive set of evaluation metrics that clarify performance differences. Our analysis identifies a large performance disparity between that reported in the literature and that seen in real-world conditions. To aid reproducibility and future comparison, all implementations have been released in an open source geoinference package.

1 Introduction

Social media provides a continuously-updated stream of data which has proven useful for predicting group behaviors and modeling populations. Associating data with the particular geolocation from which it originated creates a powerful tool for modeling geographic phenomena, such as tracking the flu, predicting elections, or observing linguistic differences between groups. However, only a small amount of social media data comes with location; for example, less than one percent of Twitter posts are associated with a geolocation. Therefore, recent work has focused on geoinference for predicting the locations of posts. One direction of geoinference using social networks has produced approaches that claim to accurately locate the majority of posts within tens of kilometers of their true locations (McGee, Caverlee, and Cheng 2013; Rout et al. 2013; Jurgens 2013; Compton, Jurgens, and Allen 2014).

Despite the significant interest in geoinference and reported success of methods, examining current evaluation practices reveals significant disparity between evaluation settings. Moreover, the conditions in which methods have been tested often do not mirror the real-world conditions in which these methods are expected to operate. In particular, data set sizes have varied by four orders of magnitude and no agreement has been seen in (1) the source of ground truth data, (2) what percentage of the dataset should already be labeled with ground truth locations when training, and (3) which evaluation metrics should be used. While these factors are not necessarily limitations of the algorithms, they critically hinder comparability and obtaining a complete understanding of how state of the art performs at its intended task.

Here, we conduct a critical evaluation of state of the art by testing nine geolocation inference techniques, all published recently in top-tier conferences. All methods were evaluated on the same prediction task: given a Twitter post from an arbitrary user, predict its geolocation as latitude and longitude. We outline three criteria for geoinference and then introduce corresponding evaluation metrics that can easily be reproduced across datasets, ensuring methodological comparability in future works. In three experiments, we systematically assess (1) the accuracy of each algorithm in identical, real-world conditions, (2) the impact of how ground truth is determined for training, and (3) the relationship between performance and time as models infer the locations of posts created after the model was trained.

Our study makes three key contributions. First, we demonstrate that real-world performance is often much less accurate than reported in original experiments; and in particular, we find that several algorithms did not scale to real-world conditions and even when algorithmic improvements are made, did not produce accurate location inferences. Second, we find that the standard practice of basing ground truth on users' self-reported locations consistently produces inferior results: The established technique for extracting users' self-reported locations identified significantly fewer user locations than reported in earlier studies –possibly suggesting a shifting user behavior– and results in significantly worse performance than if more-sparse GPS-based locations are used as ground truth. Third, we show that the number of posts able to be tagged by a trained model decreases by

half every four months due to the platform's changing user base, which given the complexity of some methods, raises the issue of how to effectively retrain methods on increasing amounts of data. As a part of the study, we release an open source geoinference framework that includes implementations of all of the nine methods and proposed evaluation metrics, lowering the bar for future comparisons.

2 Geolocation Inference Methods

Social network-based geoinference is built upon the finding that relationships in social media are strong indicators of spatial proximity (Wellman and Gulia 1999; van Meeteren, Poorthuis, and Dugundji 2009). The spatial proximity of friends is not limited to one social media platform or one definition of friendship, with works showing that the distances of friends' locations follow a power-law like distribution for Facebook friends (Backstrom, Sun, and Marlow 2010), FourSquare friends (Jurgens 2013), Twitter followers (Gonzalez et al. 2011), and Twitter mentions (McGee, Caverlee, and Cheng 2011).

Geolocation inference algorithms using Twitter require two common parameters: (1) a definition of what constitutes a relationship in Twitter to create the social network, and (2) a source of ground truth location data to use in inference. The nine geoinference methods analyzed here have varied considerably in both parameters, as shown in Table 1, columns two and five. Despite GPS-annotated posts being rare in Twitter ($<$1% frequency) (Hecht et al. 2011), the social networks used to test many methods have been constructed from only those users with known locations, which is not a representative sample of the users seen in real-world conditions. Similarly, methods have been tested on significantly different sizes of data, without discussions of scalability in cases where a method is tested on relatively small numbers of users or posts. Following, we discuss the algorithms, how these different sources of information are used, and their complexities.

Backstrom, Sun, and Marlow (2010) propose one of the first geolocation inference methods using the social network of Facebook. The ground truth is established using users' self-reported addresses in Facebook, which were required to include street numbers and therefore indicate highly-precise locations. Inference begins by building a probabilistic model representing the likelihood of observing a relationship between two users given the geographic distance between them. Then, for each user, the known locations of the user's friends are each tested to identify the location that maximizes the likelihood of it being the user's location given the geographic distribution of the friends' locations. Backstrom, Sun, and Marlow (2010) propose an efficient method for computing the likelihoods, resulting in an overall complexity of $O(|V|k^2)$ where V is the number of vertices (users) in the graph and k is the average number of vertex neighbors with known locations.

McGee, Caverlee, and Cheng (2013) extend the method of Backstrom, Sun, and Marlow (2010), citing the difficulties of adapting from Facebook to Twitter: (1) user-provided data is significantly less precise in Twitter, (2) the geographic scale of the study moves from only within the United States to a global scope, and (3) social relationships in Twitter serve multiple roles, beyond signifying friendships. As such, McGee et al. seek to classify a user's Twitter relationships according to the probability that they serve as strong predictors of that user's location. For ground truth, users with at least three GPS posts are selected and the median latitude and longitude values are selected as their location.

All relationships between located users are used to train a tree regression model that predicts the actual distance between the users. These predicted distances are treated as measurements of the predictive capabilities; a user's social relations are then sorted by capability into ten partitions to identify the friends that are potentially most informative. For inference, the edges in each partition are processed using a similar method as in the Backstrom et al. model to predict the likelihood of a user having an edge to a friend at that location. The intuition is that by separating relationships into different partitions, the location information in the partition with the most-predictive relationships will dominate the likelihood calculation. In the absence of relationship partitions, this method reduces to Backstrom, Sun, and Marlow (2010).

