# **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 dataframe. Degree and eigenvector centralities are going to be calculated independently for each group. A weight is going to be assigned to each link, which describes that intensity of traffic between pickup station and drop-off station. 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 certain types of demographic (coming from geographic areas where pick up stations are located) converge. This could be useful in determining prime locations for advertising. By knowing which areas where different target demographic may converge, advertising companies can plan where they may want to advertise.