Assignment #3: Network Visualization

1. Transforming the dataset into a network

The dataset is represented as a directed network graph where each person is a node and a directed edge from node1 to node2 is an indication that node1 sent an email to node2. I used Gephi for the Network visualization since it was much easier to set up and offered a diverse set of functions out of the box. I also removed the date column since it wasn't offering much value in terms of generating the network.

2. Questions

2.a) Do you find communities (modularity)? If yes, how many?

There were 2 communities detected. One of the communities of size 7 consists of all the users who've talked to Mohamed excluding Dr. Hsiao and the other is of size 26 which has every other user speaking exclusively to Dr. Hsiao. I set the modularity at 0.5 since there is limited communication between Mohamed and the students hence the network is not densely connected. Mohamed's community is labeled green and Dr. Hsiao's community is labeled red.

2.b) What is the average degree centrality? What does it mean in this email network?

The average degree centrality is 1.970 which means a user who's participated in at least one email thread has been involved in 1.970 emails on an average across all threads. The degree centrality of Dr. Hsiao is 54 and Mohamed is 13 and all the other 31 users are around 1-4 resulting in a massive drop in the average degree centrality.

2.c) What is the average betweenness centrality? What does it mean in this email network?

The betweenness centrality of a Node is the number of times a node appears on the shortest path between any two nodes. Every edge in the graph is always between either or both Dr. Hsiao and Mohamed. Hence the shortest path between any two student nodes will have to pass through at least one of these users as no there are no records of any two students having communicated with each other. The betweenness centrality for Dr. Hsiao and Mohamed are 775 and 114 respectively and the **average betweenness centrality of the entire network is 26.9.**

2.d) What is the average Eigenvector centrality? What does it mean in this email network?

The average Eigenvector centrality is 0.1908. This centrality measures the influence of a node in a network. Dr.Hsiao and Mohamed and the most influential nodes in the network and students who have sent emails to both of them will have a higher eigenvector centrality score. This metric will consider the influence of the email recipient, hence a high score means a student has conversed with both Dr. Hsiao and Mohamed several times and a low score implies there's barely been any communication and if any, it's probably with just one of the two influential nodes. The nodes are also sized(small to large) and colored(light to dark) in increasing order of their Eigenvector centrality

2.e) Layout algorithm

I chose the Yifan Hu layout since it incorporates a powerful force-directed algorithm to attract nodes with edges and repel nodes without edges. The resulting layout also helps in discerning between communities easily since the students who conversed with Mohamed are shifted towards Mohamed's node and Dr. Hsiao's node is at the center since she has spoken to almost every user in the dataset.