Originally, this method was tested with all users labeled with locations (cf. Table 1, col. 6), in which case the complexity grows beyond that of Backstrom et al. to $O(|V|ck^2 + k|V|\log|V|)$ where c is the number of partition for the tree regression and the latter term in the complexity is the cost of the regression. However, as the size of the network grows, computing the regression model on all pairs of located user dominates the run-time making the algorithm infeasible in practice. Therefore, we train the regression using data from a fixed number of users (50K) rather than $|V|$, which yielded similar results in practice and allows the method to scale to the size of networks used in this study.

Kong, Liu, and Huang (2014) propose several extensions to the Backstrom, Sun, and Marlow (2010) model based on strategies for weighting which of a user's friends are likely to be most predictive of their location. This method uses the friendship definition of Huberman, Romero, and Wu (2009) which is that a user a is friends with b if a has mentioned b in at least two posts. Given a user, their friends are weighted according to a social tightness coefficient, which is computed as the cosine similarity of the two user's friends.

Kong, Liu, and Huang (2014) diverges significantly from the Backstrom, Sun, and Marlow (2010) model and McGee, Caverlee, and Cheng (2013) extension by making the algorithm operate in multiple passes. Initially, the network is only sparsely labeled, preventing accurate inference on users with few labeled neighbors. Therefore, multiple passes are made through the network, using the previous pass's inferred locations as ground truth to infer the locations of other users. This multi-pass approach allows their method to identify significantly more users, though potentially at a cost of precision. The use of multiple passes and use of social closeness raises the complexity of this extension to $O(p|V|k^3)$ where p is the number of passes performed.

Method	Type	Users	Edges	Ground Truth	% Labeled	Multi-pass	Output
Davis Jr et al. (2011)	Bi-directional followers	24.7K	-[a]	GPS, GeoIP, Self-reported	40.3%	no	City Name
Li et al. (2012)	Friends, Followers	139.1K	4.1M	Self-Reported	100%	no	City Name
Li, Wang, and Chang (2012)	Friends, Followers	139.1K	4.1M	Self-Reported	100%	no	City Name
Rout et al. (2013)	Friends, Followers	206.2K	9.8M	Self-Reported	100%	no	City Name
McGee, Caverlee, and Cheng (2013)	Friends, Followers	249.6K	81.2M	GPS, Self-Reported	100%	no	GPS
Kong, Liu, and Huang (2014)	Mentions	660.0K	19.4M	GPS	22.5%	yes	GPS
Backstrom, Sun, and Marlow (2010)	Facebook friendships	2.9M	30.6M	Self-Reported[b]	25.0%[c]	no	GPS
Jurgens (2013)	Bi-directional mentions	47.8M	254M	GPS	5.34%	yes	GPS
Compton, Jurgens, and Allen (2014)	Bi-directional mentions	110.9M	1.03B	GPS	11.1%	yes	GPS

[a] Statistics on the number of edges per user were not provided.
[b] Backstrom, Sun, and Marlow (2010) use self-reported numbered street addresses from Facebook which are more precise than self-reported city-level locations found in Twitter.
[c] The network is fully labeled at the start but experiments are performed with 75% of locations removed.

Table 1: The nine evaluated network-based geoinference methods and their original testing conditions.

Li et al. (2012) (denoted `Li12a`) propose to infer geolocation by taking into account the influence of both users and of locations, capturing the intuition that some users are more informative for predicting the locations of the neighbors. The best-performing of their approaches first assigns users to random locations and then iteratively updates the locations of users from their neighbors and mentioned location names, refining the parameters in the update by measuring the prediction error of users with already-known locations. Each iteration of the algorithm requires $O(|E|)$ operations, giving a total complexity of $O(t|E|)$ for t total iterations. Li et al. provide a proof of convergence for the entire algorithm. However, our datasets are significantly larger than those used to test the method, and, as a result, we found it necessary to restrict the number of iterations to a maximum of five, allowing full completion of each updating step per iteration.

Li, Wang, and Chang (2012) (denoted `Li12b`) recognize that a user may relocate over the course of their social media history and therefore may have more than one home location from which they posts. To uncover all of a user's home locations, relationships between users and locations are modeled with a supervised extension to Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). A graph is constructed from both the social relationships in Twitter and the location names mentioned in a user's posts, with the generative model predicting the causes of edges in the graph. The model is inferred from the statistics of observed edges between users with known locations. Like other network-based approaches (McGee, Caverlee, and Cheng 2013; Kong, Liu, and Huang 2014), the Li, Wang, and Chang (2012) approach attempts to separate out the location-predicting relationships from those serving other functions in Twitter by using the topics learned in the model to specify relationship type.

Given the probability distribution over locations for a user, this method selects the user's home location as the most probable. This home location is reported as the predicted location of each of that user's tweets. The computational complexity of the model is dependent upon the complexity of the topic model; however, in practice we found that the model converged quickly but the overall runtime of the method was dominated by the task of searching for location names within a user's post.

Rout et al. (2013) approach geolocation inference as a classification task for assigning a user to one 4,295 cities in the United Kingdom (UK) using a follower-based social network. In their best-performing configuration, a SVM classifier was trained with three core features: (1) the cities of a user's friends, relative to the city's respective number of Twitter users, (2) the number of closed triads in a user's social network residing in the same city, and (3) the number of reciprocal following relationships a user has per city. The method was originally evaluated only for users in the UK; owing to the demographics of the UK, over 28% of the users in the resulting evaluation dataset were located in a single city (London).

