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

David Jurgens
HRL Laboratories, LLC
Malibu, CA USA
jurgens@cs.ucla.edu

Abstract

Social networks are often grounded in spatial locality where individuals form relationships with those they meet nearby. However, the location of individuals in online social networking platforms is often unknown. Prior approaches have tried to infer individuals' locations from the content they produce online or their online relations, but often are limited by the available location-related data. We propose a new method for social networks that accurately infers locations for nearly all of individuals by spatially propagating location assignments through the social network, using only a small number of initial locations. In five experiments, we demonstrate the effectiveness in multiple social networking platforms, using both precise and noisy data to start the inference, and present heuristics for improving performance. In one experiment, we demonstrate the ability to infer the locations of a group of users who generate over 74% of the daily Twitter message volume with an estimated median location error of 10km. Our results open the possibility of gathering large quantities of location-annotated data from social media platforms.

1 Introduction

Online social networks enable people to easily connect and maintain relationships with others independent of the individuals' locality. However, the physical location of participants in online networks has become an increasingly important factor in the analysis of social media. Modeling phenomena such as political elections, disease outbreaks (Paul and Dredze 2011), or appropriate responses to natural disasters (Mandel et al. 2012) often depends on knowing where individuals are located as they communicate about the phenomena. However, location-annotated data is often very sparse; for example, although many models leverage Twitter microtext, less than 1% of its data has been annotated with the coordinates of where the messages originated.

We propose a new method for estimating user locations in online social network that leverages social relationships and the spatial distribution of locations in an individual's local social network. Recent work has shown that geography does still matter in online social networks (Goldenberg and Levy 2009; Mok, Wellman, and Carrasco 2010; Gruzd, Wellman, and Takhteyev 2011) and this work builds upon that result to show that the spatial distribution of a user's network can be used to easily infer an individual's location. To test our hypothesis, we examine location and social relationships in the online platforms of Twitter and Foursquare, both individually and when relationships from both social networks are used.

Given the importance of location-annotated data, significant efforts have been made to infer location from a variety of information. Most frequently, user location is estimated from the content they produce by identifying geographic references or regional writing styles (Cheng, Caverlee, and Lee 2010; Ikawa, Enoki, and Tatsubori 2012). Other efforts have focused on mining the metadata associated with individuals, such as their self-reported locations (Hecht et al. 2011). More recent efforts have begun to examine the user's social network to infer location (Backstrom, Sun, and Marlow 2010; Davis Jr et al. 2011; Sadilek, Kautz, and Bigham 2012), but results have been limited to small user samples or to settings where location information is already plentiful.

Unlike previous approaches, our proposed approach is not specific to any social network and is dependent only on observable social relationships and a small amount of ground truth locations. In being based on relations, we are able to accurately estimate the locations of a larger segment of users than previous approach that rely on user-provided content. Furthermore, our approach may be combined with the location estimates from prior efforts as the ground truth locations, which we also examine.

The contributions of this paper are as follows. First, we demonstrate that a user's social network provides a powerful source of information for inferring their location. Second, we propose a new algorithm, spatial label propagation, that can effectively infer user locations given a small amount of ground truth. Third, we demonstrate multiple techniques for improving the accuracy of location information, ultimately showing that an estimated 50% of the users in a Twitter-based social network may be located to within 10km. These estimates provide a way for gathering high volumes of location tagged data, which we estimate can provide locations for over 74% of the messages being sent using Twitter on any given day. Fourth, we demonstrate the potential for leveraging location information from one social media platform to locate individuals in another, show that information from Twitter can locate 50% of the individuals in a Foursquare social network to within 25km.

2 Social Networks Data

Despite predictions that location is no longer a factor in the formation and maintenance of social relationships in the online setting (Cairncross 2001), recent studies have argued that a user's social network is still influenced by locality (Goldenberg and Levy 2009; Mok, Wellman, and Carrasco 2010). Individuals form offline relationships, which are then transferred to online social platforms and as a result, a user's social network is likely to include many individuals who are geographically close. We refer to the part of the network directly connected to an individual as their *ego network*. Therefore, we hypothesize that the ego network is a prime candidate for use in inferring an individual's location: The locations of the individuals in an ego network should predict of where that individual is. As a test, we evaluate three social networks extracted from two social media platforms, and then use the available ground truth data to measure the potential for location inference.

Twitter Twitter has provided one of the most popular platforms for social media. Individuals post short messages (tweets) and may form asymmetric social relationships, known as *following*, where one individual monitors the tweets of another individual. Additionally, Twitter provides a mechanism for specifically referencing another user by name, referred to as a *mention*, which can serve as a way to hold conversations (Honeycutt and Herring 2009).

In analyses of Twitter, Takhteyev, Gruzd, and Wellman (2012) and McGee, Caverlee, and Cheng (2011) both examined distance across multiple types of social relationship, finding a consistent pattern of users having friends that are nearby. Specifically, McGee, Caverlee, and Cheng (2011) note that the distribution of distances between mutually-following individuals had the highest percentage of short distances. Therefore, we crawled a network using the Twitter API, comprising 95,855 individuals and 16,609,095 bidirectional following relationships. The crawl was started from users that had known locations (described next in Sec. 2.1) to maximize the neighbor distance data for analysis.

Building a large-scale social network for millions of users with bidirectional following relationships is a time-intensive and potentially infeasible process due to the rate limits on accessing information from the Twitter API. Therefore, we consider a second type of social network built from users who both mention each other. We hypothesize that bidirectional mentions provide evidence of a social relationship via mutual communication. Using a 10% sample of all Twitter messages over the period of April 15, 2012 to November 16, 2012, we extracted 254,263,081 relationships between 47,760,573 unique users that had at least one bidirectional mention. Because the network is constructed from a sampling of Twitter, the network may exclude some relationships that are present in a full stream of the data; however, the network is significantly larger than what would be available by accessing the Twitter API. We refer to this network as the Mention network.

