DATA620 Homework 2 PT 2

6/13/21

Team Members

Bharani Nittala
John Mazon
LeTicia Cancel

Assignment

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 Sources

Proposal 1: Microsoft Graph

We will use the Graph Explorer - Microsoft Graph api using one of our email addresses to get graph data. We will likely use the GET - items trending around me since it does not pull personal information. We are not sure what exactly we will get with this API, it will be more of an exploratory analysis. It will be interesting to see how the data differs between each user if each team member uses the API with their own login, giving us the option to compare graphs.

Proposal 2: Using Kaggle Netflix Shows dataset

Data Source:

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

About the Data:

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report (https://flixable.com/netflix-museum/) which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Hypothesis:

The dataset can be leveraged to measure degree centrality by the nodes such as TV shows and Movies. It will be interesting to see how the data differs between the two popular content types (such as TV shows and Movies)

Methodology:

- We will 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')`

- Then will perform basic exploratory data analysis and treat rows with missing rows.
- The dataset contains history from 1925 to 2019 along with ratings. After measuring degree centrality for each of the content types individually we will compare and contrast the two measures
- Also, we will try to see the correlation with ratings if there exists any

Hypothetical Outcome:

There doesn't exist any correlation between (avg.) ratings and centrality measures obtained as there could be many confounding variables such as locale, language, cast, duration of the show/movie and others.