The choice of an SVM classifier ultimately constrains the method's scalability. Originally, the algorithm used radial basis function (RBF) kernel for the SVM, which as the authors note, significantly increases in training time as more training instances are added. Further compounding the computational cost is the number of unique city labels. Multi-class SVMs are typically trained by constructing n one-versus-rest binary classifiers, where n is the number of classes. In the global setting with hundreds of thousands of possible cities, training a multi-class SVM on all possible locations becomes computationally infeasible. To scale the algorithm two changes were made (1) the RBF kernel was replaced with a linear kernel, lowering training time to $O(|V|^2)$ and (2) the number of unique locations was restricted to the 5,000 locations most frequently seen in the training data. With these improvements to scalability, the method becomes $O(|L||V|^2)$ where L is the restricted set of user locations.

Davis Jr et al. (2011) propose one of the simplest approaches where given a user, their location is inferred by taking the most-frequently seen location among their social network. Locations are defined in terms of cities and are assigned to users on the basis of GPS or GeoIP data, when possible, or the self-reported location field of a user's profile. The method was applied to a small set of users mentioning certain terms on Twitter and restricted to only those users in Brazil, which entailed using a Brazilian gazetteer for location name matching. Although the algorithm is defined on a

single-user basis, if run on an entire network, the method is equivalent to a single round of label propagation (Zhu and Ghahramani 2002) where a user is assigned the location label of its neighbors, subject to addition criteria.

Jurgens (2013) extends the idea of location inference as label propagation by interpreting location labels spatially. Locations are inferred using an iterative, multi-pass procedure. Like Kong, Liu, and Huang (2014), during one pass, the location of each user is assigned as the geometeric median of the locations of all the user's neighbors, with ground-truth data providing the locations for the first pass. The use of multiple passes allows the inferred user locations (rather than ground truth) to be used when making new inferences, potentially overcoming the problem of making inferences when ground truth data is sparse. Calculating the geometric median from the locations of a user's friends requires k^2 operations, yielding an overall complexity of $O(p|V|k^2)$.

Compton, Jurgens, and Allen (2014) extend the method of Jurgens (2013) to take into account edge weights in the social network and to limit the propagation of noisy locations. Rather than calculate each user's location as the geometric median of their friends' locations, the Compton et al. method weights locations as a function of how many times users interacted, thereby favoring locations of friends with whom there exists a stronger evidence of a close relationship. Furthermore, during each iteration's inference, a user's location may only be updated if the new location is not too distant from the locations of the neighbors, thereby preventing the propagation of erroneous location information through the network. However, these algorithmic extensions do not change the complexity of the original algorithm, making the approach still $O(p|V|k^2)$.

3 Evaluation

Prior evaluations of geoinference methods have had little standardization in the metrics used to measure performance, which significantly reduces comparability between results. Furthermore, the existing metrics often test for different capabilities, so when only one metric is reported, an incomplete view is given of a method's performance. Therefore, we propose three essential evaluation criteria by which methods should be measured: (1) accurate inference of a post's location, (2) accurate inference of all the posts generated by a user, and (3) maximizing the number of posts whose location is able to be inferred. The first criteria provides insight into how accurate the method may be for an arbitrary post, which is in some way the most fundamental reason for performing geoinference. The second user-based criteria, while similar to the first, is necessary for two reasons. First, given that users post in different volumes, the value of any post-based metric is biased by the users who post most frequently. User-based metrics provide a way of separating volume from accuracy and can provide a reliable estimate of the expected error for an arbitrary user, which may not be evident from post-based metrics alone. As a result, post-based and user-based metrics act analogously to micro- and macro-averaged precision. Second, many downstream algorithms that require geolocated data operate on users rather than tweets (e.g., flu trends, election prediction) and therefore the accuracy with which a given user can be located quantifies the scale at which such user-based studies can be performed. Finally, the third criteria is necessary for distinguishing approaches on the basis of how much data may be located, which allows observing a method's trade-off between accuracy and coverage. Following, we summarize the current types of evaluation metrics and then propose three complementary metrics for evaluating methods by post and by user.

Prior Metrics

Five main metrics have been proposed for evaluating geolocation inference methods both at the tweet level and user level. The first approach evaluates geoinference methods that predict a city or produce a ranking of cities according to each being the city from which the tweet originates (Davis Jr et al. 2011; Rout et al. 2013). Inferences are measured using Precision@k which reports the percentage of location inferences in which correct location is within the k highest-ranked locations of the prediction; when k=1, the metric is equivalent to the traditional definition of Precision. While the metric has a clear interpretation, Precision@k lacks any notion of how distant the predicted location is from the actual location, so nearby answers are treated as equally incorrect as distant answers. This limits the metric's ability to measure the first criteria.

The second approach evaluates systems using Accuracy@k, which measures the percentage of predictions that are within k distance units (e.g., kilometers or miles) of the true location (Li et al. 2012; Li, Wang, and Chang 2012; McGee, Caverlee, and Cheng 2013; Rout et al. 2013). Accuracy@k has the advantage that it is applicable to both city- and GPS-reporting geoinference methods. However, the metric is sensitive to the choice in k: low k values can make the precision scores of methods identical, regardless of how far above k their predictions were, and similarly, large k values can create similar precision scores for two methods, even if one is significantly more accurate. As a result, many analyses report multiple values of k, making it difficult to summarize the overall performance of a system and compare approaches from analyses reporting different k. The need for multiple k values to capture performance makes Accuracy@k unsuitable for providing a single statistic to measure the post-based and user-based criteria.

A third approach reports the average error distance (Li et al. 2012; McGee, Caverlee, and Cheng 2013; Kong, Liu, and Huang 2014). The average error provides a useful single statistic for comparing approaches. However, the error distance when choosing the location of a friend follows a power law curve (Rout et al. 2013; Jurgens 2013), which does not have a well defined mean given the exponent of the error distribution (Newman 2005). Furthermore, the mean value is not robust to outliers in the error distribution (e.g., large errors from predicting the location of a traveling user) and can therefore overestimate the expected error.