Foursquare Foursquare is an online location-based service platform where users can check in to locations such as restaurants and may also form a social network. Relationships in Foursquare are bidirectional and mutually agreed upon by both parties. Using the Foursquare API, we crawled 3,976,819 users over three months, extracting their complete social network, for a total of 17,619,191 relationships.

Foursquare users also provide information about their identities in other online platforms, notably Twitter. Therefore, we used the Twitter API to map our Foursquare users' self-reported Twitter usernames to their corresponding Twitter identities, which was successful for 1,617,615 individuals. This mapping also provided a way to validate the relationship inference from Twitter mentions. We observed 14,048,788 relationships in Foursquare where both individuals also had Twitter identities. Of those relationships, 7,412,589 (52.8%) also had an inferred relationship from bidirectional mentions in Twitter. We view the high percentage of multi-network relationships as an empirical validation of inferring social relationships from Twitter mentions.

2.1 Location Data

Both of the considered platforms provide some form of ground truth data, which enables us to analyze the distances within a user's ego network.

Foursquare Foursquare users provide both self-reported location data in the user profile as well as limited amounts of publicly-accessible GPS-tagged data. The largest amount of GPS-tagged data comes from publicly shared information on the user's Mayorships, which are awards given to users for being the individual that checks in the most to a specific location. While Pontes et al. (2012) found that Mayorships provided strong evidence of the users location, we note that they used the self-reported location data as ground truth, which has far higher coverage, with 98% of users providing identifiable locations. As a result, we use the self-reported location data as ground truth. We use a conservative location-mapping procedure that labels 2,735,701 (68.7%) of the users in our network with locations based on the text reported in their profile.

Twitter Twitter users provide both self-reported location data in the profiles as well as GPS data for messages that are tagged with locations. Unlike Foursquare, self reported data is very noisy and often not localizable at the city level (Hecht et al. 2011). Furthermore, only a small percentage of Twitter users enable GPS annotation of their posts, accounting for 0.7% of the observed messages.

Because GPS-tagged data is often produced by mobile phones, users are frequently seen in multiple locations. For the purposes of our analysis, we treat a user has having only one location, which is considered to be representative of where that user is most likely to be. Therefore, we construct our set of located Twitter users by restricting it to only those with at least five GPS-tagged tweets where at least five occur within a 15km geographic radius. Each user is then assigned a single location using the geometric median, m, of their GPS locations, L,

$$m = \arg\min_{x \in L} \sum_{y \in L} distance(x, y), \qquad (1)$$

where orthnormic distance is calculated using Vincenty's formula (Vincenty 1975). Equation 1 is a specialization of

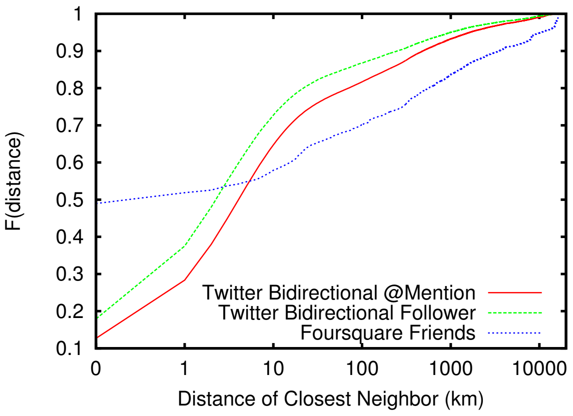

Figure 1: The cumulative distribution functions for the distance to a user's geographically closest friend.

the multivariate L_1 median to operate on spheres (Vardi and Zhang 2000). We opt to use a median location, rather than a mean, as the median represents an actual location of the user and furthermore avoids assignment a user a non-meaningful location from averaging locations. Furthermore, the geometric median is robust to location outliers, such as when an individual posts GPS-tagged messages from vacation or an atypical location far from the normal concentration of locations. Ultimately, 2,554,064 (5.34%) of the Twitter users in our network are assigned locations.

2.2 Neighbor Locality

While previous studies have examined the distribution of distances within a user's ego network, we ask what is the distribution of distances to individual's geographically nearest neighbor. If the ego network is useful for location inference, then the closest neighbor represents the maximally predictive information that is initially available. Therefore, for each network, we measure the distance between each individual and the closest neighbor in their ego network.

Figure 1 illustrates the cumulative distribution functions (CDF) for each network of the distance to the closest neighbor, where $F(x)$ denotes P$(distance \leq x)$ and x is a distance in kilometers. The CDF demonstrates that the nearest neighbor is highly predictive of the individuals location, with all three networks showing that over half of the individuals have a neighbor that predicts their location to within 4km. Despite being four order of magnitude different in size, the bidirectional Follower and Mention networks both exhibit similar trends with their error distribution. We view the superior predictive performance of the Follower network being due to the higher prevalence of users following their nearby friends without engaging in conversation with them. The Foursquare network exhibits a large probability mass for users at the exact same location (distance zero); however, this is due to the method used to assign users locations. User coordinates are derived from location names so users in the same city will have zero distance, despite possibly being several kilometers apart physically; in contrast, the Twitter network uses GPS coordinates, and therefore distances are more likely to vary on shorter scales.

3 Location Inference

Given an individual's social network, selecting the nearest individual can provide strong evidence of the individual's location, as shown in Figure 1. However, two key problems exist for using this information. First, given the ego network, the choice in which neighbor should be selected is unclear, with many potential methods. Second, location data may be sparse, as in the case of the Mention network, which only contains locations for approximately 5.34% of the users, and therefore many users will have no neighbors with locations. Therefore, we propose a new method for location inference in social networks, spatial label propagation, and then evaluate a series of heuristics for selecting which of the neighbors' locations should be used.

