Graph classification of communications over E-mail in an institution

Heer Gohil
SC435: Complex Networks
DA-IICT
Gandhinagar, Gujarat
201901135@daiict.ac.in

Aakash Desai
SC435: Complex Networks
DA-IICT
Gandhinagar, Gujarat
201901223@daiict.ac.in

***Abstract*—In this paper we are analyzing the Email Eu Core network. There are several outcomes from the analysis such as small average path length and small clustering co-efficient. The paper also discusses about finding the communities and centrality and other alike subjects. The supporting code for all these observations are on Github Repository.**

Keywords— Email-Eu-Core network, Complex Network, communication network.

I. Introduction

Networks related to different fields are now being researched along with their characteristics, and depending on these statistical characteristics, such networks can also be characterised. Additionally, a number of network models have been put out , which may be very beneficial in comprehending the statistical features. One of the four main kinds of complex networks is the communication network. The Internet is a real world complex network that has been investigated and is an example of a communication network.

II. About the Dataset

A significant European research institution's email data was used to create the Email-Eu network. All emails sent and received by individuals associated with the research organization were anonymised. The network's nodes are represented by the members, and the presence of an edge between two nodes implies that at least one email has been sent and received between the two members. The collection does not include any incoming or outgoing emails to or from the rest of the world; instead, the emails exclusively represent communication among institution members.

Total Nodes - 1005
Total Edges - 25571

III. Method and Observation

The methods used for the project are as mentioned below:

1) *Reading the data*
2) *Data Cleaning*
   - Checking nodes which are not connected to any other nodes.
   - Removing isolated node
3) *Data Analysis*
   - Generation of Adjacency matrix
   - Network Density
   - Clustering Co-efficient
   - Degree Analysis
   - Average Shortest path length
   - Diameter
   - Minimum Spanning Tree
   - Community Analysis
   - Centrality Analysis
   - Strong and Weak Components
   - Power Law

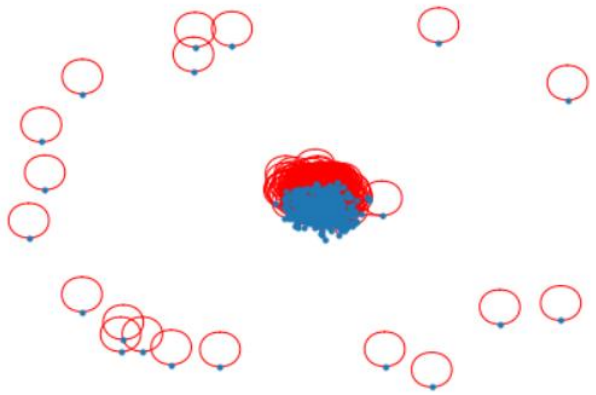
III.(1) Reading the Data.

Using pandas library, we have opened and read the .txt data file.

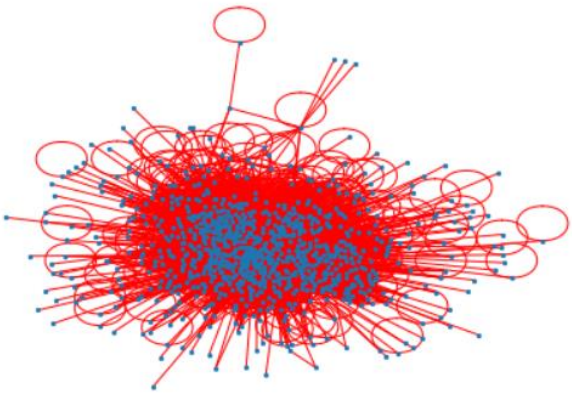
	0	1
0	0	1
1	2	3
2	2	4
3	5	6
4	5	7

III.(2) Data Cleaning

Here we have few nodes who are not contacting anyone and are not contacted by anyone. These nodes are either non active email ids or are of the people who are no more in the organization but have their email ids still active.



Here the nodes in the outer region are of those types. They are removed and the rest of the data is taken into consideration for further analysis.



