

**EMAIL NETWORK ANALYSIS USING PYTHON AND NETWORKX**

***Table of content:***

***Abstract***

***About the dataset***

***Degree centrality***

***Betweenness centrality.***

***Page rank/ eigen vector centrality.***

***Clustering coefficient***

***Closeness centrality.***

ABSTRACT :

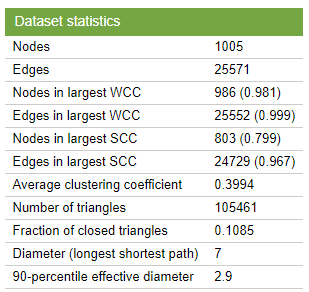
This is a dataset of various email communications between people, graph analysis on this dataset helps us to know about the important people, leaders and opinion makers in this network which will be helpful for fast communication among any community

ABOUT THE DATASET:

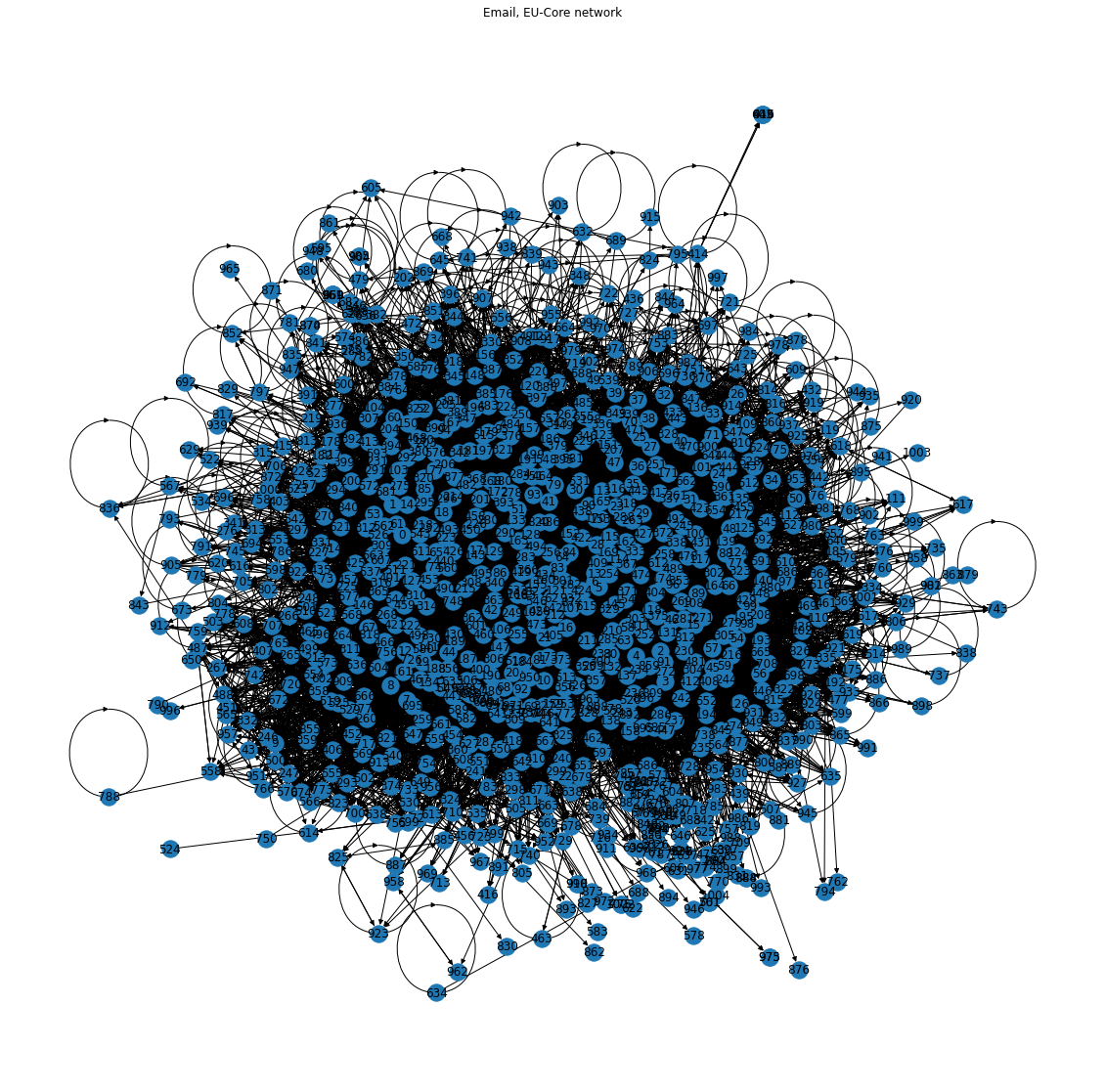
The network was said that it is generated using email data from a large European research institution. We have anonymized information about all incoming and outgoing email between members of the research institution.