3.1 Label Propagation

Label propagation is a semi-supervised, iterative algorithm designed to infer labels for items connected in a network (Zhu and Ghahramani 2002). Usually, the true labels are known for only a small number of items in the network, which serve as a source of ground truth information for estimate the labels of other nodes. The algorithm proceeds iteratively, where in each round, items receive the most frequent label from their neighbors.

Our extension to label propagation recognizes that the labels themselves may be interpreted spatially, which impacts the update procedure for each round. Rather than selecting the most frequently label of their neighbors, the geometric configuration of the neighbors can be to select the current node's new label. The algorithm is formalized as follows. Let U be the set of users in the social network and N be a mapping for each user to the other individuals in their ego network $\{u \rightarrow \{n_1, \ldots, n_m\}\}$. Let L be a ground-truth mapping from users to their known coordinates $\{u \rightarrow (latitude, longitude)\}$. Spatial label propagation then proceeds according to Algorithm 1. Algorithm 1 contains two key parameters: (1) the definition of the $select$ function that uses the spatial arrangement of the locations in N and (2) the stopping criteria. We note that traditional label propagation has a closed form solution when the most frequent label is selected (Zhu and Ghahramani 2002), and therefore requires no stopping criteria; however, no closed form exists when using the medians described next in Sec. 3.2.

3.2 Location Selection Methods

The choice in $select$ function is crucial to accurate location inference. We consider three variants and two baseline methods for selecting a location from the list of neighbors' locations. First, we consider using the geometric median (Eq. 1), as described in Sec. 2.1.

Second, we consider an alternative multivariate median definition using Oja's Simplex Median (Oja 1983). Oja's median is defined as

$$m = \operatorname*{arg\,min}_{x \in L} \sum_{y,z \in L} area(x, y, z), \qquad (2)$$

where $area$ is the surface area of the simplex defined by the points x, y, and z over the ellipsoid for the Earth's surface.

```
Data: $U$, $L$, and $N$
Let $E$ be the current mapping from user to location;
Initialize $E$ with $L$;
while Convergence criteria is not met do
    Let $E'$ be the next mapping from user to location;
    for $u \in U - domain(L)$ do
        Let $M$ be a list of locations;
        for $n \in N(u)$ do
            if $E(n) \neq \emptyset$ then
                add $E(n)$ to $M$;
            end
        end
        if $M \neq \emptyset$ then
            $E'(u) = select(M)$;
        end
    end
    $E = E'$
end
Result: Estimated user locations, $E$
```

Algorithm 1: Spatial Label Propagation defines the *select* function to use the spatial distribution of the location M in deciding the location of u in the next iteration.

Third, we consider a heuristic based on social theory. Social networks often exhibit triadic closure: Given relationships (a, b) and (a, c), it is likely that b and c will also have a relationship, i.e., forming a triangle between the three individuals (Kossinets and Watts 2006). We view these closures as evidence of a stronger social relationship from a to b and c, and subsequently hypothesize that they may provide a better source of information on the location of a. Therefore, we incorporate a heuristic that first filters the social network of an individual to only those in triadic closure and then computes the geometric median of the locations of those neighbors. Should the individual not have locations for those individuals or not have any neighbors in triadic closure, we instead compute the geometric median over all the neighbors' locations. We refer to this method as the Triangle Heuristic.

The first baseline method selects a random location from M; we refer to this baseline as Random Neighbor. As a second baseline method, we use the method proposed by Davis Jr et al. (2011) where location coordinates are converted into names (a process known as reverse geocoding), and then the most frequent name is selected.[1] To covert coordinates to names, we use the World Gazetteer database for canonical names and fall back to Google Reverse Geocoding service for coordinates that did not have a name. This second baseline is equivalent to performing label propagation where users are labeled with names instead of coordinates, and so we refer to it as Traditional Label Propagation.

3.3 Experiment

Each of the medians was tested on the Twitter Mention network, which was selected due to having the largest coverage and a small amount of labeled data. To measure performance, we used five-fold cross validation: the 2,554,064 users with associated locations were divided into five sets and then performance was measured by running each algorithm over the entire network of 47.7M users, using four sets of located users as ground truth (2,043,252 users) and then calculating the error on the remaining 510,812 held-out users. Sets are rotated so that each is held out once.

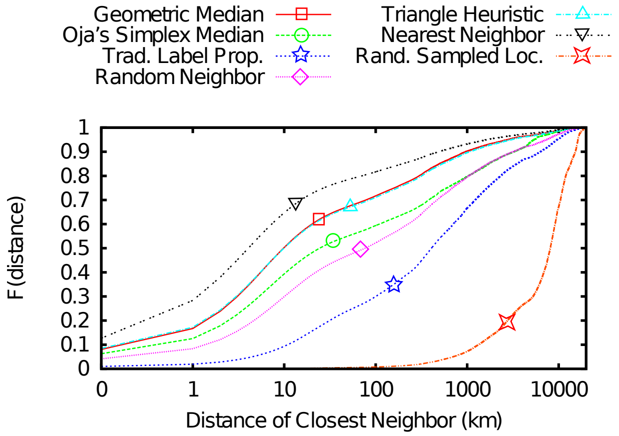

Figure 2: Error probability distribution location inference method after four iterations in the Twitter Mention network.

We include two additional non-*select*-based baselines: (1) the error distribution of selecting the nearest neighbor among the 5.34% of the network with known locations and (2) the error from assigning a random location to an individual. The random baseline was constructed by assigning all individuals' locations by sampling without replacement from the observed locations. Error was estimated from 30 complete trials of sampling and then computing the error distribution in aggregate. We note that the nearest-neighbor should not be considered the upper bound; for example, if an individual has n neighbors, only one of which has a location, multiple iterations may provide locations to the remaining $n - 1$ neighbors, which can improve performance if the only labeled neighbor is distant. However, the nearest neighbor does represent an upper bound when only a single iteration of Algorithm 1 is used.

