DATA 620 Week Four - Assignment Centrality Measures

Gabriel Santos

**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.

**An interesting set of network data using Kaggle Netflix Shows dataset.**

**Data Source:**

https://www.kaggle.com/shivamb/netflix-shows

**About the Data:**

This dataset consists of a list of TV shows and movies available on Netflix as of 2021. The dataset is collected from fixable, which is a third-party Netflix search engine.

In 2021, the Flixable.com page (https://flixable.com/netflix-museum/) published a report on the behavior of the American Netflix catalog at the beginning of 2010. In the report, we can see that TV shows have tripled, while movies have decreased by up to 30%.

**Hypothesis:**

We can measure from the data set the degree of centrality by the nodes such as TV shows and Movies.

I want to analyze if the behavior is similar for TV shows and movies made in the United States compared to those made outside the United States.

**Methodology:**

* Load the dataset by loading the kaggle api package in python and calling the library to load the dataset using the code pd.read\_csv('../input/netflix-shows/netflix\_titles.csv')
* Analyze the data, checking for missing data.
* The dataset contains historical data from 1925 to 2021 along with ratings. After measuring the degree centrality for each of the content types individually we will compare the two measures.
* Filter data for TV shows and movies made in the US, and filter data for TV shows and movies made outside the United States.
* The resulting data must be graphed, showing the nodes.
* Calculate the degree, eigenvector, Betweenness, and Closeness, of each data group.
* Carry out a statistical analysis of the data from each group. Calculate the T-score (t\_stat), degrees of freedom (df), critical values (cv), and p-values (p).
* Identify if there is a correlation between the qualifications.

**Possible conclusions:**

There is no correlation between the scores and the centrality measures obtained because there are other variables that affect these data.

The behavior is similar for TV shows and movies made in the United States compared to those made outside the United States, over the years.