III.(3) Data Analysis

A) Generation of Adjacency Matrix

Here using the given edge list, we create a directional graph G. We then create adjacency matrix using that graph.

```
array([[1, 1, 0, ..., 0, 0, 0],
       [0, 1, 0, ..., 0, 0, 0],
       [0, 0, 1, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

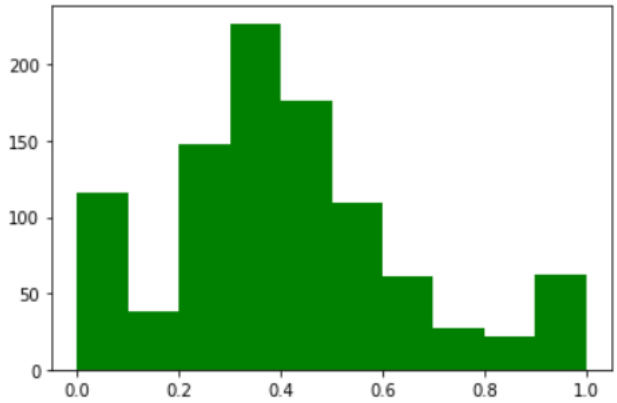
B) Network Density

The percentage of actual connections among a network's possible connections is known as "network density." A link between two "nodes" is referred to as a "possible connection" regardless of whether it really exists. [1]

Density of the Network: 0.03311

C) Clustering Co-efficient

The clustering coefficients (a measure of the density of triangles in a network) quantify the average likelihood that two neighbours of a vertex are also neighbours. [2]

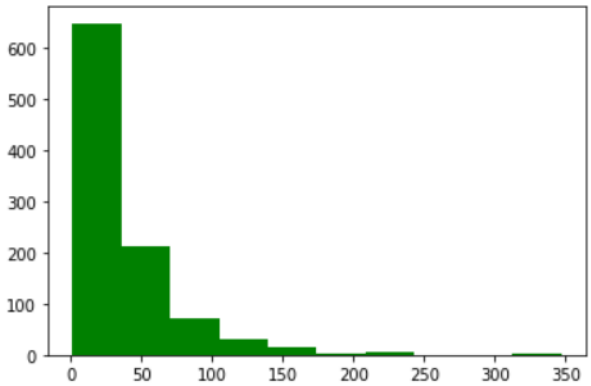


Average Clustering: `0.4070504475195386`

Clustering coefficient tells how well connected the neighborhood of the node is. If the neighborhood is fully connected, the clustering coefficient is 1 and a value close to 0 means that there are hardly any connections in the neighborhood.Here we have a clustering coefficient avg of 0.407 and hence it lies in middle.

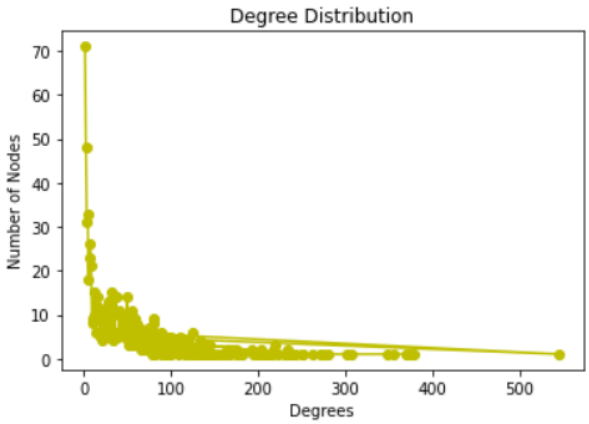
D) Degree Analysis

Here we find out the average degree of our email network.



Average Degree of the network: `33.84787018255578`

Here the average degree of the email network is 33.84 and the graph depicts that more than 600 nodes have a degree of between 0 to 50 and 200 nodes have a degree between 50 to 100.



E) Average shortest path length

