Introduction

Social media provides a rich environment where data is often posted regarding the health of an individual. This information can be used to accurately predict where epidemics are occurring globally and regionally. This data can aid in viral outbreak detection. In order to make such systems accurate there is the need to know where an individual is located geographically. Online environments such as Twitter allow a user to specify any location they wish, which can lead to a very noisy signal regarding where outbreaks are occurring. In this paper we hope to resolve this issue in the following way:

- Collect the structural and location field information contained within users egonet who is tweeting about being sick with the flu.
- Apply the directed Infomap community detection algorithm on the egonet of the user to find latent communities.
- Assign a location for each community using the Weiszfeld algorithm which iteratively reweights least squares.
- Propose new metrics to predict the geographic location of a user based on information contained within these communities.
- Evaluate the effectiveness of different metrics on geographical location prediction.

- **Dataset** – 1,317 randomly selected Twitter user egonets were collected for several months during 2018. We limited our selection to users with fewer than 500 friends/followers to avoid Twitter API limits and to avoid collecting celebrities. We had to filter this data set to remove self-reported user locations with nonexistent locations using Google Geocoding API reducing it to 1,088 egonets.
- **State Level** – we further filtered our dataset to users with at least one neighbor with a location matching their state. This reduced our data set to 936 egonets containing 78,167 users.
- **City Level** – next we filtered our dataset to users with at least one neighbor who has a location matching the focal nodes city reducing our data set to 607 egonets with 54,113 users.

Motivation

Here we demonstrate the motivation behind our method. We have a user who lives in New York, NY. They have 10 associated communities the largest of which is centered around Los Angeles, CA. The second largest in dark orange is in Philadelphia, PA, followed by the light orange community in New York, NY. Using our community-based approach combined with our community closeness metric we can accurately geotag the user to the New York, NY community.

Visualization of Geotagging Results

The Visualization of Geotagging Results shows two visualizations:

- Figure 2 – same user as in Figure 1.
- Figure 3 – a user with 19 communities located in Orlando Beach, FL.

Here we have 2 separate users figure 2 is the same user as in figure 1. Figure 2 is a user with 19 communities the furthest of which is in Romania. The green house symbol is where the user lives, purple star is the location predicted by median/average haversine distance. The red star is the predicted location using our community closeness metric.

Community Goodness Metrics

- **Nearest Community** – the community closest to the focal node if you knew the focal nodes location. You can think of this as the optimal result of our community-based method.
- **Average and median haversine distance** – the haversine distance is the arc distance between two points on a sphere given longitude and latitude coordinates. We calculate haversine distance between each pair of users and remove outliers using the Median Absolute Deviation.
- **Geometric Median Across Users** – result of running the Weiszfeld algorithm using all user locations other than focal node to predict location.
- **Random Neighbor** – randomly choose a neighbor as the predicted location.

**Equation 1: Community Closeness**

\[ C_C = \frac{\text{number of edges between two users in } C}{|C| - 1} \]

**Community Closeness** – the ratio of the pairwise users in the same community who are 25 miles from each other. Where \( d \) is the distance threshold, \( |u_i - l_j| \) is the haversine distance between two users in \( C \).

- **Conductance** – 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** – the number of edges in the community divided by the total possible edges in the community.
- **Triangle Participation Ratio** – number of nodes in a community that form a triad, divided by the total number of nodes in the community.

Understanding Community Closeness

- In order to better understand the community closeness threshold, we examined the result at thresholds of 0, 1, 5, 10, 15, 20, 25, 30, 50, 100. At 1 mile we find that threshold 5 performs the best accurately geotagging 51% of the users. Threshold 5 starts to become worst after 7 miles. Threshold 0 and threshold 30 become worst from 7 to 100 miles. We find that a threshold of 50 becomes the best after 15 miles and remains the best until all thresholds converge.

Conclusion and Future Work

In this project we defined a new way of predicting user location based on latent communities in a user egonet. We also proposed and evaluated a variety of metrics for choosing the best community for user location prediction. In the future, we plan to investigate new machine learning methods for choosing the best community.

Reference