3.4 Results

Spatial label propagation converged quickly, as shown in Figure 3, and therefore, we report results after four iterations completed in order to measure performance across the 45.8M located individuals in the network. Additional iterations did not significantly increase coverage and often reduced performance. Figure 2 illustrate the performance of each method using a CDF. For clarity, we make two notes for interpreting this CDF and those of later sections. First, a comparison of methods along the x-axis of the CDF reveals the difference in performance for a percentage of the network; for example, at $F(x)$=0.5, the expected error for 50% of the individuals is in the range $[0km, 4km]$ for the nearest neighbor baseline, compared to the range $[0km, 406km]$ for traditional label propagation. Second, a comparison of

[1] Davis Jr et al. (2011) used a non-iterative, name-based setup for location inference, equivalent to a single iteration of traditional label propagation.

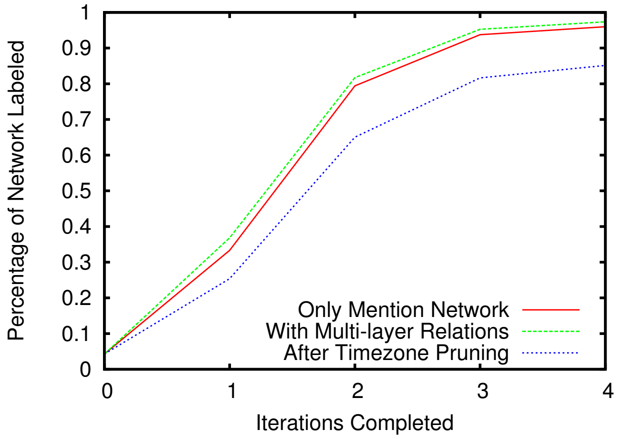

Figure 3: The percentage of users assigned a location per iteration of spatial Label propagation for three networks

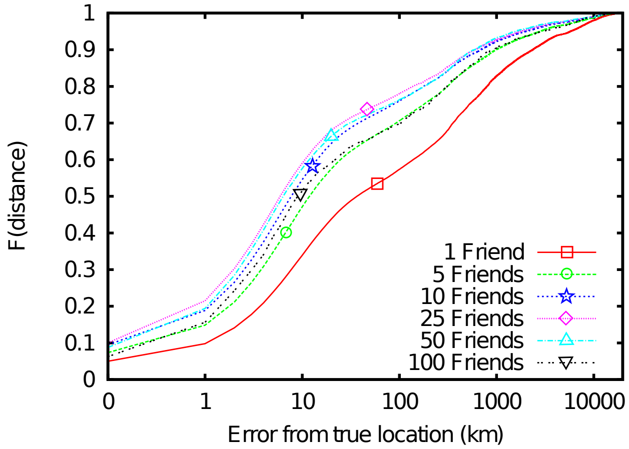

Figure 4: CDF of the error using the geometric median for different sizes of the ego network

methods along the y-axis of the CDF reveals the difference in the number of individuals that are included in the error range specified by the x-axis. Due to the scale of the network, each 0.1 increase in $F(x)$ at a specific x corresponds to a reduction in the error bound for 4.7M users.

The choice in *select* method has a significant impact in the performance of location inference. Of the methods, the geometric median offers significantly higher performance than other methods. While the geometric median does not match the performance of the nearest neighbor baseline, it does provide relatively high accuracy, locating over half of the network with an error in $[0km, 10km]$.

The triadic heuristic provided nearly identical performance with the geometric median without the heuristic. Scellato et al. (2011) note that the size of triads in social networks from three location-based services (including Foursquare) varied depending on the size of the user's ego network, where users with more friends being involved in triads across longer distances. A further analysis of all

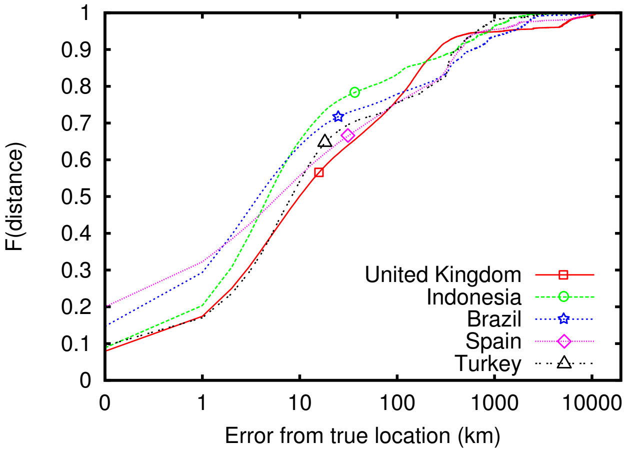

Figure 5: CDF of the error using the geometric median for the five of the six countries with the most GPS-based users

57.2M GPS-tagged triads in the Mention network revealed a Spearman's correlation of 0.278 between the length of the triad and the size of a user's ego network, which is low but statistically significant. This correlation suggests that the triadic heuristic should not be expected to improve performance given its tendency to include more distance neighbors as the size of the ego network increases.

Surprisingly, using the location of randomly selected neighbor results in a large improvement over the location name-based method used by Davis Jr et al. (2011), shown as Traditional Label Propagation in Figure 2. Analyzing the name-based method's performance further showed that the low performance was often due to having multiple individuals in the ego network labeled with the name of a large population center; because individuals are more likely to live in such locations, the frequency-based approach is more likely to select the name of a populous city rather than the geographically distributed names that are closer to the individual. In contrast, the geometric median is robust to such outliers, provided they do not constitute over half of the locations in the ego network.