The average number of steps along the smallest pathways for all potential pairs of network nodes is the definition of average path length, also known as average shortest path length, in terms of network topology. It is a way to gauge how well people can move large amounts of data through a network. [3]

Avg Shortest Path Length: 2.586933824816466

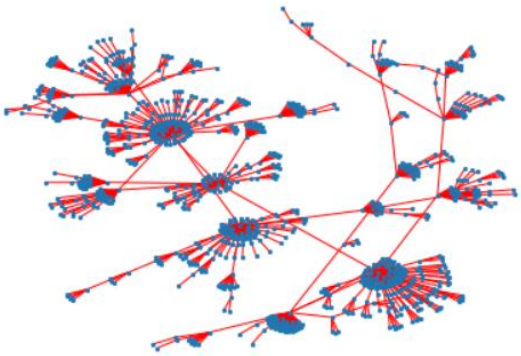
F) Diameter of a Network

It is the shortest distance in the network between the two furthest nodes.

Diameter: 7

G) Minimum Spanning Tree

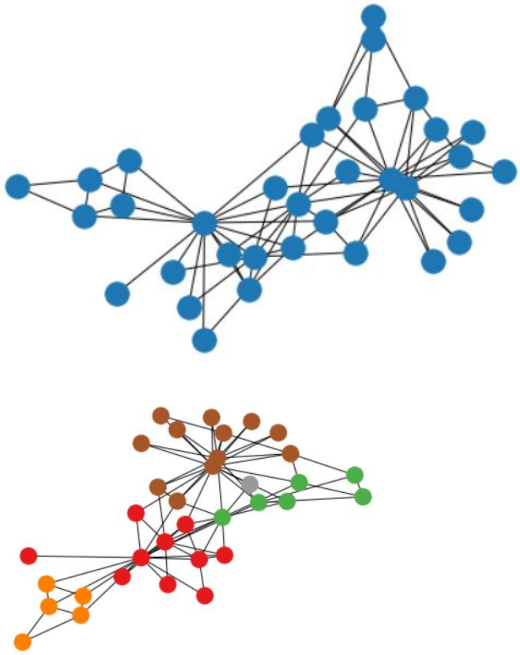
Finding the minimum Spanning Tree will help us to find out which roles the company has added for them to act as middlemen. Minimum Spanning Tree removes the extra edges and only keeps those ones which will help in achieving shortest path and also help in making graph smaller.



Here the graph depicts that there are some nodes central to many other nodes and the graph is one of the most efficient graphs to refer for node to node communication as it depicts the shortest path as well as central nodes in the graph.

H) Community Analysis

We use the karate club graph function to find the different types of communities in our network.

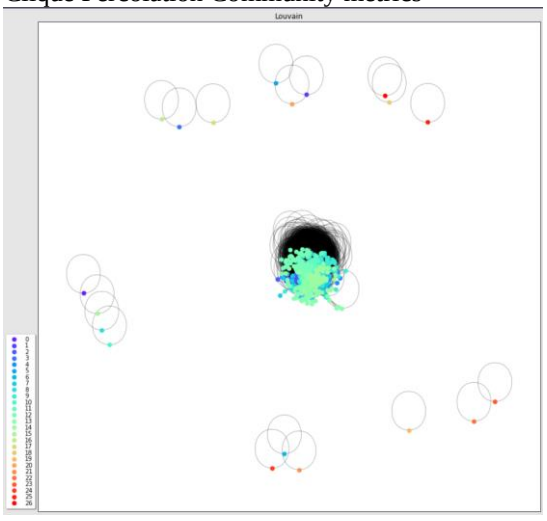


Here the graph depicts that there are some nodes central to many other nodes and the graph is one of the most efficient graphs to refer for node to node communication as it depicts the shortest path as well as central nodes in the graph.

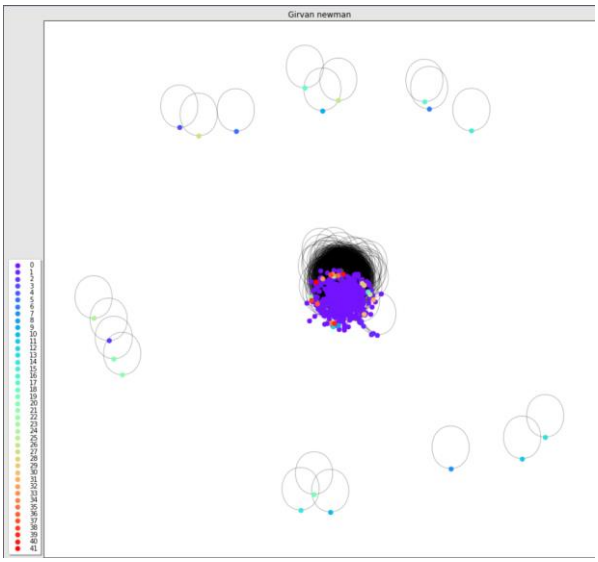
Now using Girvan Newman Algorithm, we will find out which members(nodes) are present in which community.

