Location Prediction with Communities in User Ego-Net in Social Media

Paul Wagenseller III, Adrian Avram, Eric Jiang, Feng Wang, and Yunpeng Zhao

Arizona State University

***Abstract*—Social media embed rich but noisy signals of physical locations of their users. Accurately inferring a user's location can significantly improve the user's experience on the social media and enable the development of new location-based applications. This paper proposes a novel community-based approach for predicting the location of a user by using communities in the ego-net of the user. We further propose both geographical proximity and structural proximity metrics to profile communities in the ego-net of a user, and then evaluate the effectiveness of each individual metric on real social media data. We discover that geographical proximity metrics, such as *average/median haversine distance* and *community closeness*, are strong indicators of a good community for geotagging. In addition, structural proximity metric *conductance* performs comparable to geographical proximity metrics while *triangle participation ratio* and *internal density* are weak location indicators. To the best of our knowledge, this is the first effort to infer the physical location of a user from the perspective of latent communities in the user's ego-net.**

***Index Terms*—community detection, Ego-net, geographical proximity, structural proximity, Twitter**

I. Introduction

As social media get increasingly popular and are used to host innovative applications to address real-world challenges, accurately resolving the real-world geographic location of a social media user is critical. For example, public health applications such as flu surveillance systems [1] [2], which use social media signals to infer the users' real-world flu status, require accurate user Geo-location in order to make location-related actions. Additionally, accurate user location prediction can significantly improve a user's experience through location-based personalization such as localized content and location-aware recommendations. It can also enable new location-based applications. In reality, today's social media embed rich geographic data. A user can provide their home location in the location field of their user profile. Tweets may include metadata such as the time and GPS-coordinates associated with the tweet. The content of the tweet may also indicate a location. However, the location information in social media is usually noisy and sparse. The location information in the user profile can be arbitrary strings instead of meaningful locations, and tweet content can be ambiguous as well. Moreover, [3] reports that only about 1 to 3% of all tweets are tagged with geographical coordinates as meta-information in 2013. Therefore, to correctly determine the real-world Geo-location of a social media user is a challenging problem.

Extensive research has investigated the inference of a user location by studying social relationships between users in social media [4] [5] [6] [7] [8] [9] and utilizing the content of their tweets [10] [11]. Several survey articles [12] [13] also systematically compared different approaches of location inference.

People form online social relationships for different reasons. Some of them are between relatives or acquaintances and live close. Others are formed by people from all over the world who have never met each other, but simply share interests in politics, sports, etc. This is especially true for Twitter since its main purpose is for information dissemination. These different contacts play different roles towards inferring physical location of a target. For example, a family community should provide more value in determining a user's location than a general political community that the user belongs to. A challenging question is, how to distinguish such relationships and discover the relationships that help the inference of a user's location?

In this paper, we propose a novel methodology based on *latent communities* in social media to determine the locations of social media users. Community is a structure within which users are densely connected to each other while being more loosely connected to the outside world. By identifying the communities that a user belongs to, we can divide their contacts into groups of different purposes and choose the best community to predict the location of the user.

Our methodology contains three steps. First, collect and build the ego-net of individual users. The ego-net is a network evolving around a user, called the focal node. It includes the focal node, all its neighbors, and the ties between those neighbors. Second, remove the focal node from the ego-net and then apply community detection algorithms to discover all communities in the ego-net. We adopt a well-accepted algorithm Infomap [14] to discover the latent communities. As mentioned in [15], Infomap can produce communities with desirable sizes for social media. Third, we define geographical proximity and structural proximity metrics to profile a community and choose the community which can provide the most reliable location information for the user. The location of the user is then predicted by the geometric median of the locations of all users in the best community. We propose metrics such as *average/median haversine distance* and *community closeness* to measure the geographical proximity of a community. Family communities usually have a high geographical prox-

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

imity since family members tend to live close to each other. We also use structure-based metrics including *conductance*, *triangle participation ratio*, and *internal density* to measure the structural properties of a community. In our experiments, we investigate the impact of a community's structure and geographical proximity on the accuracy of inferred locations. We discover that geographical proximity metrics including *average/median haversine distance* and *community closeness* are strong indicators of a good community for geotagging. In addition, structural proximity metric *conductance* performs comparable to geographical proximity metrics while *triangle participation ratio* and *internal density* are weak location indicators.

