K740) Assignment Four: Social Network Analysis

Submission is due Thursday, April 09, 2023, before midnight.

Please submit this report (50% of the grade) bundled with the Jupyter notebook (50% of the grade) you used to generate results.

# Downloading materials



1. Use codes from Jupyter notebooks (Social Network Analysis sessions) to complete this assignment. You will need to:
   * Change variable names, directory paths, and values for evaluation metrics using the information in each question.
   * Remember that column names in the classmatesChar.csv file are different from column names in the file we used in the class.

# The problem:

## Describing a Network (10).

Consider an undirected network for individuals A, B, C, D, and E. A is connected to B and C. B is connected to A and C. C is connected to A, B, and D. D is connected to C and E. E is connected to D.

1. Produce a network plot for this network. *(2%)*

See Jupyter notebook

1. What node(s) would need to be removed from the network for the remaining nodes to constitute a clique? *(1%)*

Nodes D and E should be removed.

1. What is the degree for node A? *(2%)*

2

1. Which node(s) have the lowest degree? *(1%)*

Node E has the lowest degree as 1.

1. Tabulate the degree distribution for this network. *(1%)*

See Jupyter notebooks

1. Is this network connected? *(1%)*

Since every node is connected to every other node in the network, either directly or indirectly, we can conclude that the network is connected.

1. Calculate the betweenness centrality for nodes A and C. *(1%)*

Betweenness centrality for node A: 0.0

Betweenness centrality for node C: 0.6667

1. Calculate the density of the network. *(1%)*

0.5

## Network Density and Size (10%).

Imagine that two new nodes are added to the undirected network in the previous exercise (F and G).

1. By what percent has the number of nodes increased? *(2%)*

Previously we had 5 nodes. Now we have 7 nodes. Therefore, the percent of the number of nodes increased equal to 40% (7-5/5).

1. By what percent has the number of possible edges increased? *(2%)*
   * This is an undirected network with one edge for each possible pair of nodes.

The possible edges for the original network of 5 nodes: 5\*4/2 = 10

The possible edges for the newl network of 7 nodes: 7\*6/2 = 21

(New Edges - Old Edges) / Old Edges \* 100% = (21 - 10) / 10 \* 100% = 110%

Therefore, the number of possible edges increased by 110%.

1. Suppose the new node has a typical (median) number of connections. What will happen to network density? *(2%)*

To find the median number of connections, we must, firstly, list the number of connections for each existing node:

A: 2 connections

B: 2 connections

C: 3 connections

D: 2 connections

E: 1 connection

F: not connected yet

G: not connected yet

The median value is 2, which means the new node should have 2 connections. Therefore, we can add two connections for node F as F-A, F-B, for node G as G-A, G-B.

Density (before) = 5 / 10 = 0.5; Density (after) = 9 / 21 = 0.43.

In conclusion, after adding the new node with a median number of connections, the network density has decreased from 0.5 to approximately 0.43.

1. Comment on comparing densities in networks of different sizes. *(2%)*

The density droped in previous case because, while the number of actual edges increased by 4, the number of possible edges increased by 11, which in turn has decreased the network density.

When comparing network densities across networks of different sizes, it is essential to consider that density is a relative measure, and the number of possible connections grows quadratically with the number of nodes. A higher density in a small network does not necessarily imply more interconnectedness than a larger network with a lower density. To make meaningful comparisons, consider additional network measures such as average degree or clustering coefficient, which can provide more insights into the interconnectedness and structure of the networks being compared.

1. Tabulate the degree distribution for this network. *(2%)*

See Jupyter notebooks

## Social Network Analysis (80%).

Download and use the following two .csv files to answer the questions in this section:

* + The friendship network (classmates.csv) shows who in a class with 26 students are friends with whom.
  + The characteristics of classmates (classmatesChar.csv), which contains names, gender, and the number of sessions they attended the class.

### Network characteristics *(50%)*.

1. Create the network *(7.5%)*.
   * Read classmates’ characteristics from classmatesChar.csv and add attributes to nodes.
     + Node labels will be students’ names.
     + Each node will have gender and attendance attributes.
   * Create the graph with different node colours for males and females and node size equal to students’ attendance (multiplied by 200)。

See Jupyter notebook

1. What is the density of the network? What does it mean? *(2.5%)*

The density in this network is 0.22. This means not all but just 22% of total possible connections (friendship) between classmates in the class are actually present.

1. What is the transitivity of the network? What does it mean? *(2.5%)*

The transitivity is 0.4947. In our case, the transitivity means how interconnected the students are in terms of forming triangles (three students are friends with each other).

1. Who are the top five most and the bottom five least popular people in the class? *(2.5%)*

Top 5 most popular people: 'Nora', 'Yves', 'Chloe', 'Grace', 'Quinn'

Bottom 5 least popular people: 'Ben', 'Leah', 'Riley', 'Stella', 'Emma'

1. Who are the top five people that could introduce students to one another? *(2.5%)*

They are 'Nora', 'Yves', 'Quinn', 'David', 'Frank'.

