Assignment: 3
CSE 578 Data Vizualization
Spring 2020

1. The dataset given to us is an email network. To import the dataset into Gephi to visualize it, I modified the headings 'From' and 'To' to 'Source' and 'Target' respectively so that the dataset can be imported as an edge list. I modified the date column to a date format acceptable by Gephi. This allowed me to use the dates as a timestamp. As we inherently don't provide weights to these edges and Gephi doesn't support parallel edges, multiple edges from one node to another where transformed into a weighted edge, with the weight signifying the number of emails sent from that node to a particular recipient.

2. Modularity:

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

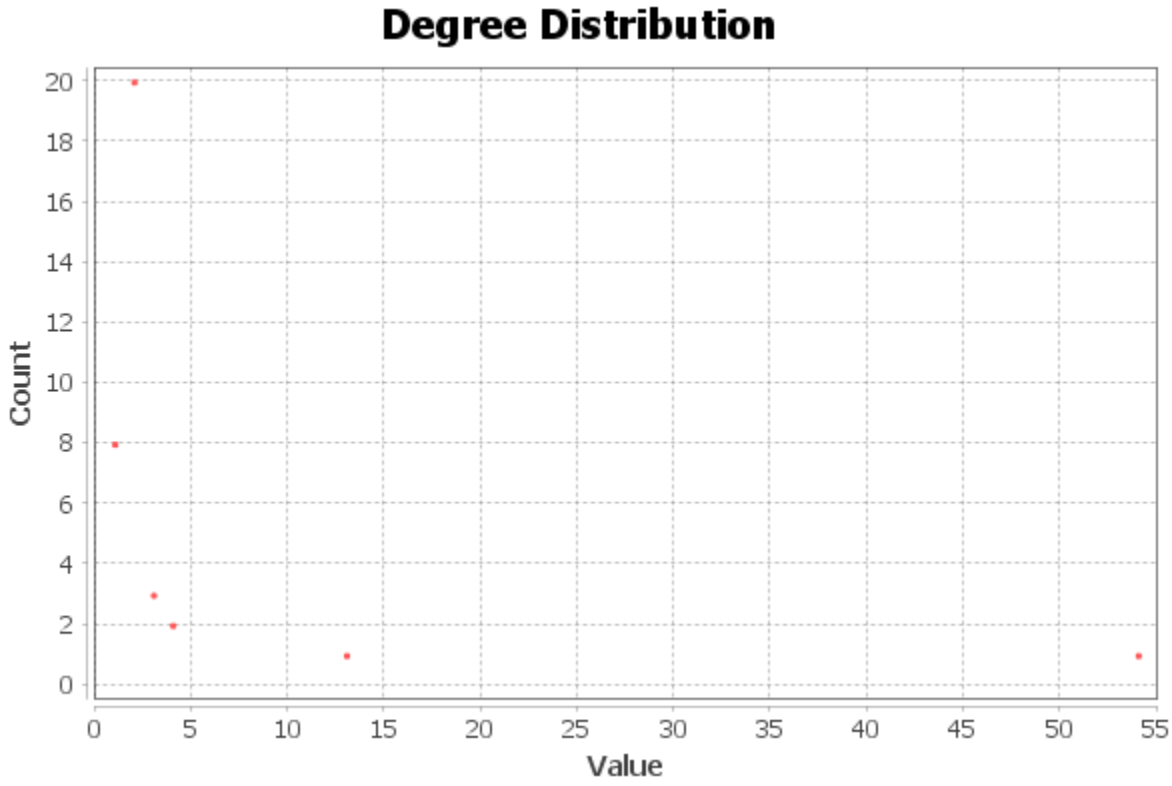
Modularity: 0.088
Modularity with resolution: 0.088
Number of Communities: 2