In a secondary analysis, we measured the performance of the geometric median for two partitions of the data: Figure 4 shows performance according to difference sizes of the ego network, and Figure 5 shows performance for the six countries with the highest number of GPS-tagged users. Location inference performance increases significantly as the ego network size grows; the difference in $F(10km)$ between a user having one friend and twenty five friends is 0.259, which corresponds to approximately 12M individual having their location estimate improved to under 10km. However, we note that for performance does degrade for highly popular users whose ego networks are larger than one hundred individuals; such individuals often have widely distribution social networks, which decreases the likelihood of the geometric median selecting a location near their actual location.

Country-specific performance was related to the geographic distribution and density of its major population centers. For example, despite Brazil being the fifth geograph-

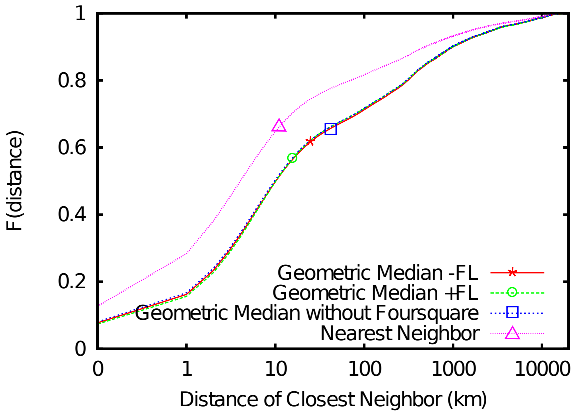

Figure 6: Error probability distributions for location inference after four iterations in the multi-layered network. The inclusion and exclusion of location data from Foursquare is denoted using +FL and -FL, respectively.

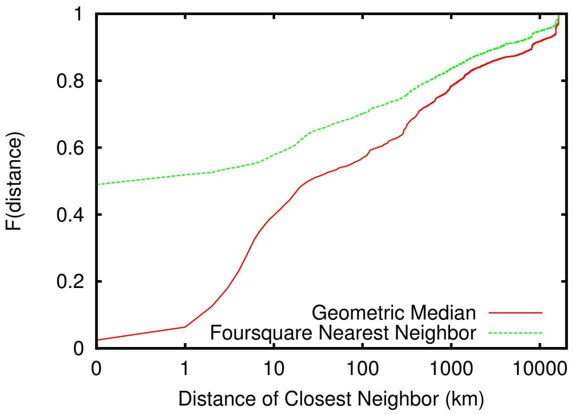

Figure 7: Error probability distributions for location inference for Foursquare users in the merged Twitter-Foursquare network using only location information from Twitter.

ically largest country, inference performs well because the majority of its individuals in the Mention network reside in its largest cities, São Paulo and Rio de Janeiro, and therefore individuals in the ego network are likely to be nearby.

Last, the number of users in the Mention network is under 10% of all Twitter users, with estimates of the total number of users on the Twitter platform above half a billion. Therefore, as a measure of the location annotations' practicality, we tested what percentage of the daily Twitter volume was generated by these users. Using a held-out set of three days in November, 2012, we found that the individuals that had been located generated approximately 74.7% of the total volume during that period, yielding two orders of magnitude more location-tagged data than available with GPS data alone.

4 Experiments

Given the established accuracy in estimating user location via spatial label propagation, we consider its performance in four additional setting: (1) combining information from multiple social media platforms, (2) using noisy self-reported locations instead of GPS data, (3) leveraging linguistic similarity to predict more local neighbors in the ego network, and (4) pruning geographically distant relationships across which location should not be propagated.

4.1 Location Inference in Multi-layered Networks

Individuals in online social network platforms often provide information on their identity in other platforms, acting as a way of linking the multiple platforms into a single multi-layered network. These multi-layered networks have the advantage of both significantly increasing the number of individuals and location information, but also providing more evidence of social relationships between them, using indicators from each platform. Specifically, for the spatial label propagation, the additional relationships can potentially remove bottlenecks in the graph, where too few relationships exist to accurately update the next iteration. Second, merging social networks also enables location inference for individuals in other platforms that do not contain easily-accessible location information for their users, e.g., Tumblr; location information may be propagated from one platform to another using individuals with multi-platform identities.

Methodology We propose two experiments: (1) merging two social networks to examine the potential improvement in estimating user location, and (2) leveraging the location information in one network to estimate the locations of the users in the other. Accordingly, we selected the Twitter and Foursquare platforms, as both have a significant overlap in the number of shared users and both have location information, which is required for the second experiment. While our Twitter Mention network is an order of magnitude larger than our Foursquare network, the resulting combination still provides a significant number of new individuals and relations through which locations may be propagated.

For the first experiment, we consider two variations. In the first, only Twitter-derived locations are used, which measures the impact of including additional social relations. In the second variation, we include additional location data for all 846,079 Twitter users that have a known Foursquare-based location but not a Twitter-based one. This second variation measures the potential for increasing performance by adding additional ground truth data. Performance is measured the same as in Section 3.3, with five-fold cross validation. To maintain direct comparability between experiments, we report the performance for the first experiment only for those Twitter users with locations.

For the second experiment, all location information in Twitter is used to infer the locations of users in the merged network. Next, we measure the accuracy of the locations for the 1,106,647 Foursquare users who self-report their location but do not have no reported Twitter identities, as these are the additional nodes in the network beyond what is present in Twitter alone. As a baseline, we compare against selecting the nearest-neighbor in the network (Fig. 1).

Results In the first experiment, the inclusion of both Foursquare relationships and locations did not significantly impact the location inference accuracy, with only minor differences in performance, as shown in Figure 6. However, the inclusion of the data did improve the convergence of the algorithm and its total coverage. Figure 3 shows the increase in coverage in only users in the Twitter Mention network as the algorithm proceeds. After four iterations, the new relationships from the Foursquare network result in an additional coverage of 659,240 individuals in the Mention network (1.3% absolute increase) with no loss of accuracy.