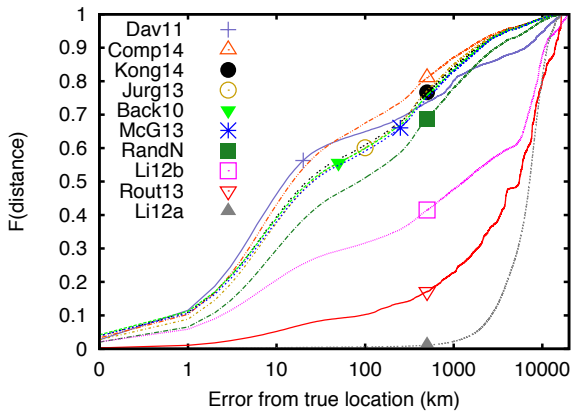

Figure 1: The CDF of location prediction errors for all nine methods, generated from five-fold cross-validation on one month of data.

A fourth approach reports performance using a cumulative distribution function (CDF) which shows the percentage of inferences having an error less than x kilometers away from the true location (Backstrom, Sun, and Marlow 2010; Rout et al. 2013; Jurgens 2013; Compton, Jurgens, and Allen 2014). Hence, the CDF represents Accuracy@k for all values of k and reports the full underlying error distribution of the method. While the CDF lends itself to easy visual comparison, the lack of a quantifiable statistic of performance limits its ability to compare approaches from different analyses and subsequent use as an evaluation metric for the proposed criteria.

Finally, a fifth recent evaluation approach considers the cases in which the predictions of geoinference method are probability distributions over a geographic area (Priedhorsky, Culotta, and Del Valle 2014). However, as no network-based method predicts probability distributions, we do not consider this type of evaluation further.

Proposed Metrics

We propose one metric for each of the three desired criteria identified earlier.

AUC The first metric measures post accuracy and extends the intuition behind Accuracy@k by computing a single statistic from the CDF of a method's error per post, which is equivalent in definition to a curve representing Accuracy for all values of k. For clarity, we define the metric in terms of a CDF $F(x)$ = P($distance \leq x$) where x is a distance, though we note that $P(x)$ and k are fully analogous to Accuracy and k. As shown in Figure 1, the area under the curve (AUC) of a CDF provides a statistic for quantifying a system's overall performance, where systems which make highly accurate predictions in the majority of cases (i.e., have high values of $F(x)$ for low values of x) maximize their AUC. Furthermore the AUC provides an easily comparable and reproducible value which removes the need to use visual comparison as done in prior works using CDFs for evaluation (Backstrom, Sun, and Marlow 2010; Rout et al. 2013; Jurgens 2013; Compton, Jurgens, and Allen 2014).

Two adjustments are made when calculating the value of the AUC. First, as predictions become further from the actual locations, the degree to which they are incorrect becomes less important in measuring the total accuracy of the system. For example, the difference in performance (as indicated in the metric's value) between system a with $F_a(1km) = 0.7$ and a system b with $F_b(100km) = 0.7$ should be larger than the difference between two systems c and d with $F_c(9000km) = 0.7$ and $F_d(9100km) = 0.7$, respectively, reflecting the intuition that the first system is significantly better than the second, while there is little practical difference between the third and fourth systems. Hence, the penalty for an inaccurate prediction should scale inversely with distance. Therefore, the CDF curve is computed from a log-scaled x value, $F(x)$ = P($distance \leq \log x$), in which case the majority of the AUC value comes from prediction accuracy for low values of x; this log scaling matches that used in prior work when visualizing CDFs for comparing geoinference methods. Second, because the maximum error for any prediction is bounded by the Earth's circumference, for ease of interpretation, we normalize the AUC to be in $[0, 1]$, in which case the area under the normalized curve has a maximal value of 1 if all of a method's predictions have no error and the area has a minimal value of 0 if all predictions are the furthest distance possible from the actual location. As a result, the normalized AUC provides a single measure for post-based accuracy for the first criterion.

Median-Max To quantify user-based accuracy for the second criteria, user-based error, we adopt a statistic that reflects the expected worst-case distance error per user. The highest error is identified from all predictions for a user's posts and then the median of these errors across all users is reported. This error value provides an intuitive way of characterizing the expected accuracy of a user's predictions: half of the users have a maximum error of at most this distance.[1] For simplicity, we refer to this metric is as Median-Max.

Post Coverage The third metric measures, Post Coverage, is defined as the percentage of tested posts for which a geoinference method can predict a location. Post Coverage is essential for measuring how much of the total posting volume a geoinference method is capable of tagging. Post Coverage also represents a challenging metric for network-based geoinference methods, which are only able to predict locations for users in their underlying social network and therefore may be unable to infer locations for frequently-posting users that do not have social relationships (i.e., are not in the network).

4 Experimental Setup

Geoinference systems were built from identical training data to control for (1) the size and type of social network and (2) the source of ground truth used to seed the network with

[1] Other user-based errors were considered, such as the median of the median user errors, which yielded identical system rankings. We opted for the Median-Max based on ease of interpretation.

user's initial locations. Following, we describe the data collection procedure and resulting dataset.