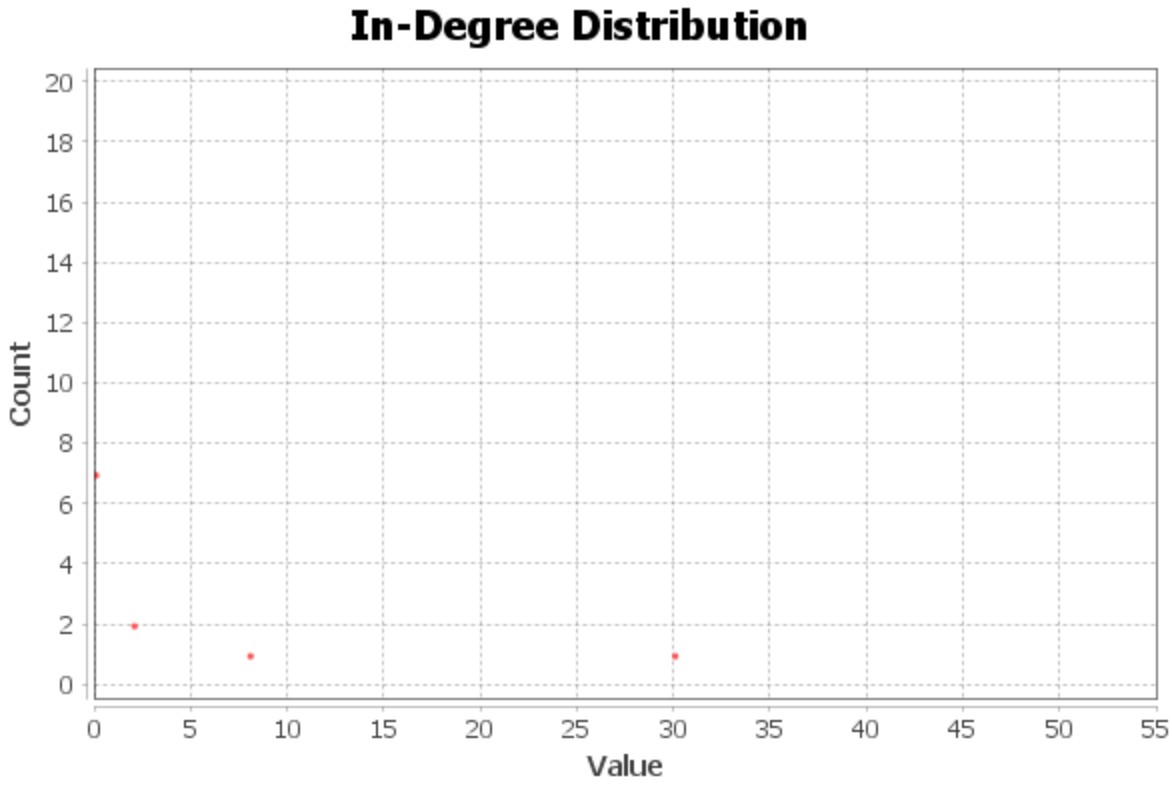
3. Average Degree Centrality

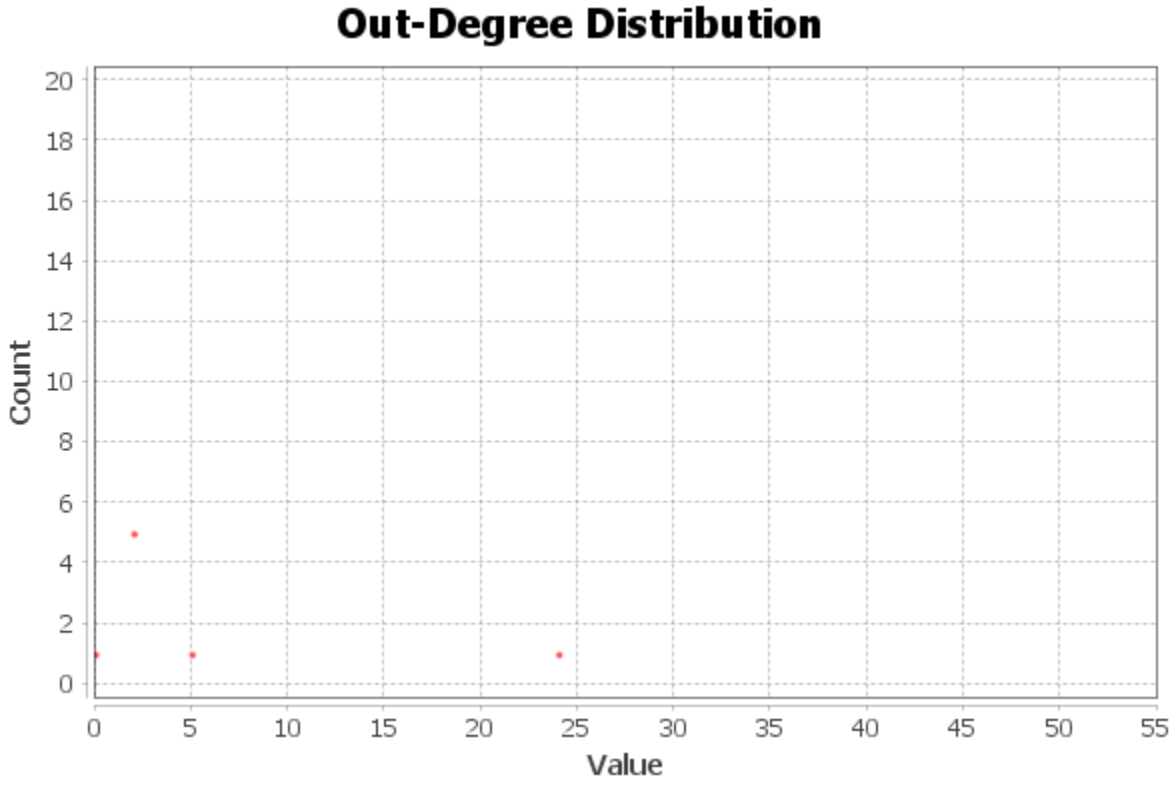
Results:

Average Degree: 1.886

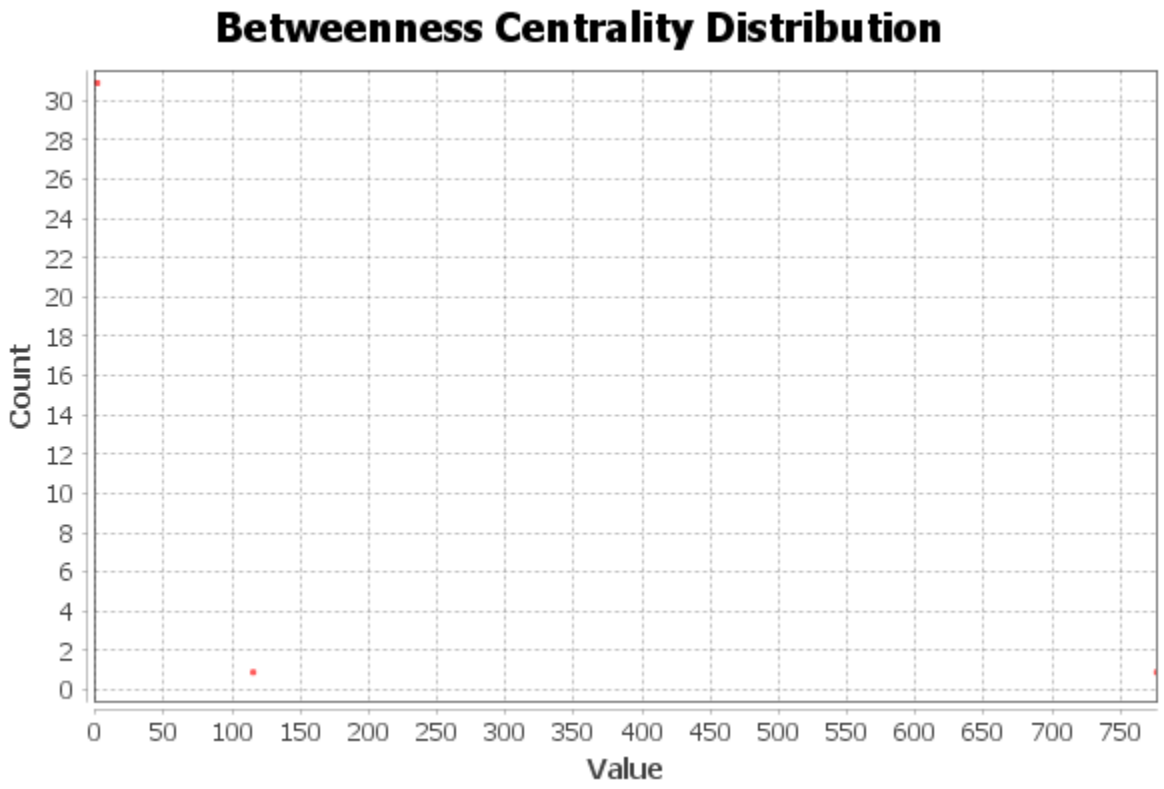
The degree centrality of an undirected network means the average number of nodes accessible by a node. In an email network like ours, which is directed, the degree centrality will be the average of the in-degree as well as the out-degrees. This means that on an average, people in this network have contacted or have been contacted by someone else 1.886 times, i.e, their average in-degree and out-degree is 1.886.