The second experiment indicates that accurate location inference across platforms is possible, as shown in Figure 7. Notably, after four iterations, all users in the Foursquare network were located. Although performance is not as high as the nearest-neighbor baseline, the error bound is low for the majority of users: 39.7% of the users are located within $[0, 10]$ kilometers of their self-reported location and 50.1% within $[0, 25]$.

We hypothesized that the performance is constrained by the spatial distribution in the connections between the two platforms. The multi-platform individuals serve as the only gateways for propagating information from one platform to the other. However, if these individuals are not spatially distributed, their relative geographic concentrations bias the locations that flow into the other platform. To test this hypothesis, we examined the geographic distribution of the gateway individuals compared with the distribution of the individuals in the Foursquare network that were being located. Figure 8 shows the country-level distributions for the gateway individuals and individuals in the held-out set for nine of the ten most-represented countries in both groups. Here, we see significant differences in geographic distributions of the two sets: the second and third most represented countries in the held-out set are significantly underrepresented, having half of their expected frequencies in the gateway individuals. Accordingly, individuals in these countries are constrained by smaller bottlenecks through which nearby locations are propagated to them, and as a result, are assigned more distance locations from the over-represented locations.

4.2 Self-Reported Locations

In many social network platforms, individuals may provide a self-reported location, which has served as the source of location information for prior work, e.g., (Hecht et al. 2011; Pontes et al. 2012), despite the potential for noise. For some platforms, GPS data may be unavailable and therefore we consider the impact of using noisier seed locations in the absence of ground truth information. Furthermore, we hypothesize that spatial label propagation may help smooth out errors in the location field, ultimately producing better results than may be obtained through using it directly.

Methodology Self-reported locations were extracted from the Twitter profiles of all users in the mention network. To prepare the self-reported location names for spatial label propagation, we first attempt to convert each name into specific GPS coordinates. The Google geocoding service was used to map each name to specific coordinate. As Hecht et al. (2011) note, converting location names to coordinates is highly imprecise, where under two thirds of users provide valid locations, and even fewer specify locations at the city level. Ultimately, this produced coordinates for 11,319,349 users. We seed the mention network with all locations and then proceed to run spatial label propagation using the geometric median. Because no ground truth data is used for seeding, we do not perform cross validation as in previous experiments, instead using all self-reported data for a single run and then evaluating based on ground truth.

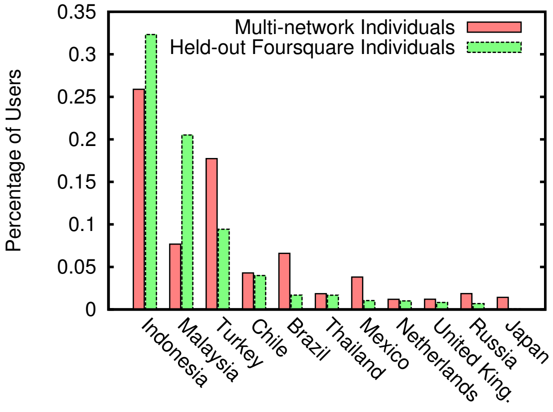

Figure 8: The percentage of users in nine of the ten most-represented counties in the held-out set of Foursquare users, relative to their percentages in the set of gateway individuals bridging the two platforms.

Results As a baseline for the error prior to spatial label propagation, we computed the CDF for the distance between the self-reported location and the GPS-derived location for the 793,604 users that had both. As Figure 9 shows, self-reported locations can serve effective seed data for location inference. While the performance is not as high as when using GPS-based locations as seed, the results show that inference is still competitive for platforms where GPS data is unavailable. Furthermore, spatial label propagation is able to infer the locations of 97% of the individuals in the network using self-reported data from only 23.7% of the users with only a few kilometer loss in precision.

Regarding the second hypothesis, Figure 9 shows no evidence of error smoothing to improve performance beyond the initial conditions. However, some error smoothing is seen in subsequent iterations; the second iteration sees a reduction in error at $F(x) = 0.5$ from $[0, 21]$ in the first iteration to $[0, 15]$, which is a direct result of the geometric median having more locations from which to select.

4.3 Predicting Proximity from Linguistic Similarity

Frequently users in social relationships communicate with each other, both offline and online, which potentially encourages homophily, where individuals adopt traits and interests of others in their social group (McPherson, Smith-Lovin, and Cook 2001). The content of the conversations can provide evidence of social influence (Danescu-Niculescu-Mizil et al. 2012), and we hypothesize that users who live nearby

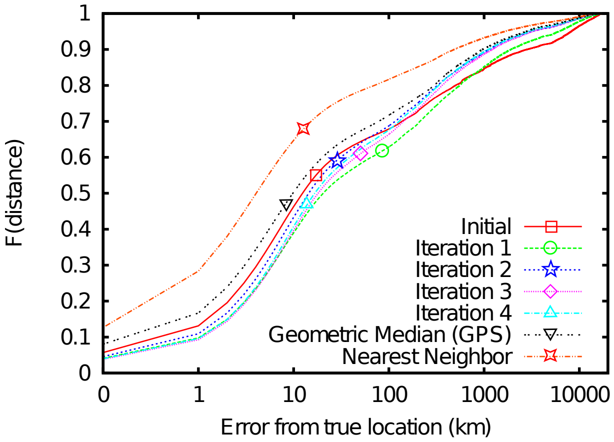

Figure 9: Cumulative distribution of errors after each iteration of spatial label propagation using only self-reported locations for initialization

were more likely to influence each other in addition to discussing more local topics and therefore, the similarity of their language may provide an important clue as to which individuals in the ego network may be most proximate.

