Using a community-based approach, we geolocated social media user locations (city, state) to within a 10-mile radius with an accuracy of 65%, where the nearest community accuracy at this distance is 80%. This translates to a community-based accuracy of approximately 4 out of 5 people being correctly geotagged to within 10 miles when using their communities. This was possible using the Infomap community detection algorithm, combined with newly defined metrics, which aids in viral outbreak detection and response.

**Problem Statement**

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. 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.

**Methodology**

- **Nearest Community** – optimal result of community-based approach.
- **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.
- **Random Neighbor** – randomly choose a neighbor as the predicted location.
- **Geometric Median Across Users** – result of running the Weiszfeld algorithm using all user locations other than focal node to predict location.
- **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.
- **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, \( l_u - l_v \) is the haversine distance between two users in \( C \).

\[
CC = \frac{\sum_{u \in C} \sum_{v \in C} \text{if} \ l_u - l_v \leq d \ \text{then} \ 1 \ \text{else} \ 0}{|C|^2 - |C|}
\]

**Definitions & Results**

Using a community-based approach combined with the community closeness metric, we accurately geotag the user to the New York, NY community. The colors indicate the different communities, and the large yellow node is the user being predicted.

**References**