The contribution of this paper are twofold: 1) we propose a novel community-based approach for location inference in social media. To the best of our knowledge, this is the first time that a community-based approach is introduced and systematically evaluated on a real data set. 2) We propose metrics to measure the geographical proximity and structural proximity of a community and carry out extensive experiments to evaluate the effects of these metrics on the accuracy of location inference.

The remainder of this paper is organized as follows. Section II describes the proposed community based location predication approach and the metrics for community proximity and community structure. Section III explains our data collection and filtering process and shows the statistics of the data. Experimental results are given in Section IV. Section V summarizes related work in location inference in social media and Section VI concludes this paper and outlines our future work.

II. Community Based Location Prediction Approach

In this section, we first illustrate the motivation of the community-based location prediction approach, and then propose metrics to profile a community and decide the coordinates of the focal node of the community.

A. Communities in User Ego-Net

Social media users form online social relationships which can be used as indicators for predicting a user's physical location. However, social relationships are formed for a variety of reasons. Family and friends in real life will be natural friends on social media, while relationships can also form when two social media users share similar interests in politics or sports, even if they are from distant physical locations and have never met each other. Different contacts play different roles towards inferring the physical location of a target. We propose a novel methodology based on *latent communities* in social media to determine the locations of social media users. To be more specific, our approach utilizes the social relationships embedded in the communities of the ego-net of a user and then choose the best communities to infer the geographic location of the target node. The ego-net is defined as the network evolving around a focal node. It includes the focal node and all its neighbors and the ties among all its neighbors. We choose the ego-net because the location of a user is more related to the locations of their friends and how their friends are connected among themselves, rather than the locations of others that are not directly related to the user. Communities in an ego-net are then identified with the directed Infomap [14] algorithm, which is a random walk algorithm based on information theory and produces high quality communities [15].

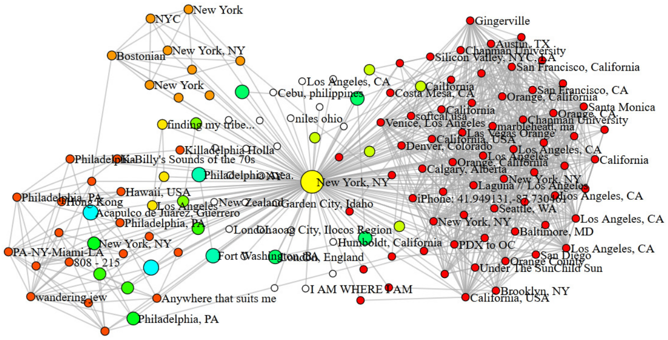


Fig. 1: Ego-net of a user in New York, NY

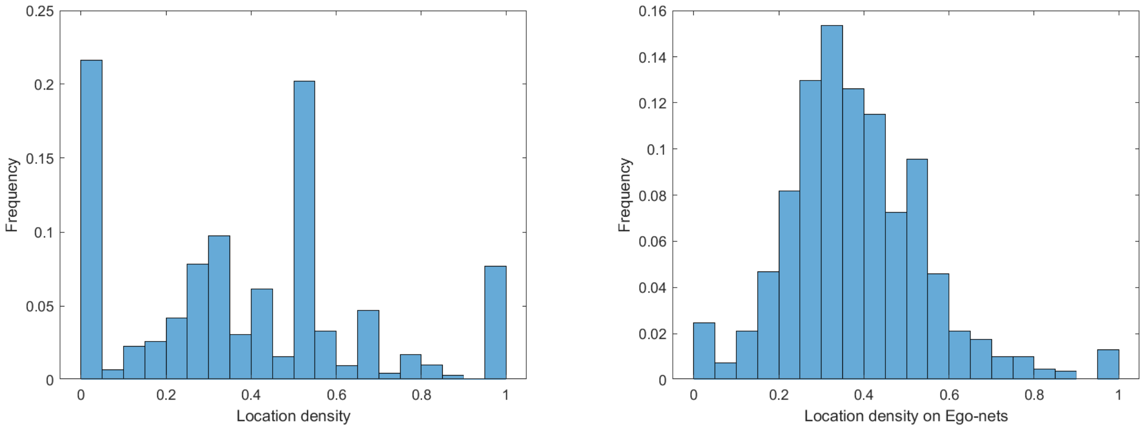
Figure 1 illustrates the motivation of our community-based approach. The central yellow node is a user who lives in New York, NY with 10 associated communities. The largest community colored red is around Los Angeles, California. The second largest community colored dark orange is in Philadelphia, PA, followed by the light orange community in New York, NY. If the location of the focal node is determined based on the most common locations of its contacts, this user would be geotagged to Los Angeles, California. With community-based approach, by choosing the communities with the best *c*ommunity closeness value, we are able to infer the user's correct location in New York, NY.