```
[({0, 1, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21},
  {2, 8, 9, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33}),
 ({0, 1, 3, 4, 5, 6, 7, 10, 11, 12, 13, 16, 17, 19, 21},
  {2, 8, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33},
  {9}),
 ({0, 1, 3, 7, 11, 12, 13, 17, 19, 21},
  {2, 8, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33},
  {4, 5, 6, 10, 16},
  {9}),
 ({0, 1, 3, 7, 11, 12, 13, 17, 19, 21},
  {2, 24, 25, 27, 28, 31},
  {4, 5, 6, 10, 16},
  {8, 14, 15, 18, 20, 22, 23, 26, 29, 30, 32, 33},
  {9}),
 ({0, 1, 3, 7, 12, 13, 17, 19, 21},
  {2, 24, 25, 27, 28, 31},
  {4, 5, 6, 10, 16},
  {8, 14, 15, 18, 20, 22, 23, 26, 29, 30, 32, 33},
  {9},
  {11}),
 ({0, 1, 3, 7, 12, 13, 17, 19, 21},
  {2, 24, 25, 27, 28, 31},
  {4, 5, 6, 10, 16},
  {8, 14, 15, 18, 20, 22, 23, 29, 30, 32, 33},
  {9},
```

Clique Percolation Community metrics



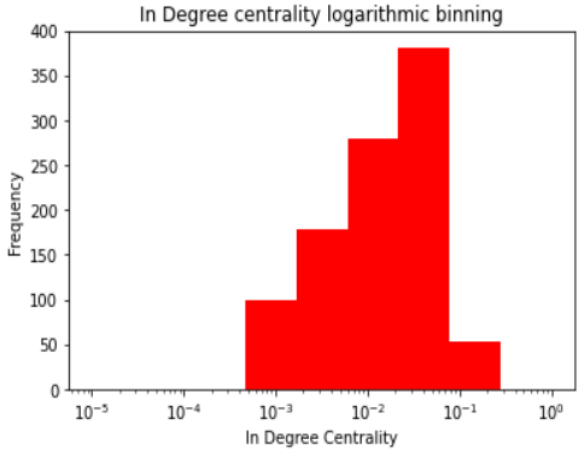
Louvain metrics:
Recall = 0.7790519877675841
Precision = 0.25784050494116983
Purity = 0.4626865671641791
Modularity = 0.4341910304658381



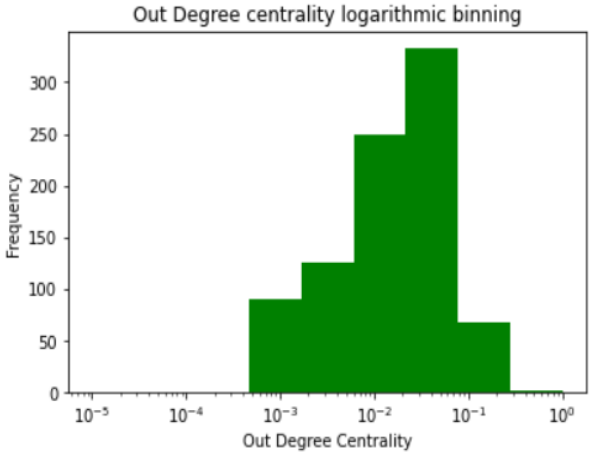
Girvan newman metrics:
Recall = 0.8983180428134556
Precision = 0.0460407161018425
Purity = 0.15024875621890546
Modularity = 0.003583879527853089

I) Centrality Analysis

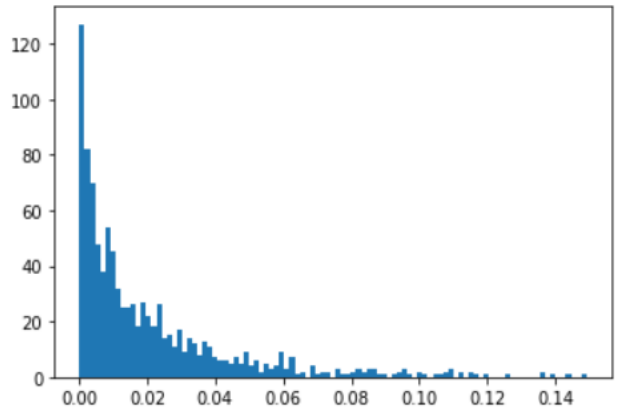
➔ In degree and Out Degree Analysis
The percentage of other nodes that a node v's incoming edges are connected to determines its in-degree centrality.



The percentage of nodes that a node v's outgoing edges are linked to is known as the node's out-degree centrality.



➔ Eigen Vector Analysis
Based on the centrality of its neighbours, eigenvector centrality calculates a node's centrality. For directed graphs, we may assess a node's "importance" and "prestige" using the Eigen Vector centrality (based on the out-degree Eigen Vector) (through the in-degree Eigen Vector) [4]



The graph depicts that the centrality of major nodes lies between 0.0 and 0.02.
➔ Katz Centrality
By counting the number of immediate neighbours (first degree nodes) as well as all other nodes in the network that are connected to the node under consideration through these immediate neighbours, Katz centrality calculates the relative importance of a node within a network. [5]