In-Degree Distribution
Count
Value
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30
35
40
45
50
55


Out-Degree Distribution
Count
Value
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30
35
40
45
50
55

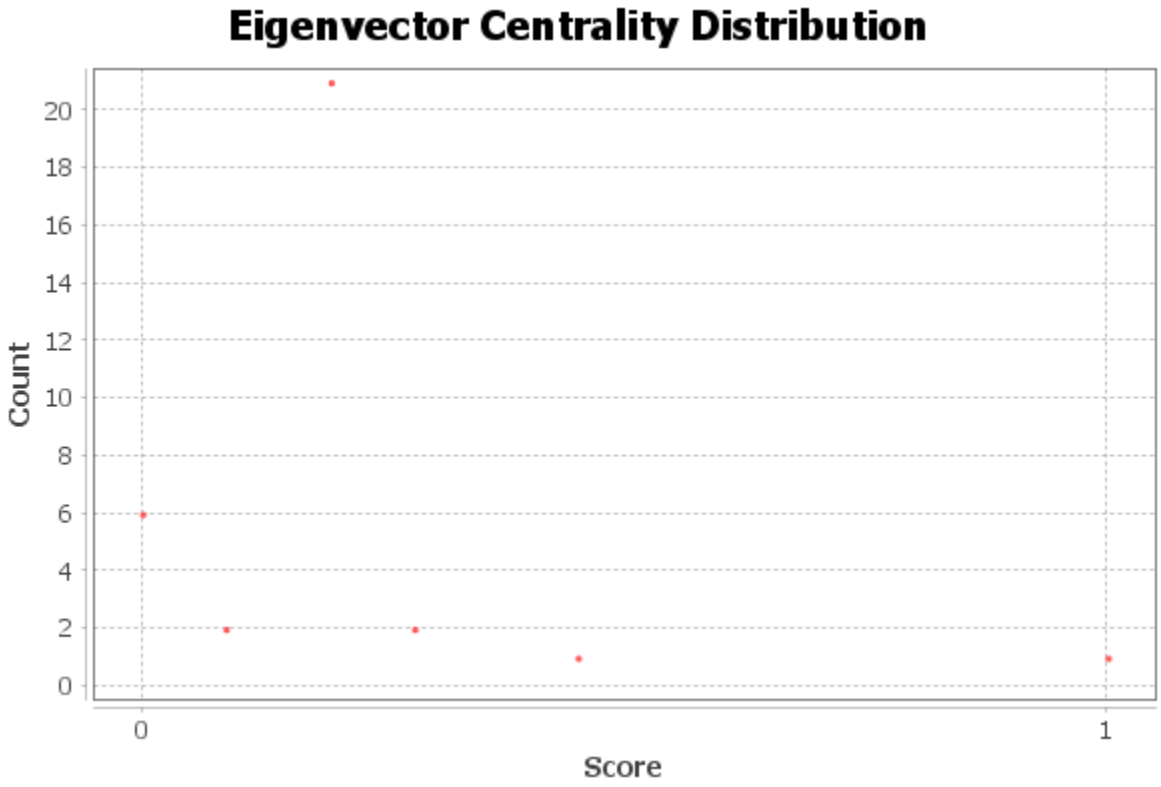
4. Betweeness Centrality:
   Results:
   Diameter: 3
   Radius: 2
   Average Path length: 2.028935185185185
   Average betweeness centrality: 25.1837
   We can see the betweeness centrality distribution as below:



5. Eigenvector Centrality:

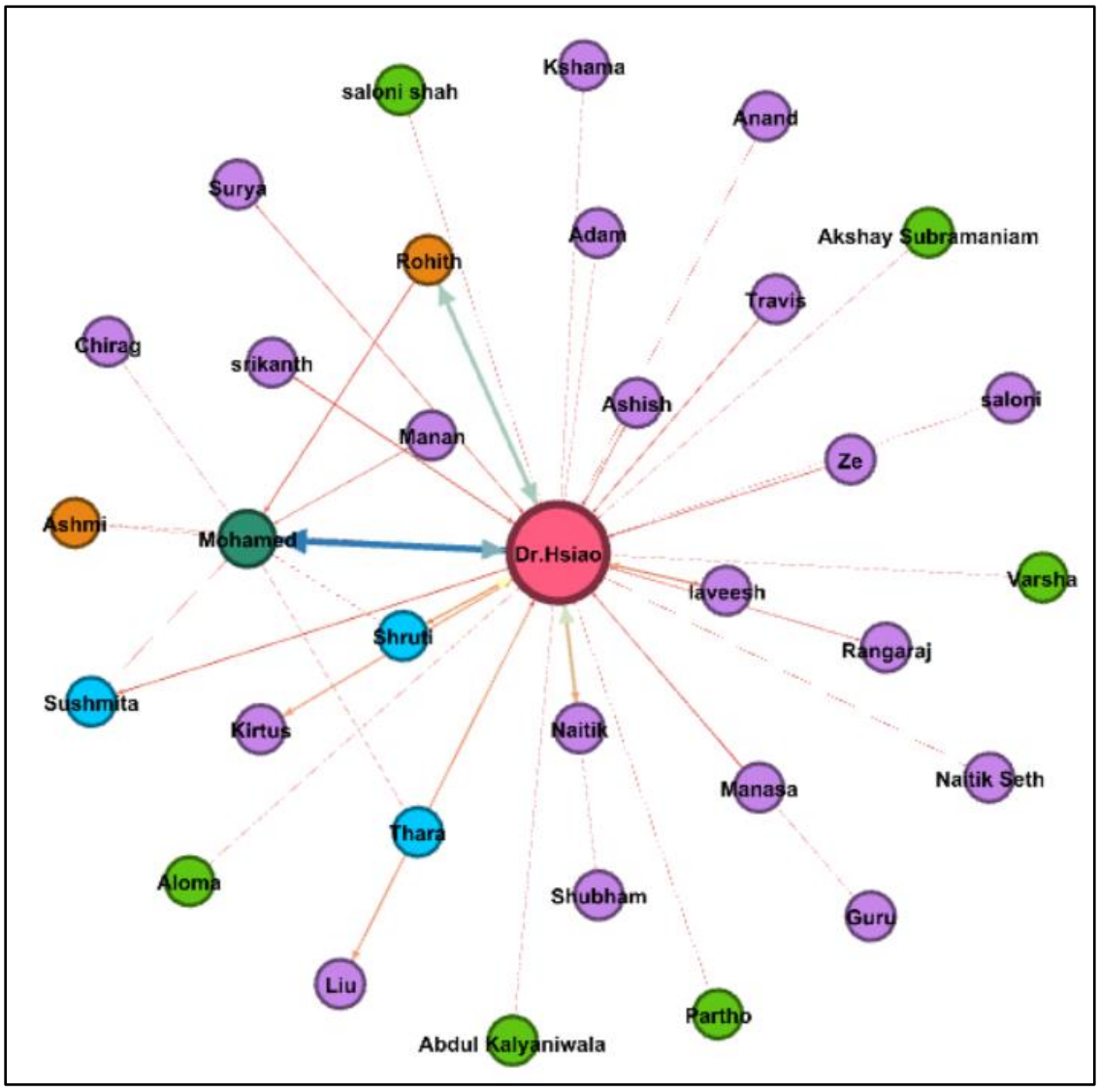
   Parameters:
   Network Interpretation: directed
   Number of iterations: 100
   Sum change: 2.0925821254005883E-4

   Results:

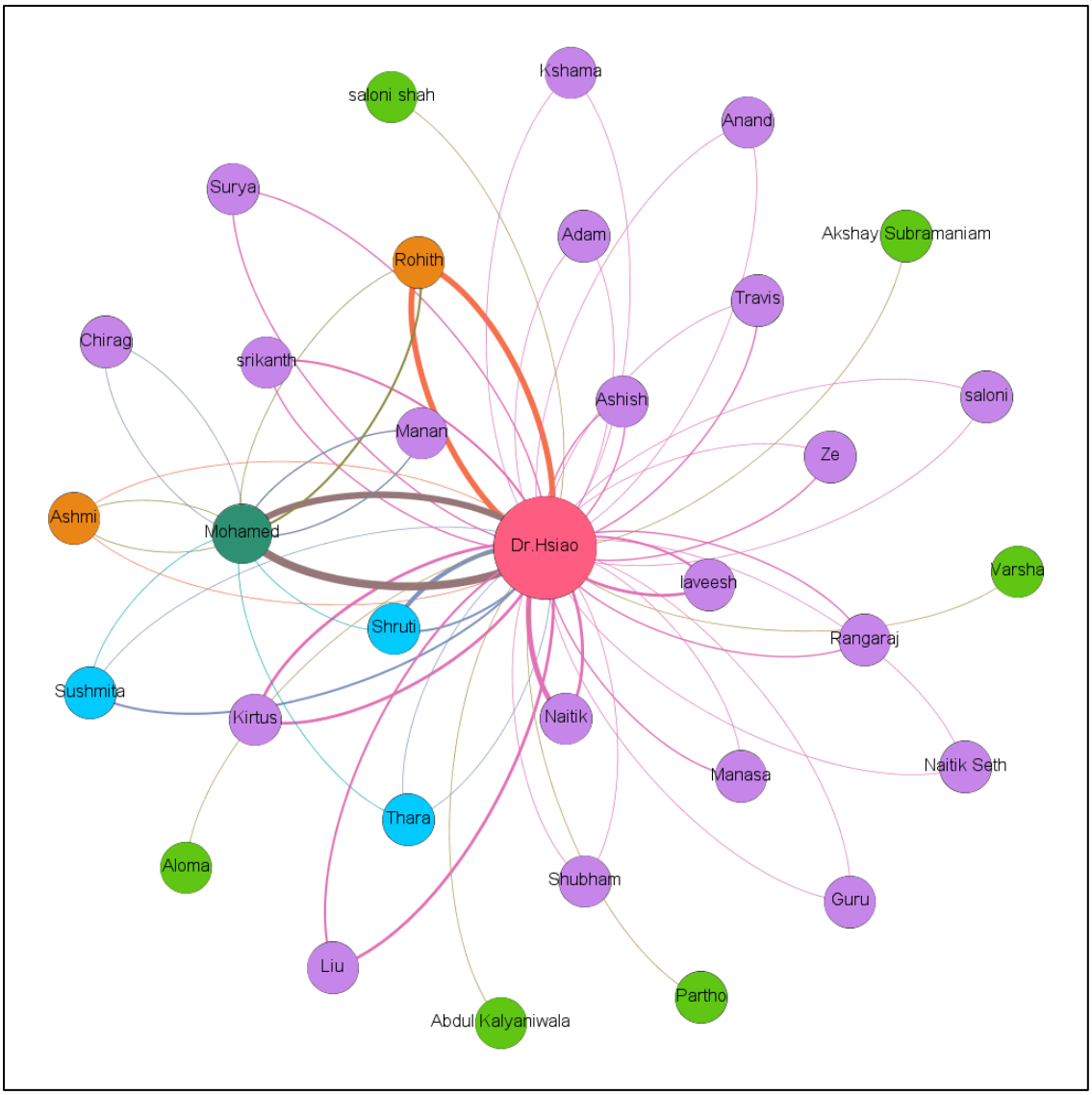


Now, if we check each nodes eigenvector centrality, we get that 'Dr. Hsiao' is the most important node with an eigenvector centrality of 1.0.
The average eigenvector centrality of the network will give us on an average how important is each node in the network. For this network, it comes out to be 0.2371.

6. Reference Layout:



The visualization generated on exporting is shown below:



The nodes have been colored and sized based on the eigenvector centrality. The minimum node size has been set to 15 and the maximum node size to 60. This allows the nodes to be sufficiently big, hence visualized easily. The label names are the people in the network who have sent emails.

I decided to use the Fruchterman-Reingold algorithm for this workspace because the Fruchterman-Reingold Algorithm is a force-directed layout algorithm. The idea of a force directed layout algorithm is to consider a force between any two nodes. Since our dataset is consisting of all the email comminications between the professor and the students almost all the nodes would be connected to Dr Hsiao so it makes sense keep Dr Hsiao's node is at the center. This design lets us know the connection between Dr and all the students. The edge weight determines the number of correspondences front and back. The minimum node size is 25% and maximum node size is 60%. The label for each node is the name of the person. The nodes are partitioned by modularity and colored accordingly.