B. Community Goodness Metrics

We propose to measure the goodness of a community from two perspectives: community geographical proximity and community structural proximity. Community geographical proximity is measured with *a*verage/median haversine distance and *c*ommunity closeness. Community structural proximity is measured with *c*onductance, *i*nternal density, and *t*riangle participation ratio. We list and describe these metrics as below:

Average/Median haversine distance: The haversine distance calculates the arc distance between two points on a sphere given their longitudes and latitudes. For every pair of users in a community, we calculate the haversine distance between them, filter the long-distance outliers using the Median Absolute Deviation (MAD), and calculate the average/median haversine distance of all remaining distances.

Community closeness: This metric is the ratio of the pairwise users in the same community who are within 25 miles from each other. Equation (1) gives the formula of the Community Closeness (CC) of community C, where d is a distance threshold and $|l_u - l_v|$ denotes the haversine distance between two users u and v in C.



(a) Location Density in Communities

(b) Location Density in Ego-Nets

Fig. 2: Histograms of Location Frequencies in Ego-Nets

$$CC_C = \frac{|u, v \in C : |l_u - l_v| \leq d|}{|C||C-1|} \tag{1}$$

Conductance, internal density, and triangle participation ratio: These are three common metrics to measure the closeness of a community in terms of the structure of its social relationships. Conductance is a metric that takes both external and internal connections of a community [16] into consideration. It is defined as the ratio of the number of edges between the community and its complement over the sum of degrees of nodes within the community. Internal density [16] is a measure of the internal structure within a community. It is defined as the number of edges in the community divided by the total possible edges in the community. Triangle Participation Ratio (TPR) [17] is defined as the number of nodes in a community that form a triad, divided by the total number of nodes in the community.

C. Location Prediction

After the best community is chosen based on a goodness metric, the geographic coordinates of its focal node are predicted by the geometric median of all users in that community. The coordinates are then mapped to a (city, state, country) tuple using Google's Reverse Geocoding API. The geometric median of a community is defined as follows: Given a community and a set of coordinates (latitude, longitude) of all users in the community, the geometric median is the point that has the minimal sum of haversine distances to all the other nodes in the community. We adopt Weiszfeld's algorithm [18] for this purpose. This algorithm iteratively re-weights least squares and is robust to location outliers. The python implementation of the Weiszfeld algorithm has a time complexity of $O(nI)$ where n is the number of points and I is the number of iterations. The number of iterations is set to 50 by default.

III. Datasets

In this section, we first explain our approach of ego-net collection and filtering then present the density of location information in the filtered ego-nets.

A. Ego-Net Collection

We collected the ego-nets of $1,317$ Twitter users during winter 2018 using Twitter's REST API. The Twitter users were randomly selected from a flu tweet stream gathered during the 2017-2018 Flu season. We limited our collection process to users with less than 500 followers and 500 friends. This restriction was to reduce the number of API calls because Twitter imposes rate limits on their REST API. This decision is also consistent with the findings in [4], which state that users who follow too many others or have too many followers are not good sources for geo-tagging.