Data collection method A training dataset was collected from a 10% sample Twitter posts during August 2014. This resulted in 1.3B tweets from which we constructed a social network. Though many types of networks are possible for representing the users in these posts, bi-directional mentions were selected to form the social network due to (1) their known efficacy in predicting locality (McGee, Caverlee, and Cheng 2011; Jurgens 2013; Compton, Jurgens, and Allen 2014), which makes them functionally equivalent to the follower networks used in some methods prior evaluations, and (2) being easier to obtain in constructing large social networks than when using following relationships, the collection of which is limited by Twitter API rate restrictions. The final social network used for training the models contained 15,238,513 users and 26,429,346 edges.[2] The choice was made to limit the training dataset to only one month due to practical reasons. While some methods have been tested on larger networks, our training network still contains three orders of magnitude more users than have been used in the evaluations of most methods (cf. Table 1) and represents a realistic amount of data with which a geoinference method would be expected to produce accurate results.[3]

Ground Truth Geoinference methods have used two sources of ground truth for training the models. The first source is the user's self-reported profile location, which has been reported to frequently contain city-level location names (Hecht et al. 2011). Cheng, Caverlee, and Lee (2010) propose the most widely-used heuristic of parsing a user's self-reported location by matching it with a "City, State" patterns. However, Cheng, Caverlee, and Lee (2010) and the several network-based methods using the heuristic have each evaluated their methods only in a single English-speaking country (either the United States or the United Kingdom), Therefore, to account for countries in which users would do not report their state or province when describing the city, the pattern was extended to match "City, Country." Furthermore, we also allow self-reported locations to match the pattern of "City" if the city name is unambiguous and occurs in only one country.

As an initial gazetteer, we adopt Geonames, a large geographical database that covers much of the world and is free to use under open source licenses.[4] Importantly, the Geonames database also includes (1) alternative names of the city, including a limited set of alternatives in different languages or the population's native script, and, (2) where possible, population sizes. These two piece of information enable creating a larger set of possible city expressions and, following Rout et al. (2013), resolving ambiguous city names to the most populous city. Combining all alternative names resulted in a set of 7,587,458 possible city names, with 2,810,211 ambiguous names resolved to a single location based on population size.

Our extension of Cheng, Caverlee, and Lee (2010) with the Geonames gazetteer identified locations for 515,653 users (3.4%); this percentage is significantly lower than the coverage reported in other works (Cheng, Caverlee, and Lee 2010; Jurgens 2013); we attribute this lower coverage to testing an increasingly-global user base who may report locations in multiple languages (Liu, Kliman-Silver, and Mislove 2014) and increased privacy concerns that may inhibit users from providing their location (Jeong and Coyle 2014).

As an alternate, second source of ground truth user location, we use GPS-annotated posts to infer a home location for users. Following prior work (Jurgens 2013; Compton, Jurgens, and Allen 2014; Kong, Liu, and Huang 2014), we require users to have at least five GPS-annotated posts within 15km of each other and the median absolute deviation of the distances in these posts is no greater than 30km, ensuring that home locations are only reported for mostly geo-stationary users. GPS-derived home locations were identified for 402,009 users (2.6%).

The choice in self-reported or GPS-derived locations highlights a trade-off in the source of ground truth: although GPS data potentially provides a highly-accurate picture of where a user was, GPS-based locations cover 20% less of the network than the potentially-noisier self-reported locations. Models were trained using each source of data in order to measure this effect on performance.

Baseline Fundamentally, social-network based geoinference methods vary in how they incorporate the location information of a user's immediate social network for deciding that user's location. Therefore, as a baseline for comparison, we consider a system called *Random Neighbor* that performs a similar operation but removes any knowledge from how neighbors' locations are used. The baseline (1) assigns a user's location from a randomly-selected already-located neighbor and (2) repeats this assignment process for all users for k iterations in order to propagate locations through the network. The value of k was set to four in order to allow locations to propagate through the network for most users.

5 Experiment 1: Cross-Validation

In prior evaluations of the methods considered here, each method was compared with only one other system, making it impossible to know how these different methods actually compare. Therefore, as a first experiment, we evaluate all nine methods and the baseline in a common setting, on identical data, using consistent sources of ground truth, and measuring performance with the proposed evaluation metrics.

Setup Methods were tested using five-fold cross validation: users with ground-truth locations are partitioned into

[2]For comparison, constructing a bi-directional follower-based network for the same set of users would take approximately 29 years to complete given Twitter's current API rate limits.

[3]All experiments were also performed using a range of larger training sets from the one-month size reported here to a much larger six-month size with a 44M user network and 1.5B tweets. All of these training datasets yielded similar results to those presented here (with only changes to metrics' values but not method ranks) and therefore, for ease of presentation and replicability, we report only results for the one-month August dataset.

[4]http://www.geonames.org/

five sets, with four used for training and one for testing. Accuracy for users in the test set is always measured according to deviation from the GPS locations of their tweets, irrespective of the users' ground truth; i.e., accuracy is never measured in terms of how close a prediction was to a self-reported location.

Because this experiment uses a subset of the available data, we report User Coverage instead of Tweet Coverage. User Coverage measures the percentage of users in the network for which a location was inferred. Because the methods are limited to predicting locations only for those users in the social network, User Coverage provides a way of assessing how much of the network it can use for prediction.

Results Quite surprisingly, cross-validation revealed significant performance difference between systems, with some performing well below the baseline. Table 2 shows the performance of the nine systems and baseline across the three metrics, with Figure 1 showing the CDF used to compute the corresponding AUC value when using GPS-based locations.

Four trends are notable. First, six of the methods outperform the baseline, differing only by 0.016 in AUC and 44.9km in their Median-Max error when using GPS-based ground truth. However, the six methods vary widely in terms of coverage. While the method of Davis Jr et al. (2011) has the lowest Median-Max error, it finds locations for only 13.1% of the network; in contrast, the methods of Compton, Jurgens, and Allen (2014) has a 4km increase in Median-Max error but identifies locations for an absolute increase in user coverage of 47.5%, amounting to locations for 6.96M more users. Indeed, the two other multi-pass methods of Kong, Liu, and Huang (2014) and Jurgens (2013) both attain higher coverage while still maintaining relatively high accuracy.

