Data Overview and Motivation

The data measures the frequency of long ties in different geographic locations of the US and Mexico. Long ties, the edges in the connectivity network without any mutual contact, have been shown to play an important role in access to economic opportunities. Clustering coefficient is yet another related measure that captures the level of structural diversity in a network. Both these measures are provided at two different levels: zip codes and counties (or municipalities in Mexico). The data is described in more detail in the accompanying paper "Long ties, disruptive life events and economic prosperity."

Overview of Methodology

We use the commenting activity on the Facebook platform to construct a network among all users in the US (or Mexico). In this network, an edge exists between two users if they have reciprocally posted comments to each other during a fixed snapshot period (180 days prior November 2021). The edge weight from user x to user y is the fraction of comments x has sent y among all comments x has made toward all of its neighbors. In other words, edge weights are the normalized count of comments made along each edge. Edge weights adjacent to each node sum up to 1. Even though all edges are reciprocal, edge weights along a single pair vary in each direction.

To measure the frequency of long ties in each geographic location, we first construct a network corresponding to each geographic unit. In this context, geographic units refer to either zip codes or counties in the US and zip codes or municipalities in Mexico. For simplicity from now on, we only refer to zip codes as a geographic unit, while keeping in mind that similar networks can also be constructed at the county or municipality level. Each zip code network is a subnetwork extracted from the full country network. The nodes in each zip code network consists of all users who reside in that zip code and all their contacts. The network contains all the edges among these nodes, which includes the edges among nodes who reside outside the zip code but have a connection to at least one user in the zip code.

For each zip code network, we construct two related measures of structural diversity. First, we use the fraction of long ties among all edges that involve at least one node inside the zip code as our primary measure of zip code structural diversity as defined below.

$$l_z = \frac{\sum_{i \in Z} \sum_{j \in N_i} I(\mathrm{mc}_{ij} = 0)}{\sum_{i \in Z} |N_i|}$$

Where Z is the set of nodes residing inside the zip code, N_i is the set of i's neighbors, regardless of whether they reside inside or outside the zip code, $I(.)$ is the indicator function and mc_{ij} is the number of mutual contacts between i and j. It should be noted that long ties among two users who both reside outside the zip code are not included in the definition above, since we are interested in measuring the benefits of long ties conferred to users inside the zip code.

Second, we use the clustering coefficient, or the fraction of closed triad in the zip code network, as the secondary measure of structural diversity. Similar to the case of the fraction of long ties above, we only consider triads that involve at least one node inside the zip code and discard triads whose all three nodes reside outside the zip code. With this explanation, we can measure the clustering coefficient in each zip code network as below.

$$c_z = \frac{\sum_{i \in Z} \sum_{j \in N_i} mc_{ij}}{\sum_{i \in Z} |N_i|(|N_i| - 1)}$$

A separate set of indicators per each zip code involves the weighted version of the variables above. These weighted measures enable us to examine the relationship between the strength of long ties and outcome variables. First, the weighted fraction of long ties, defined as below, captures the fraction of total edge weight on long ties from the total among all edges that involve at least one user in the zip code.

$$l_z^w = \frac{\sum_{i \in Z} \sum_{j \in N_i} w_{ij} I(\mathrm{mc}_{ij} = 0)}{|Z|}$$

Where w_{ij} corresponds to the (normalized) edge weight from node *i* to node *j*. It is important to note that w_{ij}'s sum up to 1 within each node *i*. This is why the denominator in the above equation is the count of nodes in the zip code. The second measure is the weighted clustering coefficient which simply weighs each triad by the weight of edges adjacent to the node(s) inside the zip code as defined below.

$$c_z^w = \frac{\sum_{i \in Z} \sum_{j \in N_i} w_{ij} mc_{ij}}{|Z|(|N_i| - 1)}$$

In addition to the measures above, we also release the average degree of users who reside inside the zip code. This would allow us to adjust for network size when conducting statistical analysis on the networks.

Privacy Preserving Steps

To ensure individual users cannot be re-identified, we take several steps to preserve user privacy.

1. As the first step, we have not released any data on geo units that are too small. Network measures from a geographic unit (e.g. zip code or county) will be removed from the data release if 1) it has less than 150 users or nodes who reside in the unit or 2) if its network has less than 1000 edges from a node inside the unit to a node outside.

The released value for each measure corresponds to the average of 10 random repetitions of the procedure described below:

2. We add N(0, 40) distributed random noise to the following measures:
    a. The count of long ties involving at least one node in the region (numerator in the fraction of long ties)
    b. The count of all ties involving at least one node in the region (denominator in the fraction of long ties)

    The value of the fraction of long ties from each random iteration is the ratio of two above noisy measures.

3. We also add N(0, 80) distributed random noise to the following measures:
    a. The count of closed triads (numerator in the clustering coefficient)
    b. The count of all triads (denominator in the clustering coefficient)

    The value of the clustering coefficient from each random iteration is the ratio of two above noisy measures.

4. We add N(0, 2) distributed random noise to the following measures:
    a. The weighted count of long ties involving at least one node in the region (numerator in the weighted fraction of long ties)
    b. The weighted count of all ties involving at least one node in the region (denominator in the weighted fraction of long ties)

    The value for the weighted fraction of long ties from each random iteration is the ratio of two above noisy measures.

5. We add N(0, 6) distributed random noise to the following measures:
    a. The weighted count of closed triads (numerator in the weight clustering coefficient)
    b. The weighted count of all triads (denominator in the weighted clustering coefficient)

    The released value for the weighted clustering coefficient is the ratio of two above noisy measures.

6. We also add N(0, 0.6) distributed random noise to mean degree of nodes inside the geographic unit.

7. Finally, in each random iteration we only include 99% of nodes from each geographic unit. In particular, for each network corresponding to a geographic unit, first we randomly remove 1% of nodes with their adjacent edges and then compute all the measures with

the added noise explained above. Note that removed nodes from one draw of a geographic unit network are independent of other draws or other geographic unit networks.

The final released measures for each geo-unit network are averages from 10 random networks each generated according to the procedure above. We also ensure that none of the above average measures can be less than 0 or greater than 1 (if they correspond to a ratio statistic) due to the random noise added, by censuring them at those values.

Codebook: Column Definition

1. **Unit_id:** The ID of the unit. If unit is zipcode, this column contains the zipcode associated with the geo-unit. If the unit is county, it contains the 5-digit FIPS code of the county for US or the municipality code for MX.
2. **Mean_degree:** Mean degree of nodes residing in the geo-unit (to other nodes residing inside or outside the geo-unit)
3. **clustering_coef:** Clustering coefficient of all triads in the network. It is the fraction of closed triangles whose at least one node resides in the geo-unit and two nodes can reside either inside or outside the geo-unit
4. **weighted_clustering_coef:** Weighted clustering coefficient of all triads in the network. It is the fraction of closed triangles whose at least one node resides in the geo-unit and two nodes can reside either inside or outside the geo-unit, weighted by fraction weight within each node.
5. **Fraction_long_edges:** Fraction of long edges, with zero mutual friends, among edges whose at least one end resides inside the geo-unit.
6. **Weighted_fraction_long_edges:** Weighted fraction of long edges whose at least one end resides inside the geo-unit. Weights are fractions that sum to 1 over all edges of each node.