B. Ego-Net Filtering

Due to the extreme ambiguity of the self-reported user profile location fields, we use the following steps to further filter the $1,317$ collected ego-nets: 1) remove the ego-nets whose focal nodes do not provide a (city, state) pair in the location field of their user profile. This reduced our data set to $1,088$ ego-nets with 82,511 users; 2) remove the ego-nets whose focal nodes do not have valid GPS coordinates returned by Google Geocoding API. This reduced our data set to $1,064$ ego-nets; 3) remove the ego-nets whose focal nodes do not have any neighboring node in the same city, state, or country as themselves. We believe that it is unlikely for a user's profile location to be genuine if they do not share a city or state with any of their friends. This reduced our data set to 936 ego-nets with 76,167 users whose focal nodes have at least one state-level matching neighbor and 607 ego-nets with 54,113 users whose focal nodes have at least one city-level matching neighbor. The 607 city-matched ego-nets contains 29,426 users with city-level location and 723,017 edges between them.

C. *Ego-Net Location Density*

Figure 2 illustrates the sparsity level of location information in communities and ego-nets at the city level. Fig 2a is a histogram of the distribution of location densities on communities. Generally speaking, the histogram is right-skewed, which means the overall frequency of communities with more than half of users with locations with being disclosed is smaller than the frequency of communities with less than half of such users. Furthermore, the graph contains a number of spikes. For example, there are 21.66% of communities with location density being exactly 0, and 18.97% of communities with location density being 0.5. This is due to the average community size being low (on average approximately 5 users for each community). Fig 2b is a histogram of the distribution of location densities on ego-nets. The histogram is right-skewed, consistent with the pattern in the left panel. Nevertheless, the plot is smoother because the locations hardly concentrate on a specific ratio with the sizes of ego-nets being larger.

IV. Performance Analysis

In this section, we investigate the effects of using different proximity metrics to predict a user's (focal node's) physical location. The optimal approach and two baseline approaches are defined as below. *Nearest Community:* choose the community that is closest to the actual location of the focal node as the best community. This is the optimal approach. *Geometric Median:* use the geometric median of all the user's neighbors as the focal node's location. *Random Neighbor:* use a randomly chosen neighbor's location as the focal node's location.

A. *The Effect of Geographical Proximity Metrics and Structural Proximity Metrics*

Figure 3 compares the impact of geographical proximity and structural proximity metrics, where the x-axis is the distance between a user's predicted and actual location in miles, and the y-axis is the fraction of users geotagged to a location within x miles of their actual location. As can be seen in Fig 3a, the median haversine distance metric outperforms other metrics in the range of 1 to 10 miles. It can geolocate 52% of users within 1 mile and 76% of users within 10 miles. There is not much difference between community closeness, average haversine distance, and geometric median in this range. In the 10-30 mile range, we find that the community closeness metric becomes the best, after which geometric median performs slightly better until the distance becomes approximately 70 miles. All of these metrics performed significantly better than the random neighbor baseline.

Fig 3b shows the effect of the structural proximity metrics of conductance, triangle participation ratio, and internal density. We find that conductance outperforms other structural metrics and has comparable performance to geographical proximity metrics. For example, it geotags 48% of users within 1 mile and 60% users within 10 miles, which is close to the performance of average haversine distance and community closeness metrics. A surprising discovery is that random neighbor baseline outperforms internal density, which measures the connectivity within a community. Linear regression also indicates a strong negative correlation between the prediction accuracy and internal density. Intuitively, users that are densely connected to each other in a community should have a higher chance to live close. We plan to carry out in-depth statistical analysis on our dataset to understand this scenario better.

B. *The Effect of Community Closeness Threshold*

We plotted the community closeness distance thresholds of 0, 1, 5, 10, 15, 20, 25, 30, 50 and 100 miles respectively, in Fig. 4a in order to understand what is a proper distance threshold to define community closeness. Fig. 4b includes representative thresholds to allow clear comparison. At 1 mile, threshold 5 performs the best by geotagging 51% of users within 1 mile. Threshold 5 quickly starts to perform worse than the others after approximately 7 miles. Threshold 0 and threshold 30 become the worst from 7 to 100 miles and threshold 5 tends to stay between all other thresholds. We find that the threshold of 50 miles performs the best after 15 miles and remains the best until all the closeness thresholds begin to merge. Specifically, threshold 50 geolocates 70% of users within 15 miles and 83% of users within 50 miles which is about the size of a large city. Therefore, we have chosen threshold 50 as the standard distance threshold for community closeness.