Second, the three metrics effectively capture the strengths and weaknesses of the models. For example, while Backstrom, Sun, and Marlow (2010) and Jurgens (2013) have similar AUC values, the former has a Median-Max error 30.2km lower, indicating the method provides tighter prediction error bounds per user; however, the latter method covers 43.4% more users (6.6M). Presenting all three statistics allows end-users to make an informed decision on which methods match their accuracy or volume requirements.

Third, the scale of testing conditions originally used with the methods was predictive of their performance in this real-world setting. The methods originally tested with fully-labeled networks or on small network sizes performed below the baseline in their geoinference ability (cf. Table 1). While this result does not negate their performances in the original experiments, our findings reiterate the need to match experimental conditions with those seen in the real world.

Fourth, the use of self-reported locations as ground truth degrades both accuracy and user coverage, compared with GPS. Despite having 30% more ground truth for individuals, these locations were not distributed throughout the network in such a way to increase the overall number of individuals for which a location could be inferred. The lone exception to this trend is Davis Jr et al. (2011), which does see an improvement in AUC and Median-Max, but has a dramatic drop in coverage and is only able to infer locations for 1% of all users. Multi-pass methods such as Kong, Liu, and Huang (2014) were less affected by the error from self-reported locations; because these methods may infer new user locations from previously-inferred locations, they were more robust to the arrangement of self-reported locations within the network and did not see as large of a decrease in user coverage.

	GPS			Geonames		
Method	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.
`Dav11`	0.633	**57.2km**	0.131	**0.725**	**28.3km**	0.014
`Li12a`	0.123	7738.9km	**0.788**	0.108	8028.0km	**0.788**
`Li12b`	0.386	4312.0km	0.617	0.217	6711.3km	0.591
`Rout13`	0.225	6942.7km	**0.788**	0.217	6520.8km	**0.788**
`McG13`	0.617	74.6km	0.223	0.492	293.0km	0.078
`Kong14`	0.596	127.4lkm	0.609	0.447	418.8km	0.534
`Back10`	0.622	71.9km	0.223	0.503	278.5km	0.026
`Jurg13`	0.620	102.1km	0.657	0.477	333.5km	0.623
`Comp14`	**0.656**	61.2km	0.606	0.509	250.6km	0.556
Rand. N.	0.559	182.5km	0.657	0.421	509.7km	0.623

Table 2: Performance during cross-validation of the training data, using either GPS-based or Geonames-based sources of ground truth.

6 Experiment 2: Self-reported Locations

Many geoinference methods have incorporated self-reported locations as ground truth. However, Experiment 1 revealed that performance using the standard approach to self-reported locations yielded much worse performance than GPS-based locations. This large decrease in performance raises the question of whether self-reported locations could still be used as a reliable source of information if a different gazetteer was used. No existing work has examined the impact of gazetteer choice on performance and therefore in Experiment 2, we test the performance impact of gazetteer choice using four large gazetteers: Geonames, GeoLite, DBpedia, and Google.

Setup All nine methods and the baseline were trained using an identical cross-validation setup from Experiment 1, only varying which gazetteer is used to map self-reported locations to coordinates according to the procedure described in Section 4. The Geonames gazetteer data remains unchanged from prior experiments and included for comparison. The GeoLite gazetteer[5] is a collection of GPS coordinate for 459,943 cities and is publicly available, licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License. All city names are reported in Latin characters and no population information is available for resolving ambiguous locations references. In cases of ambiguity, a self reported location is not matched to any city.

DBpedia is an structured form of Wikipedia that is queriable. DBpedia is a rich source of information, as it contains all the geographic coordinate information from Wikipedia as well as localizations of some cities in their native script, which can potentially match additional self-reported locations. The final list of cities was built by finding all DBpedia entries of type `Settlement` that had population, geographic coordinates, and were located in a country, which

[5] http://dev.maxmind.com/geoip/legacy/geolite/

Method	Geonames			DBpedia			GeoLite			Google		
	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.	AUC	Median-Max	User Cov.
Dav11	0.725	28.3km	0.014	0.741	22.7km	0.003	0.690	33.1km	0.019	0.712	20.2km	0.028
Li12a	0.108	8028.0km	0.788	0.094	9375.4km	0.788	0.110	7971.4km	0.788	0.545	8112.0km	0.788
Li12b	0.386	4312.0km	0.617	0.164	7378.3km	0.576	0.221	6720.7km	0.598	0.315	2069.8km	0.261
Rout13	0.217	6520.8km	0.788	0.234	6120.1km	0.788	0.234	6120.1km	0.788	0.231	6167.8km	0.158
McG13	0.492	293.0km	0.078	0.419	344.4km	0.010	0.512	249.4km	0.144	0.520	199.7km	0.133
Kong14	0.447	418.8km	0.534	0.337	806.9km	0.280	0.469	329.8km	0.548	0.483	312.8km	0.538
Back10	0.503	278.5km	0.026	0.428	338.0km	0.024	0.520	246.3km	0.144	0.527	193.7km	0.133
Jurg13	0.477	333.5km	0.623	0.331	948.1km	0.452	0.490	295.7km	0.629	0.512	256.9km	0.624
Comp14	0.509	250.6km	0.556	0.338	916.3km	0.462	0.517	240.8km	0.573	0.543	171.4km	0.566
Rand. N.	0.421	509.7km	0.623	0.310	1215.0km	0.452	0.437	399.7km	0.629	0.455	381.7km	0.624

Table 3: System performance when varying the gazetteer used to identify self-reported locations.

produced 255,094 cities. Recording all the variants resulted in 39,621,877 initial lexicalizations. Resolving ambiguities and removing naming variants of mixed orthographic scripts produced a final set of 11,735,961 location names.