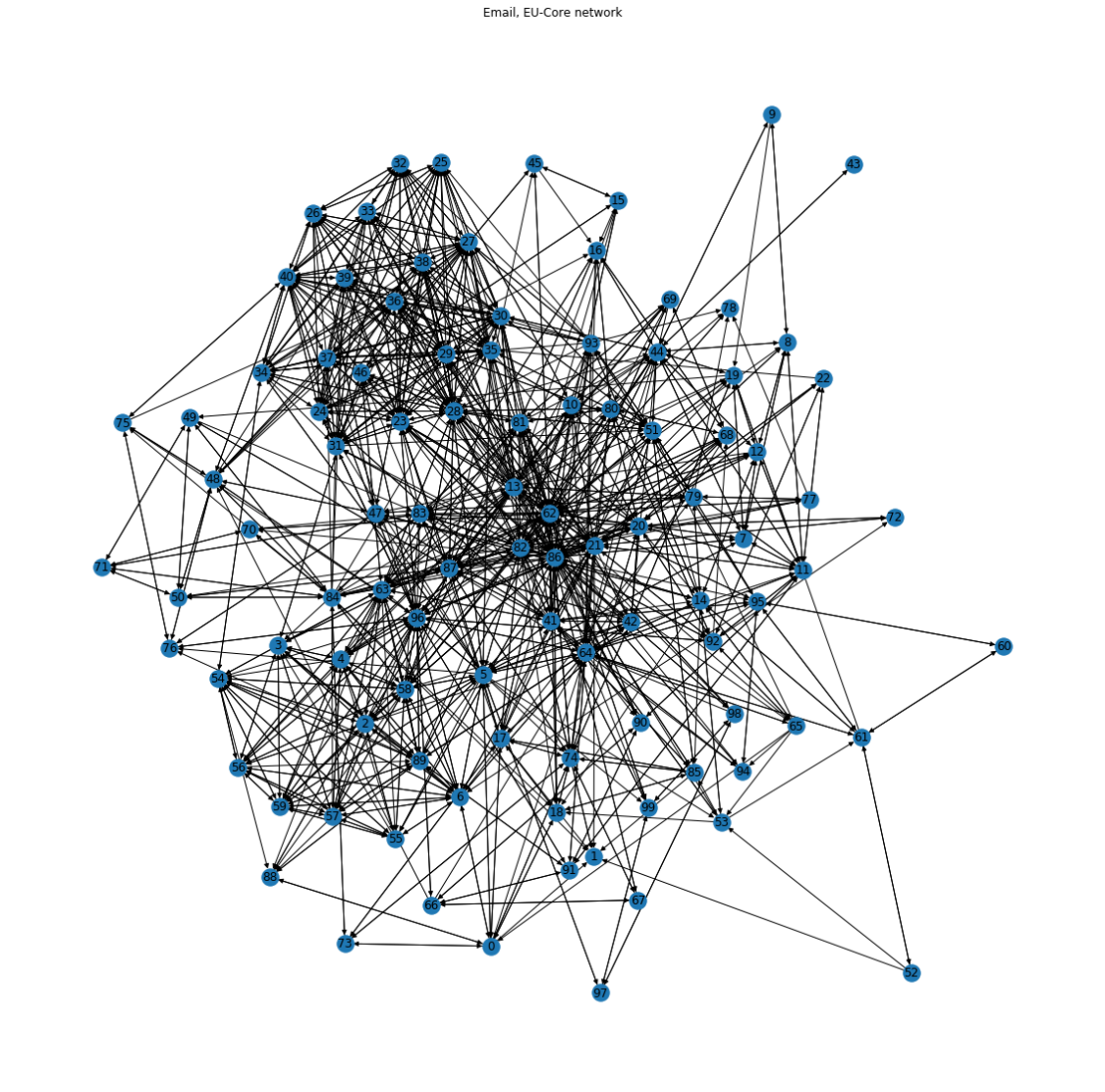
There is an edge (u, v) in the network if person u sent person v at least one email. The e-mails only represent communication between institution members (the core)

Each individual belongs to exactly one of 42 departments at the research institute.



1. Email communication links between members of the institution
2. Department membership labels





Nodes -100 ; edges - 1315

We can see that there are two densely connected networks and many connections between them.

**Nodes:** The individuals whose network we are building.

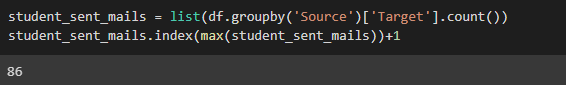
**Edges**: The connection between the nodes. It represents a relationship between the nodes of the network.