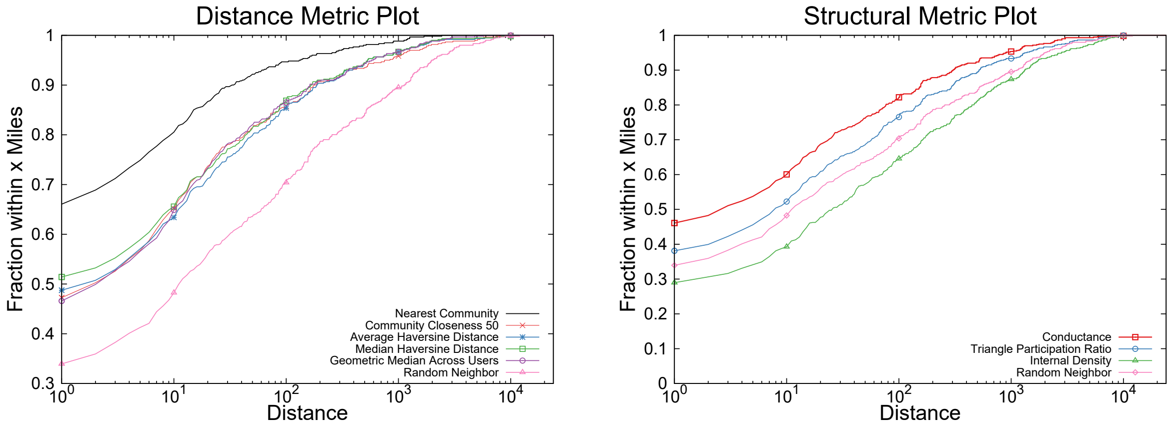
C. *Visualization of Geotagging Results*

We visualize the geotagging result for a Twitter user that lives in Ormond Beach, FL on a map in Figure 5. The focal node location is marked by a green *h*ouse symbol. This user has 19 different communities in their ego-net, and these communities contain users spread all over the world. We plot the geometric medians of each nearby community on the map by using blue markers. As can be seen, the ego-net of this user has communities in Atlanta, Georgia, New Orleans, Louisiana, and many locations in Florida, such as Tampa. The furthest community (not shown on the map) is located in Romania. We then highlight the predicted location using community closeness metric as a *r*ed star marker. Both median haversine distance and average haversine distance output the same location, marked by a *p*urple star marker. We find that the community closeness metric performs the best by geotagging the user to Deltona, FL, which is about 20 miles from the user's actual location.

V. Related Work

Existing research on social media user location inference can be divided into three categories: social relationship based approaches [5] [4] [6] [7] [8] [9], content-based approaches [10] [11], and a few comprehensive surveys [12] [13].

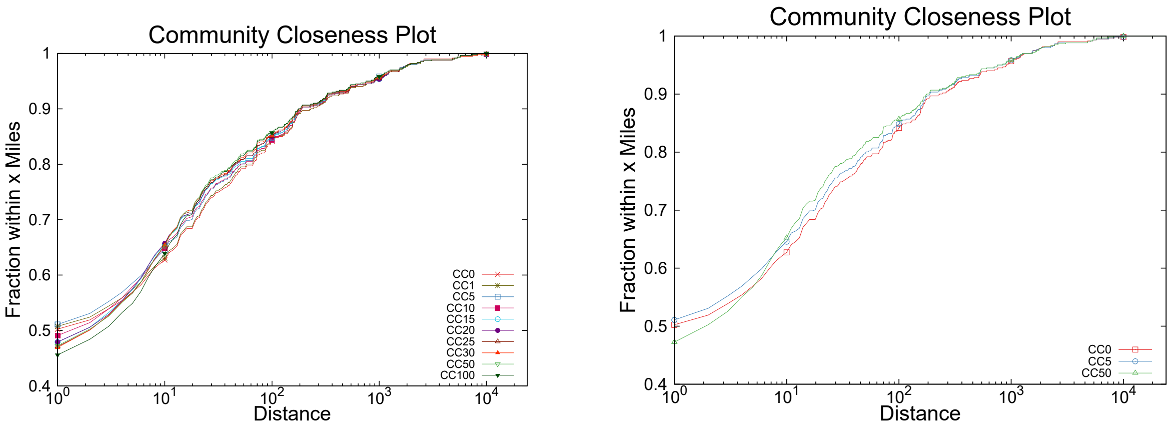
[5] is the earliest effort to study the relationship between the distance of two users and their friendship and is based on the precise street addresses of Facebook users. It gave empirical evidence that the probability of friendship decreases as