A point of concern for the existing three gazetteers is that their list of locations does not match the most likely cities from which people tweet. Therefore, for the fourth source, we manually construct our own gazetteer using the Google reverse geocoder service, which returns a descriptive name for a latitude and longitude. Using a held-out set of three months of a 10% sample of Twitter data, the latitude and longitude values of all GPS-tagged tweets were aggregated and sorted by frequency. Building the gazetteer for all 89.2M unique GPS locations is infeasible due to the service's rate limit of 2,500 queries per day. Therefore, the reverse geocoder service was queried for the name of each location, beginning with the most frequently seen locations to maximize coverage. Our final gazetteer was created after seven months of querying, comprising locations for 536,265 points, with 12,324 unique city names. We refer to this gazetteer as Google.

Results Despite the three orders of magnitude difference in the number of location name variations between the gazetteers, roughly the same number of self-reported location names were matched: 515,653 using Geonames, 75,253 using DBpedia, 589,449 using GeoLite, and 527,638 using Google. Table 3 shows the methods' resulting performances.

Three important trends emerge. First, DBpedia extracted the fewest locations, highlighting an important limitation the resource. The inconsistent markup of cities within Wikipedia results in some large cities not being recorded as Settlements within DBpedia. As a result, methods using the DBpedia gazetteer observe lower recall from omitting these cities; furthermore, because these locations are never present in the network, multi-pass methods see a greatly increased error rates from propagating the remaining inaccurate locations to individuals whose true locations cannot be extracted. To be fair, additional engineering could potentially recover more locations from DBpedia, but considering the extensive efforts made with this resource beyond that of other papers, future works must take care in how this resource is used to ensure replicability.

Second, despite the Google gazetteer having the fewest number of names (12,324), methods using it have the best performance on all three metrics. An analysis of the naming matches revealed two reasons for the improved performance.

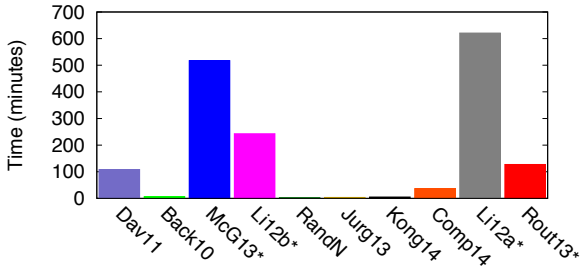

Figure 2: Training times on the one-month training data set. Methods marked with a * were those for which we had to provide algorithmic performance improvements.

First, because the gazetteer was created from the most frequent GPS locations seen in the data, the city name list included all the major cities omitted in the other gazetteers. Second, because the gazetteer is small, it contains few cities with ambiguous naming; the city names that are ambiguous in other gazetteers are automatically resolved to the city most likely to be tweeted from by virtue of the gazetteers construction. Nevertheless, the gazetteer does omit many moderately-sized cities and because the API-based data is proprietary, the gazetteer cannot be freely shared, limiting its utility.

Third, in the most-accurate setting using the Google gazetteer, methods obtained nearly identical User Coverage as with GPS-based locations (Table 2, col. 3) but in most cases had more than double the Median-Max error. This result suggests that with the current method of extracting self-reported locations, this source of information results in uniformly-worse accuracy. However, future work may potentially still obtain better results with self-reported locations by targeting only highly-accurate self-reported locations or identifying more initial users' locations through resolving short ambiguous city references in the location field.

7 Experiment 3: Temporal Decay

Geoinference models are expected to perform on a dynamic platform, with new users regularly joining and with existing users posting again after a dormant period. A natural solution to the challenge of new users is to add new training data. However, training a geoinference model can be computationally expensive. Figure 2 shows the average training times for constructing a model from the full one month

dataset described in Section 4.[6] Furthermore, none of the models scale linearly with the amount of data, so increasing the amount of training data can make retraining prohibitively expensive and raises questions about how practical is a particular approach. Therefore, in this third experiment, we evaluate how each geoinference method changes in accuracy and coverage when predicting locations for novel Twitter data created at increasingly-distant points in time after the last date seen in the training data. Further, this experiment examines the degree to which cross-validation performance is predictive of future performance by the models.

Setup All nine methods and the baseline were trained on the full August 2014 dataset (Sec. 4) using GPS data for ground truth, which provided the best performance in earlier experiments. Then, a test set was built from a 10% sample of Twitter data for each day in the following month (September) for a total of 30 distinct test sets with an average of 41.7M tweets per day. Each trained method then predicted the locations of all tweets seen in each of the 30 days in the test set. AUC, Median-Max, and Post Coverage were calculated for each day of the test set. To control for the influence of ground truth in the accuracy measurements, we calculate AUC and Median-Max only for those tweets produced by users who did *not* have ground truth locations in the training data. Since most methods retain the ground-truth locations of users in the model, were tweets from these users to be included, AUC and Median-Max would reflect the degree to which the ground truth predicted these users' tweet locations, rather than the predictive ability of the user locations inferred by the model.

Results Geoinference performance uniformly declined across all methods. Figure 3 shows the decline in AUC and Post Coverage for each method; Median-Max performances were similar to AUC and we omit them for brevity. Geoinference methods saw decreases between 14.6% and 25.8% in Post Coverage, which for all methods was highly negatively correlated with time, with a mean Pearsons r=-0.88. This decay in performance suggests that in the best case, a method's Post Coverage half life is roughly four months before it is only able to infer half of the tweets compared to when it was initially trained.

However, unlike Post Coverage, AUC decreased only slightly, indicating that the location predictions made by the method are robust across time. Methods' change in AUC varied from $+0.8\%$ to -9.4%. Surprisingly, the AUC did not drop significantly for methods with the highest coverage (e.g., Jurgens (2013) and Compton, Jurgens, and Allen (2014)) indicating that even though these multi-pass methods use inferred locations to infer new locations, the quality of their predictions remains high.

Finally, we note that Post Coverage improved as much as 5% for all methods on the weekends (seen as the small spikes in Fig. 3). This variance highlights the need for considering the temporal scale of the testing data in future geoinference evaluations in order to produce fully-comparable performance measurements.