1. Ben wants to ask Riley out. However, he wants to know her better first. Who are the people that can introduce Ben to Riley? With what sequences? *(2.5%)*

Emma, David, Quinn, Grace are the people that can introduce Ben to Riley

The shortest path sequence to introduce Ben to Riley: Ben -> Emma -> David -> Quinn -> Grace -> Riley

1. What is the minimum number of people between two students that are not friends? *(2.5%)*

1 (it's 1 person but it takes 2 edges to get to them)

1. What is the maximum number of people between two students that are not friends? *(2.5%)*

4 (detail see as Jupyter Nootbook)

1. Who are the people who hang out the most? Why these students? *(2.5%)*

Quinn is the people David hand out most. David has degree of 13 and Quinn has a

degree of 11. They both are popular students who have many friends, therefore, there is a high chance for them to hang out with each other.

1. Which members are usually excluded from gatherings? Why these students? *(2.5%)*

Ben and Leah are the most usually excluded people. Because they have the lowest degree centrality. Detail see Jupyter Nootbook.

1. What is Eigenvector-based centrality, and how does it differ from the other measures we studied? *(10%)*

Eigenvector-based centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to a node's score than connections to low-scoring nodes. In other words, a node is considered influential not only by the number of its connections but also by the quality of those connections. The key difference between Eigenvector-based centrality and degree centrality is that only eigenvector-based centrality considers the importance of connections. Betweenness centrality measures the extent to which a node lies on the shortest paths between other nodes. Unlike eigenvector centrality, betweenness centrality does not consider the influence of the nodes a node is connected to but rather their position in the network's overall structure.

1. A new textbook was launched a few months ago. The publisher wants to select a couple of students in each class and give them a free copy and ask them to promote the book. Which two students should be chosen? Why these two? *(10%)*

We should choose Yves and Nora as they are the two students with the top two degree centrality. Since they have more friends in a class, they can promote the book to more students in class.

### Network Analysis *(30%)*.

1. Add the following columns to the data frame you created from classmatesChar.csv. *(10%)*
   * Closeness centrality
   * Degree centrality
   * Betweenness centrality
   * Eigenvector-based centrality
2. Using linear regression, argue what the relationship between students’ characteristics and different centrality measures is. Is there any statistically significant relationship? *(20%)*

Based on the output of the first linear regression, we can see that there is a statistically significant relationship between the dependent variable (Closeness Centrality) and the independent variable (Attendance). This is indicated by the significant p-value (0.0002) associated with the Attendance coefficient. However, the p-value for the Gender coefficient is not statistically significant (0.4772 > 0.05), suggesting that there is no significant relationship between Gender and Closeness Centrality. The adjusted R-squared value of 0.434 also indicates that the model explains 43.4% of the variability in the dependent variable. Therefore, we can conclude that there is a significant relationship between Attendance and Closeness Centrality, but not between Gender and Closeness Centrality.

Based on the second output of the linear regression, we can see that there is a statistically significant relationship between the dependent variable (Degree Centrality) and the independent variable (Attendance). This is indicated by the significant p-value (0.0003) associated with the Attendance coefficient. However, the p-value for the Gender coefficient is not statistically significant (0.5218), suggesting that there is no significant relationship between Gender and Degree Centrality. The adjusted R-squared value of 0.417 indicates that the model explains 41.7% of the variability in the dependent variable. Therefore, we can conclude that there is a significant relationship between Attendance and Degree Centrality, but not between Gender and Degree Centrality.  
Based on the third output of the linear regression, we can see that there is a statistically significant relationship between the dependent variable (Betweenness Centrality) and the independent variable (Attendance). This is indicated by the significant p-value (0.0108) associated with the Attendance coefficient. However, the p-value for the Gender coefficient is not statistically significant (0.3712), suggesting that there is no significant relationship between Gender and Betweenness Centrality. The adjusted R-squared value of 0.210 indicates that the model explains only 21.0% of the variability in the dependent variable. Therefore, we can conclude that there is a significant relationship between Attendance and Betweenness Centrality, but not between Gender and Betweenness Centrality.

Based on the output of the linear regression, we can see that there is a statistically significant relationship between the dependent variable (Eigenvector Centrality) and the independent variable (Attendance). This is indicated by the significant p-value (0.0004) associated with the Attendance coefficient. However, the p-value for the Gender coefficient is not statistically significant (0.6046), suggesting that there is no significant relationship between Gender and Eigenvector Centrality. The adjusted R-squared value of 0.400 indicates that the model explains 40.0% of the variability in the dependent variable. Therefore, we can conclude that there is a significant relationship between Attendance and Eigenvector Centrality, but not between Gender and Eigenvector Centrality.

Overall, we can conclude that there is a significant relationship between Attendance and different centralities, but there is no significant relationship between Gender and either centrality.