```
Katz centrality
(84, 0.17271020446671365)
(433, 0.1604154493979983)
(71, 0.14641608049868257)
(50, 0.12314093428610086)
(288, 0.11261831611288756)
(431, 0.10931162148284765)
(49, 0.10756510659487527)
(494, 0.10501241955701036)
(283, 0.10285840675021578)
(217, 0.1026592199303793)
```

➔ Closeness Centrality
In a linked graph, a node's closeness centrality (or closeness) is a measure of its network centrality and is computed as the inverse of the lengths of all its shortest

routes to all other nodes. As a result, a node is closer to all other nodes the more central it is.

```
closeness centrality
(160, 0.4496688397114823)
(62, 0.4367960817756949)
(107, 0.43313263076725356)
(434, 0.42843368072064014)
(121, 0.42761501763646054)
(86, 0.423034390478414)
(64, 0.4214410406084389)
(129, 0.4190733943128858)
(183, 0.4185508589085057)
(128, 0.41725020366018833)
```

➔ Betweenness Centrality

Betweenness centrality, which is based on shortest routes, is a measure of centrality in a graph in graph theory. In a linked graph, there must be at least one shortest path between every pair of vertices such that the path's number of edges is kept to a minimum. The number of these shortest routes that travel through a vertex determines its betweenness centrality. [6]

```
Betweenness centrality
(160, 0.07212078608028884)
(86, 0.037432912122184775)
(5, 0.026984804243671952)
(121, 0.024532102889508717)
(62, 0.02451110558180135)
(107, 0.02185778938072534)
(64, 0.01866532308007618)
(82, 0.018229609890582157)
(377, 0.016220465196907407)
(129, 0.015569466534401019)
```

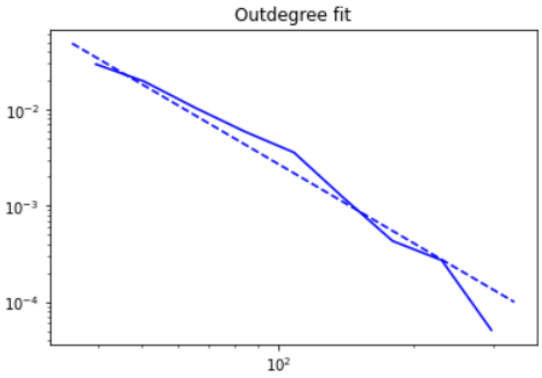
J) Strongly Connected Components
Total Number of Strongly Connected Components: 203

K) Weakly Connected Components
Total Number of Weakly Connected Components: 20

L) Power Law

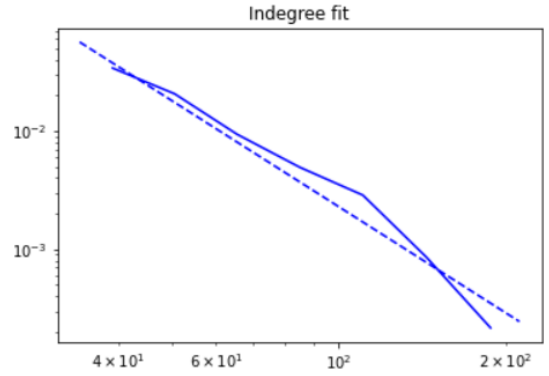
➔ Out Degree Fit

Best Minimal Value for power law fit
2.7453320333942326
0.1084496557184573



The out-degree distribution is one of the most reported topological properties to characterize real complex networks. This property describes the probability that a node in the network has a particular number of outgoing links.

➔ In Degree Fit



References

Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. "Local Higher-order Graph Clustering." In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017

J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007

Reiser, M., 1979. A queueing network analysis of computer communication networks with window flow control. *IEEE transactions on Communications*, *27*(8), pp.1199-1209.

Tichy, N.M., Tushman, M.L. and Fombrun, C., 1979. Social network analysis for organizations. *Academy of management review*, *4*(4), pp.507-519.