Data 620 – Project 1: Centrality Analysis of a Network

Team: Data Owls

Members: Keith Folsom & Valerie Briot

Dataset: Yelp Business Review Dataset

The dataset can be downloaded from: https://www.yelp.com/dataset/download

Yelp is a crowd-sourced local business review and social networking site. Users can submit a review for products and services using a 1 to 5 start rating system.

The dataset can be downloaded in either JSON or SQL format. It is comprised of multiples files, some fairly large in size. Given the volume of data, we'll be using the SQL format to locally import the Yelp database of tables and relationships. With 156,639 businesses across the United States and globally, we will then subset the businesses to be analyzed geographically based on a zip code or small set of zip codes within a city.

For the purposes of this project, we will focus on Business, User, and Review. User based reviews such as the Yelp dataset make for a natural bipartite graph with a 2-mode network. However, in this project, we will focus on the business entity (which features more attributes) as a 1-mode graph analysis.

We will build relationships between businesses as follows: Business A is connected to Business B if the same user reviews both. The score assigned by the reviewer will become some attributes of the business node (average rating).

The categorical variables being considered on the business node are:

- **Neighborhood**: what role does the neighborhood play in the rating of a business
- **Business Category**: the Yelp dataset is comprised of businesses of all types. The most frequently used categories are Restaurant, Shopping, Food, Beauty & Spa, Home Services, Health & Medical, Nightlife, and bar.
- **Attributes**: additional attributes are captured related to the business in the dataset. These include whether or not a restaurant has takeout or indicators of parking options.

Degree of Centrality

A node with a high degree of centrality would indicate that this business has been reviewed by many people. We would expect these reviews to be more likely polarizing reviews, either on the negative or positive side. We'll look at degree centrality in this light as well analyze against the selected categorical variables. Due to the volume of data we're still formulating the approach for the categorical outcome analysis**.**