[6] As a further example, training all the models for Experiments 1 and 2 required over 37 days of compute time on a state-of-the-art dedicated server –ignoring development and testing compute time.

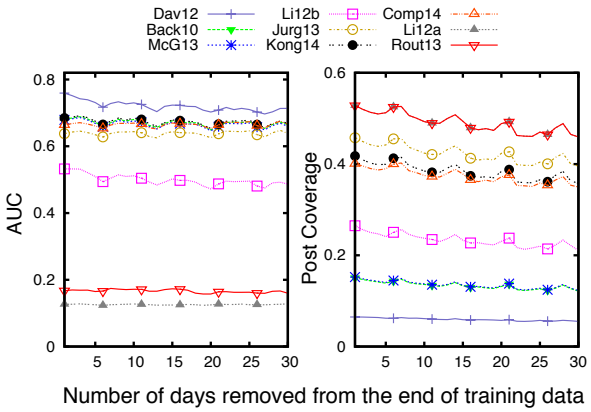

Figure 3: Performance on each day of the test data

8 Future Directions

Our work has identified significant performance differences among current state-of-the-art network-based geoinference techniques and as a result, three key area for future work in the area.

Self-reported location extraction: Prior approaches to extracting locations from self-reported location fields no longer perform well on current Twitter data (Cheng, Caverlee, and Lee 2010), with self-reported location data only moderately more prevalent than GPS data, independent of the choice in location gazetteer. Future work may increase the quantity and quality of self-reported locations by incorporating new toponym resolution techniques –especially those that consider both the user's posting content to aid in disambiguating their location field.

Twitter's temporal dynamics: Methods have yet to address the temporal aspects of the social network. As Experiment 3 revealed, these social-network based methods decline in recall, with an estimated recall half-life of four months. As such, they must be retrained periodically on new data. What effect new social relationships should have on location prediction remains an open question that is increasingly important at longer time scales as users move and form relationships with new individuals nearby.

Hybrid network- and text-based geoinference: The results of Experiments 2 and 3 show that network-based methods alone cannot infer locations for 37% of users –in the best case– leading to roughly half of all posts being unlocated. While gathering large amounts of longitudinal data may improve Post Coverage, methods that learn on other sources of information that the social network are needed to infer locations for asocial users for which social data is yet unavailable. Such text data may further assist network-based methods for challenging user types, such as disambiguating the present location of a mobile user. Hybrid methods combining both social and text information represent a promising class of solutions, with very recent approaches in this

direction showing promising results (Rahimi et al. 2015).

9 Conclusion

This paper has analyzed the current state of the art for network-based geoinference techniques. What began as a routine re-implementation of existing work revealed widespread disparity in the conditions, data, and metrics used to evaluate these works. Our work has provided three key contributions towards addressing these issues. First, we have created the first comprehensive assessment of state of the art for network-based geoinference. Using a common real-world test setting, we demonstrated that current methods differ significantly, with the best performance by methods whose original testing conditions more closely mirrored real-world conditions. To compare methods, we identified three key criteria and corresponding metrics, which fully capture the performance behavior of each method and allow for meaningful comparison, both in this and future work. Furthermore, a number of scalability issues were identified in these methods for which proposed solutions. All implementations, testing framework, and evaluation metrics are released as an open source package developing geoinference techniques, available at http://networkdynamics.org/resources/geoinference/. Second, we show that the current approach of using self-reported locations as ground truth yields far fewer ground truth locations than reported in earlier studies and even in the best configuration for matching these locations to coordinates, systems trained on self-reported locations suffer a large drop in accuracy, with the Median-Max error doubling for all but the lowest-performing systems. Third, we perform the first assessment of how well geoinference methods do at predicting locations of future posts, finding that, while accuracy rates only decrease slightly, Post Coverage was estimated to drop by half after a four month period.

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.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of CIKM*, 759–768. ACM.

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

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.

Gonzalez, R.; Cuevas, R.; Cuevas, A.; and Guerrero, C. 2011. Where are my followers? Understanding the Locality Effect in Twitter. *arXiv preprint arXiv:1105.3682*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of CHI*, 237–246. ACM.

Huberman, B. A.; Romero, D. M.; and Wu, F. 2009. Social networks that matter: Twitter under the microscope. *First Monday* 14(1).

Jeong, Y., and Coyle, E. 2014. What are you worrying about on facebook and twitter? an empirical investigation of young social network site users' privacy perceptions and behaviors. *Journal of Interactive Advertising* 14(2):51–59.

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

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

Liu, Y.; Kliman-Silver, C.; and Mislove, A. 2014. The Tweets They are a-Changin': Evolution of Twitter Users and Behavior. In *Proceedings of ICWSM*.

McGee, J.; Caverlee, J. A.; and Cheng, Z. 2011. A geographic study of tie strength in social media. In *Proceedings of CIKM*, 2333–2336. 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.

Newman, M. E. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics* 46(5):323–351.

Priedhorsky, R.; Culotta, A.; and Del Valle, S. Y. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of CSCW*.

Rahimi, A.; Vu, D.; Cohn, T.; and Baldwin, T. 2015. Exploiting text and network context for geolocation of social media users. In *Proceedings of NAACL*.

Rout, D.; Bontcheva, K.; Preoţiuc-Pietro, D.; and Cohn, T. 2013. Where's @Wally?: a Classification Approach to Geolocating Users Based on Their Social Ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*.

van Meeteren, M.; Poorthuis, A.; and Dugundji, E. 2009. Mapping communities in large virtual social networks. In *Proceedings of the 1st International Forum on the Application and Management of Personal Electronic Information*.

Wellman, B., and Gulia, M. 1999. Virtual communities as communities. *Communities in cyberspace* 167–194.

Zhu, X., and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University.