Data 620 - Week 4: Project 1 Proposal
Centrality Measures: Facebook Network

Bikram Barua and Leticia Salazar

February 19, 2023

Task:

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:

The dataset selected for project 1 was obtained from Network Repository consisting of people (nodes) and their friendship ties (edges) on Facebook. The data is available in a matrix market, typically a sparse format used to represent a matrix (.mtx file). With these types of files, the first line contains a header with information about the matrix, dimensions, and symmetry of the matrix. The following lines are non-zero entries of the matrix in row-major order, with each following line containing the row index, column index, and a value of a single non-zero entry.

Below are some specs from this dataset provided by the repository:

Nodes	18.7K
Edges	790.8K
Density	0.00454239
Maximum degree	3.2K
Minimum degree	1
Average degree	84
Assortativity	0.01805

Number of triangles	18.3M
Average number of triangles	982
Maximum number of triangles	77.2K
Average clustering coefficient	0.219051
Fraction of closed triangles	0.135615
Maximum k-core	85
Lower bound of Maximum Clique	13

Project Objectives:

- Which centrality measure would be the most relevant in the dataset?
- What would these centrality measures help us to predict?

Plan for Analysis:

Our analysis on centrality will be performed by the following:

Step 1: Leticia will download the data from the Network Repository to be able to upload it to GitHub. Since the data is available as a .mtx file different than a .csv file, we will have to format the dataset to be able to read it as a .csv file into our jupyter notebook. There is a way to also read the .mtx file directly from GitHub into a jupyter notebook which will serve as another option for us to explore.

Step 2: Once the data has been uploaded, we will clean, explore, analyze, and visualize the data to get a sense of what we are working with. Within the exploratory and analysis process we will look to answer the first question of our project objective: Which centrality measure would be the most relevant in the dataset?

Some common centrality measures include:

- Degree Centrality: based on the number of connections a node has in the network. Nodes with high degree centrality are those that have many connections to other nodes and are often considered to be important hubs within the network.
- Betweenness Centrality: based on the extent to which a node lies on the shortest path between other nodes in the network. Nodes with high betweenness centrality are often those that act as bridges between different parts of the network.

- Closeness Centrality: based on the distance between a node and all other nodes in the network. Nodes with high closeness centrality are those that are close to many other nodes in the network and are often considered to be important for efficient communication and information flow.
- Eigenvector Centrality: based on the idea that the importance of a node is proportional to the importance of its neighbors. Nodes with high eigenvector centrality are those that are connected to the other nodes with high eigenvector centrality and are often considered to be influential within the network.
- PageRank: used to evaluate the importance of web pages in a network of hyperlinks. It is based on the idea that a web page is important if it is linked to by other important web pages.

Step 3: Using networkx we will create graphs to show the connections between the people and their friendships on Facebook. At this point we should know what type of centrality measure best describes the data and be able to answer the second question of our objective: What would these centrality measures help us to predict? As a hypothesis, Leticia has predicted that the centrality measures will help us predict that eigenvector centrality will be a better fit to our dataset. Since there must be a common interest to form networks (form friendships) on Facebook the higher the number of commonalities (the more friends or interests in common) the higher the chances these two people will be connected.

References:

- Rossi, R. A., & Ahmed, N. K. (2015). *The Network Data Repository with Interactive Graph Analytics and Visualization*. Network Data Repository. Retrieved February 16, 2023, from https://networkrepository.com
- AOMAR, A. A. I. T. (2020, August 2). *Notes on graph theory - centrality measures*. Medium. Retrieved February 17, 2023, from https://towardsdatascience.com/notes-on-graph-theory-centrality-measurements-e37d2e49550a