# **Data 620 – Week 4 Assignment**

**Team No. 6**

* Yohannes Deboch
* Sherranette Tinapunan

**Requirements**

Centrality measures can be used to predict (positive or negative) outcomes for a node.

Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes.

As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.)

In addition to identifying your data source, you should create a high-level plan that describes how you would load the data for analysis and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

**Data Set**

**Chicago Divvy Bicycle Sharing Data**

<https://www.kaggle.com/devisangeetha/divvy-bike-share-eda-network-analysis/data>

Divvy is Chicagoland’s bike share system, with 6,000 bikes available at 570+ stations across Chicago and Evanston. Divvy provides residents and visitors with a convenient, fun and affordable transportation option for getting around and exploring Chicago.

The team is interested in analyzing the difference between bike pickup and drop networks of male and female riders. The columns in this data set that we would be investigating are *gender*, *from\_station\_id*, *from\_station\_name*, *to\_station\_id*, and *to\_station\_name*.

The categorical variable we will use is gender.

**High-level Plan for Loading Data for Analysis**

Pandas package will be used to load the data set into a dataframe. The data set is going to be divided into male and female dataframes. Degree and eigenvector centralities are going to be calculated independently for each group. For each group, weight is going to be calculated for each edge. The weight describes the amount of traffic between pickup and drop-off station. Th weight is going to be used when calculating the eigenvector centrality for each node.

**Hypothetical Outcome that could be Predicted from Comparing Degree Centrality Across Categorical Groups**

This data set is a directed graph. It tells us the station a bike rider picks up a bike rental and the station where the bike is dropped off. By calculating the in-degree centrality of each station, we can get an idea of the number of incoming stations a particular drop off station has. A station with a high in-degree centrality could be a location where various demographic groups that come from different geographic areas converge together.

A station with high eigenvector centrality could be a station where many bike riders frequent and/or is a station that is connected to several stations with high centrality.

One possible application for this information is to help determine key advertising locations that target certain demographics.