(a) Geographical Proximity Metrics

(b) Structural Proximity Metrics

Fig. 3: Geographical Proximity Metrics vs. Structural Proximity Metrics



(a) All thresholds

(b) Thresholds 0, 5, and 50

Fig. 4: Community Closeness

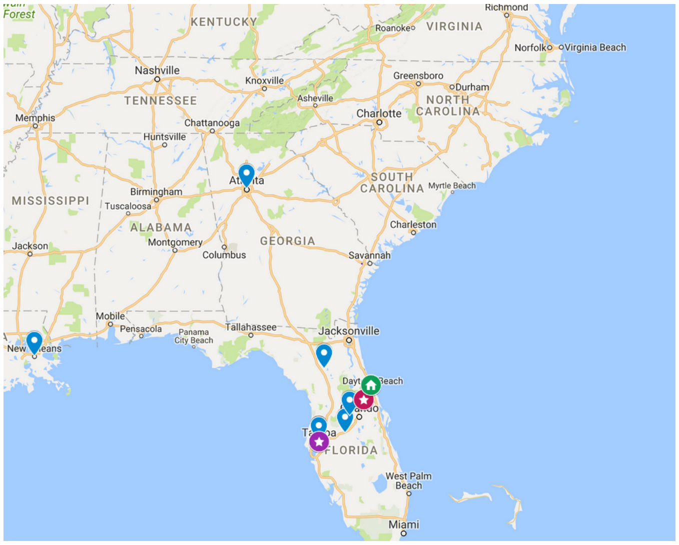


Fig. 5: Map of Geotagging Results of a Twitter User

distance increases in Facebook. It then proposed a maximum-likelihood approach which associated a probability calculated based on the distance of two endpoints of an edge with each edge, calculated the likelihood of all possible locations of a user, and chose the location with maximum likelihood as the location of the user. It further improved the algorithm by pruning the geographic search space based on the observation that the likelihood of a location is almost always maximized at the location of a friend of the user. [4] considered the strength of the social relationship between users for improved location estimation. It identified several factors which can reveal the distance between a pair of users, such as the number of followers, the reciprocity of the relationship, the locality of a user called local the contact ratio, and whether the location is a big or small city. It used these factors to train a decision tree to distinguish between pairs of users who are likely to live nearby and pairs of users who are likely to live in different areas. The results of the decision tree served as the input to a maximum likelihood estimator to predict a user's location as in [5]. [6] introduced the problem of global inference of location, that is, given a network with a small subset of users with initial locations, decide the locations of all users in the network through iteratively using the inferred location to predict for the next round in a multiple pass approach. It built the mention network in Twitter to simulate a user's social relationship, and used the geometric medium of a set of GPS-tagged tweet locations to decide the location of the user, then it proposed spatial label propagation algorithm to infer the location of all users. It discovered that geometric median approach performs well.

[10] and [11] proposed content-based approaches for geotagging which rely on tweet contents for location prediction. Given a set of prespecified cities, [10] counted the frequency for each word that the word is issued by a user located in a specific city. Then given the set of words extracted from a certain user, the aforementioned paper predicted the likelihood of the user being located in a city by averaging the frequencies for each word, weighted by the appearance frequency of words. This paper then improved the prediction accuracy by two additional techniques: 1) select words with a strong local geo-scope using a probabilistic model of spatial variation and certain machine learning techniques. 2) Smooth the distribution of words over cities by variants of smoothing techniques to overcome the location sparsity of words in tweets. [11] used a language modeling approach to model the likelihood of a tweet being issued from a certain location given the content of the tweet. It is worth noticing that both methods in [10] and [11] can only make predictions from a prespecified set of locations since their methods compute the posterior probability for each location given the tweets.

