Data 620 Assignment 4

Centrality Measures

Natalie Mollaghan

Violet Stoyanova

Sudhan Maharjan

**Assignment:**

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.

For this week’s assignment, you are not required to actually load or analyze the data. Please see also Project 1 below.

You may work in a small group on the assignment. You should post your document to GitHub.

**Data:**

The data that we will be using for this project comes from an API provided by Opensecrets.org. Open Secrets is a website that is considered to be the most comprehensive resource for information on federal campaign contributions and lobbying data. The data returned from the API that we will be analyzing for this assignment is a dataset on individual political members, their parties, and which PACs they accept money from. The objective is to see which PAC holds the most weight in politics and what that PAC represents.

**Analysis:**

*Step 1* – Our first step in developing comparing centrality measurements across nodes is to develop a python app that will create queries based on the parameters of outlined in the data section. In order to do this, we will access and API access code and token from Open Secret with OAuth based requests. We will need to define a variable for the access token, URL, the graph objects, request, and connections.

*Step 2* – Once we are able to establish a connection and access the metadata, we will need to structure it into a data frame for centrality measurements and build various functions for analysis. To do this, we will utilize python modules: Pandas, NetworkX, and CSV. Our data will structure will include the following elements:

*Step 3* – We will conduct an initial exploration of the data, in order to QA our data structure and make sure that this data will be able to be graphed. Next, we will calculate the centrality measurement of the chosen categories.

*Step 4* – We will build a network structure graph where each elected official will be the nodes of the network and the PACs as the edges. Another possible edge would be industry which the elected official is associated with.

**Hypothetical Outcomes**:

The objective is to see which PAC holds the most weight in politics and what that PAC represents. Outcomes could include PAC engagement across parties, how PAC contributions may impact what industry the elected official is associated with, or the overlaps in which PACs give political donations to elected officials and if there is any juxtapositions in where those PACs stand.