Methodology To test whether linguistic similarity can predict proximity, we first construct a document representation for each individual, using the concatenation of all of their messages. Due to the high amount of lexical variety, we clean and normalize messages by lower-casing all tokens, removing all punctuation (with the exception of the # sign in hashtags), replacing all numbers with "NUM," and replacing all hyperlinks with "URL." Messages are then tokenized by whitespace to create a distribution over the types of tokens used by that individual. We limit our analyses to all users who generated at least ten unique tokens.

Two methods are used to compare users. First, we adopt the representation of Cheng, Caverlee, and Lee (2010), who in a similar setting, treat the individual's token distribution as a unigram language model, which is a probability distribution over the terms. Individuals are compared by computing the Jensen-Shannon (JS) Divergence between the two distributions. For two users' distributions P and Q, the JS Divergence is calculated as

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)$$

where M is the average of the two distributions and D_{KL} denotes the Kullback-Liebler Divergence. The JS Divergence is bounded in $[0, 1]$, with zero and one denoting identical and maximally dissimilar distributions, respectively.

Second, we consider an alternate method based on the traditional vector space model of documents, treating the concatenated messages of each user as an n-dimensional vector, where n is the vocabulary size, and $n(i)$ indicates the individual's usage frequency for the unique token mapped to dimension i. Individuals are then compared using the cosine similarity of their vectors. Because all token counts are positive, the cosine similarity will range in $[0, 1]$; however, to enable better comparison as a distance function, we negate this value without loss of generality such that with lower values indicate similar message content and values near zero indicate mutually exclusive content.

Performance for both measures was then computed using Pearson's correlation, r between the distances of the neighbors in the ego network and the measures' respective values. The average correlation was constructed by applying a Fisher transformation to each r to produce z and then computing the average correlation from the average z; this method has less potential for bias than averaging to the r values directly (Corey, Dunlap, and Burke 1998). Last, we note that while both lexical representations are admittedly simple because of a need for computational efficiency for the scale this evaluation, both have proven effective in measuring similarity in many domains (Turney and Pantel 2010).

Results Surprisingly, both comparisons methods yielded near-zero correlations, with $\overline{r}_{cosine}$=0.011 and $\overline{r}_{JS}$=0.030. Additionally, we tested whether the correlations may change based on the size of the ego network; however, all network sizes in $[5, 211]$ followed the same trend. Similarly, an analysis of the correlations specific to the individual's country of origin showed near zero correlation for all countries. These results suggesting that a surface-level comparison of two users' lexical content is insufficient at predicting their geographic proximity. While our lexical representation is simple, we view this negative result as important for motivating deeper analyses that compare semantic content and discourse structure in order to predict nearby locality.

4.4 Using Time Zone Boundaries

Individuals in online social platforms may also reveal geographically-distinguishing features other than their location. Specifically, we consider time zone information provided on Twitter profiles as a way of potentially improving location inference by prevent location propagation between users in different timezones. However, Hale, Gaffney, and Graham (2012) note that because time zone data, specified as a UTC-offset, is self-reported, the data is very noisy; based on an analysis of GPS-tagged tweets, they estimate that only 69.2% of users setting the correct UTC offset. Indeed, in our longitudinal analysis, we found that 58.8% of users set their timezone, with an average of 1.02 timezones for users that provided the information due to changes in their profile. Nevertheless, we hypothesize that pruning cross-timezone relationships may reduce noise in spatial label propagation by prohibiting the adoption of a location from a knowingly-distant neighbor.

Methodology Each user is mapped to the set of timezones that they self-reported over six month duration of the dataset. We observed at least one timezone for 28.1M users in the Mention network. We then remove edges between users who timezone sets are disjoint. However, edges were not removed if either user had never specified any timezone. This pruning removed 96,774,420 of the edges in the network, resulting in a network with 157,488,661 edges.

Results Location inference performance with timezone pruning was nearly identical to the performance with the full

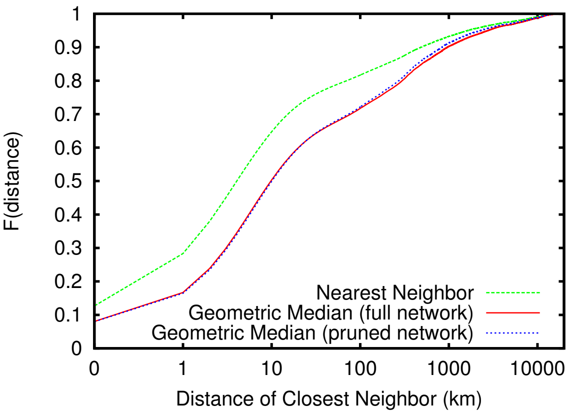

Figure 10: Inference performance when pruning relationships between individuals in different time zones

network, as shown in Figure 10, despite 38% of the edges being removed; because the geometric median is robust to outliers, performance without pruning was high even with the inclusion of locations from cross-timezone edges.

Timezone pruning does still provide two significant effects. First, from a performance standpoint, pruning edges reduces the size of the ego networks and as a result, reduces the overall run-time of four iterations of spatial label propagation to approximately one third of the time required for the full network. Because performance did not decrease with the pruning, such pruning may be essential for larger networks with billions of edges when the run-time becomes infeasible.

Second, the pruning does impede label propagation and leads to a 10.9% absolute reduction in network coverage, as shown in Figure 3. Specifically, pruning creates bottlenecks through which locations cannot propagation and also creates singletons by removing all the edges to a user. Additionally, using more than four iterations of the algorithm did not significantly increase coverage due to the majority of the loss in coverage coming from user who had all of their edges removed. Future applications may consider the trade-off of losing data from creating singletons via pruning versus including noisier location estimates for those users that are only accurate at larger geographic levels.