Several survey articles systematically compared different approaches. [12] grouped the methods into home location predication, tweet location predication, and mention location prediction categories and summarized over 50 studies on the inference of the location of a Twitter user, the location of a tweet, and the location mentioned in a tweet. [13] conducted an empirical comparison of nine location inference approaches. All nine methods were performed on the same dataset with the ground truth and were evaluated with the same standardized performance metrics.

Our work belongs to the social relationship based approach category and we are the first effort to introduce a community based approach for location inference.

VI. Conclusions and Future Work

In this paper, we first introduced a community-based approach for inferring a user's geographic location. We then proposed both geographical proximity and structural proximity metrics to profile communities in the ego-net of a user and evaluated the effectiveness of each individual metric on real social media data. We discovered that geographical proximity metrics, average/median haversine distance and community closeness, and structural proximity metric conductance are all strong indicators of a good community for geotagging while triangle participation ratio and internal density are weak location indicators. In the future, we will propose and evaluate new metrics for geographical proximity, adopt machine learning models to use a combination of metrics to determine the best community, continue the ego-net collection and carry out experiments on larger data sets, and study the global location inference problem from the perspective of communities.

Acknowledgment

This project is supported by NSF grant ATD #1737861.

References

[1] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, "Carmen: A Twitter Geolocation System with Applications to Public Health," in *Workshops at the 27th AAAI Conference on Artificial Intelligence*, 2013.

[2] J. Chon, R. Raymond, H. Wang, and F. Wang, "Modeling Flu Trends with Real-Time Geo-tagged Twitter Data Streams," in *Wireless Algorithms, Systems, and Applications, Lecture Notes in Computer Science*, vol. 9204, Springer, 2015.

[3] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E, Shook, "Mapping the global Twitter heartbeat: The geography of Twitter," in *First Monday*, vol. 18, no. 5, 2013.

[4] J. McGee, J. Caverlee, and Z. Cheng, "Location Prediction Social Media Based on Tie Strength," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM)*, 2013, pp. 459-468.

[5] L. Backstrom, and E. Sun, C. Marlow, "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity," in *Proceedings of the 19th international conference on World wide web (WWW)*, 2010, pp. 61-70.

[6] D. Jurgens, "That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships," in *ICWSM '13*, 2013, pp. 273-282.

[7] R. Li, S. Wang, H. Deng, R. Wang, and K. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *KDD '12*, 2012, pp. 1023-1031.

[8] S. Abrol, L. Khan, and B. Thuraisingham, "Tweeque: Spatial-Temporal Analysis of Social Networks for Location Mining Using Graph Partitioning," in *Proceedings of International Conference on Social Informatics*, 2012, pp. 145-148.

[9] S. Abrol and L. Khan, "TweetHood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining," in *Proceedings of IEEE Second International Conference on Social Computing (SocialCom)*, 2010, pp. 153-160.

[10] Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users," in *Proceedings of the 19th ACM international conference on Information & Knowledge Management (CIKM)*, 2010, pp. 759-768.

[11] S. Kinsella, V. Murdock, and N. O'Hare, "I am Eating a Sandwich in Glasgow?: Modeling Locations with Tweets," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, 2011, pp. 61-68.

[12] X. Zheng, J. Han, and A. Sun, "A Survey of Location Prediction on Twitter," *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[13] D. Jurgens , T. Finnethy , J. Mccorriston , Y. Xu , and D. Ruths, "Geolocation prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice," in *ICWSM '15*, 2015, pp. 188-197.

[14] M. Rosvall and C. T. Bergstrom, "Maps of Random Walks on Complex Networks Reveal Community Structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4. pp. 1118-1123, 2008.

[15] P. Wagenseller, III, F. Wang and W. Wu, "Size Matters: A Comparative Analysis of Community Detection Algorithms," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 951-960, 2018.

[16] J. Leskovec, K. J. Lang, and M. Mahoney, "Empirical Comparison of Algorithms for Network Community Detection," *Proc. of the 19th Int. Conf. on World Wide Web*, 2010, pp. 631-640.

[17] J. Yang and J. Leskovec, "Defining and Evaluating Network Communities Based on Ground-Truth," in *ICDM '12*, 2012, pp. 745-754.

[18] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Annals of Operations Research*, vol. 167, no. 1, pp. 7-41, 2009.