5 Related Work

Given the utility of location-tagged data, several works have examined location inference for users and messages, both in Twitter and other online services. Due to its wide-spread use, most work has be performed for Twitter. Cheng, Caverlee, and Lee (2010) Mahmud, Nichols, and Drews (2012) and Ikawa, Enoki, and Tatsubori (2012) have examined using the text content produced by a user for inferring their location. While having good results, these approach are often limited to only those users who generated text that contain geographic references, whereas our approach works independent of language because it is based on the social network alone.

Sadilek, Kautz, and Bigham (2012) estimate the users future location through the locations of users in their ego network. However, their approach requires that both users locations be known in order to estimate the social relationship, which limits the approach to only those individuals with known locations. Furthermore, the approach is limited to only users with users with highly active GPS data.

Davis Jr et al. (2011) use a user's Twitter follower network to perform location inference. While their approach is based on location information in an individual's ego network, it uses location names only, which in our analysis was the least-precise location inference method for label propagation. Furthermore, their approach is non-iterative, which limits their method's ability to infer location for most users due data sparsity.

Hecht et al. (2011) and Pontes et al. (2012) infer user locations from self-provided location information in Twitter and Foursquare respectively. While Pontes et al. (2012) reported more than 90% coverage of users in Foursquare with this method, no attempt was made to infer the locations of the remaining users. Hecht et al. (2011) found significantly lower information in Twitter profiles, with a high error rate. These approaches could be considered orthogonal to the provided approach and could be potentially used together to leverage multiple sources of information, as in Sec. 4.2.

Backstrom, Sun, and Marlow (2010) propose a location inference method for the Facebook social network using probabilistic inference to select the location from a user's friends. Although designed for a network with very dense location information, they do perform an analysis where the algorithm is iterated on a social network with artificially sparse location data for only 25% of the users. Although operating on a different graph, we note that their iterative performance predicted the locations for 57.3% of the users within 40.2km, whereas our method predicts for 57.2% within 16km and 65.9% within 40km despite having nearly an order of magnitude less location data to start with. However, we do note that the accuracy of their method improves as the size of the ego network increases, in contrast to the results seen in Section 3.4; future work may consider a hybrid approach that switches inference methods as the location data becomes denser.

6 Conclusion

The social relationships in online platforms provide strong evidence of an individual's location. We have presented a new algorithm, spatial label propagation, that leverages the geographic distribution of an individual's ego network to infer their location, showing that after multiple iterations, nearly all of the users in the social network are located, with an estimated median error under 10km. Further, we demonstrated that the method is accurate both for individuals in different countries and ego network sizes.

In a series of related experiments, we analyzed variations on the method showing that (1) multiple social media platforms can be leveraged to gain additional social relationships, resulting in higher coverage; (2) in the absence of GPS data, noisy self-reported location information can effectively be used with a small loss in precision; (3) spatial proximity in the ego network is not predicted by users' lexical content at the token level, and (4) the computational efficiency of the method can be significantly improved by pruning the social network based on external information that users are not physically proximate, such as timezone data.

Our proposed method is scalable, being tested on networks with tens of millions of nodes and hundreds of millions of edges. Furthermore, our method is complementary to many existing approaches to location inference, which can be used to provide the initial location data for propagation. In addition, our method can be used to leverage location data from one social media platform to those with no location information, which enables geospatial analyses on new social media platforms. As a result, the method opens up the potential for gathering large volumes of location-annotated social media data for location-based phenomena.

Acknowledgements

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.

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.

Cairncross, F. 2001. *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press.

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.

Corey, D.; Dunlap, W.; and Burke, M. 1998. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *The Journal of general psychology* 125(3):245–261.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, 699–708. 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.

Goldenberg, J., and Levy, M. 2009. Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*.

Gruzd, A.; Wellman, B.; and Takhteyev, Y. 2011. Imagining twitter as an imagined community. *American Behavioral Scientist* 55(10):1294–1318.

Hale, S.; Gaffney, D.; and Graham, M. 2012. Where in the world are you? geolocation and language identification in twitter. Technical report, Working paper.

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.

Honeycutt, C., and Herring, S. 2009. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of HICSS*, 1–10. IEEE.

Ikawa, Y.; Enoki, M.; and Tatsubori, M. 2012. Location inference using microblog messages. In *Proceedings of WWW*, 687–690. ACM.

Kossinets, G., and Watts, D. 2006. Empirical analysis of an evolving social network. *Science* 311(5757):88–90.

Mahmud, J.; Nichols, J.; and Drews, C. 2012. Where is this tweet from? inferring home locations of twitter users. In *Proceedings of ICWSM*, volume 12.

Mandel, B.; Culotta, A.; Boulahanis, J.; Stark, D.; Lewis, B.; and Rodrigue, J. 2012. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of NAACL-HLT*, 27.

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

McPherson, M.; Smith-Lovin, L.; and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

Mok, D.; Wellman, B.; and Carrasco, J. 2010. Does distance matter in the age of the internet? *Urban Studies* 47(13):2747–2783.

Oja, H. 1983. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters* 1(6):327–332.

Paul, M., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM*.

Pontes, T.; Vasconcelos, M.; Almeida, J.; Kumaraguru, P.; and Almeida, V. 2012. We know where you live: Privacy characterization of foursquare behavior. In *UbiComp '12*.

Sadilek, A.; Kautz, H.; and Bigham, J. 2012. Finding your friends and following them to where you are. In *Proceedings of WWW*, 723–732. ACM.

Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM* 11:329–336.

Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2012. Geography of twitter networks. *Social Networks* 34(1):73–81.

Turney, P., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1):141–188.

Vardi, Y., and Zhang, C. 2000. The multivariate l_1-median and associated data depth. *Proceedings of the National Academy of Sciences* 97(4):1423–1426.

Vincenty, T. 1975. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review* 23(176):88–93.

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