**DEGREE CENTRALITY :**

Degree centrality - is a measure of the number of connections a particular node has in the network.

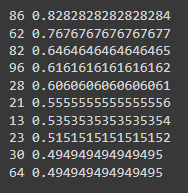
This way of looking at the importance of a person based on number of connections (Degrees) is called degree centrality.

The person with the maximum number of interactions is:

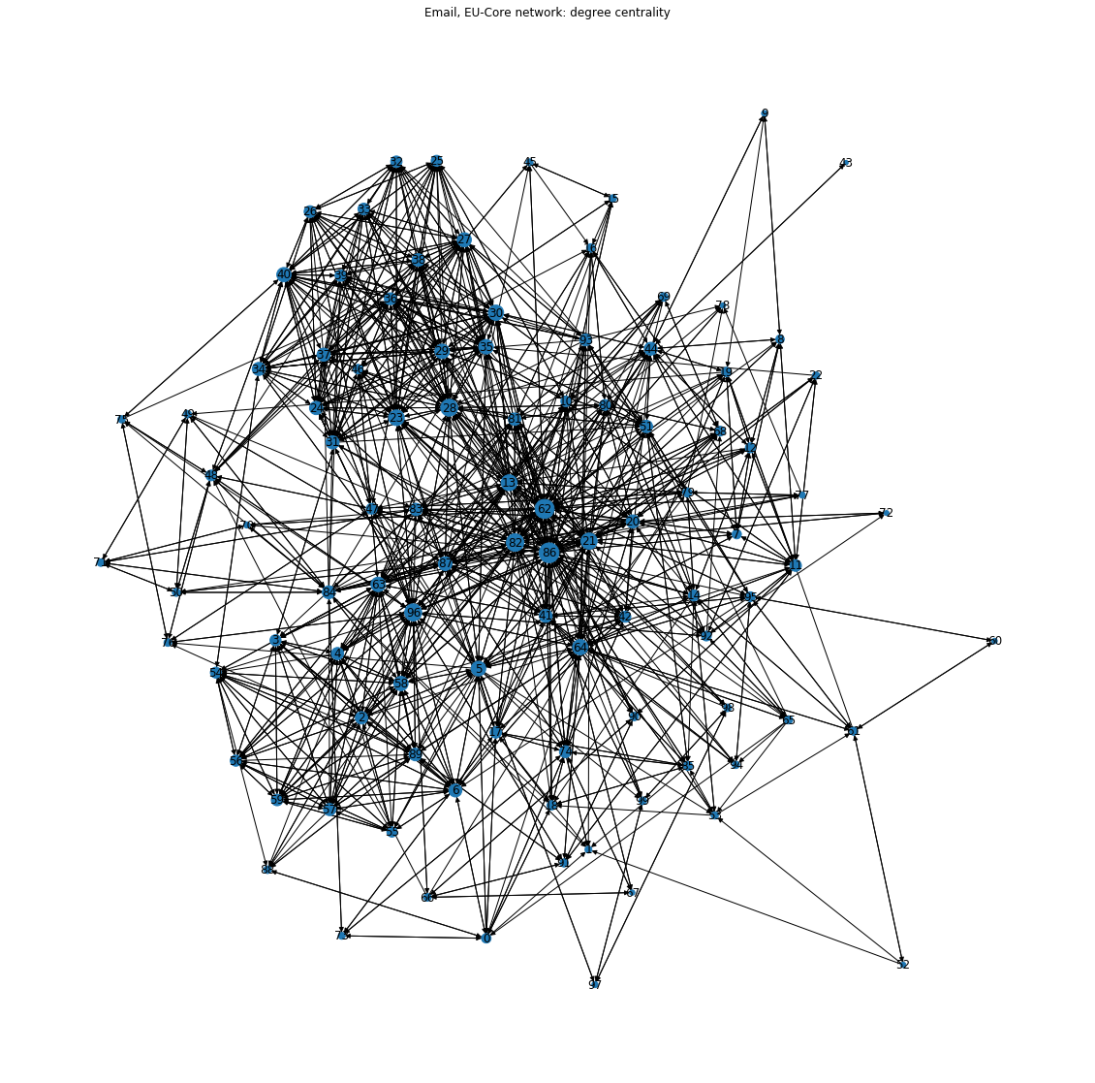


We can say that the person with index 86 has the most number of interactions and is one of the most important persons. He would be the go-to person who has connections with maximum number of people. We can see this in the graph also.

In ascending order, the top 10 people with maximum number of connections are:



This network can be visualised with most size proportional to the importance according to degree centrality.

.

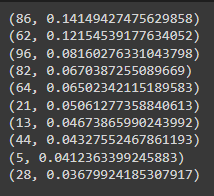
**BETWEENESS CENTRALITY :**

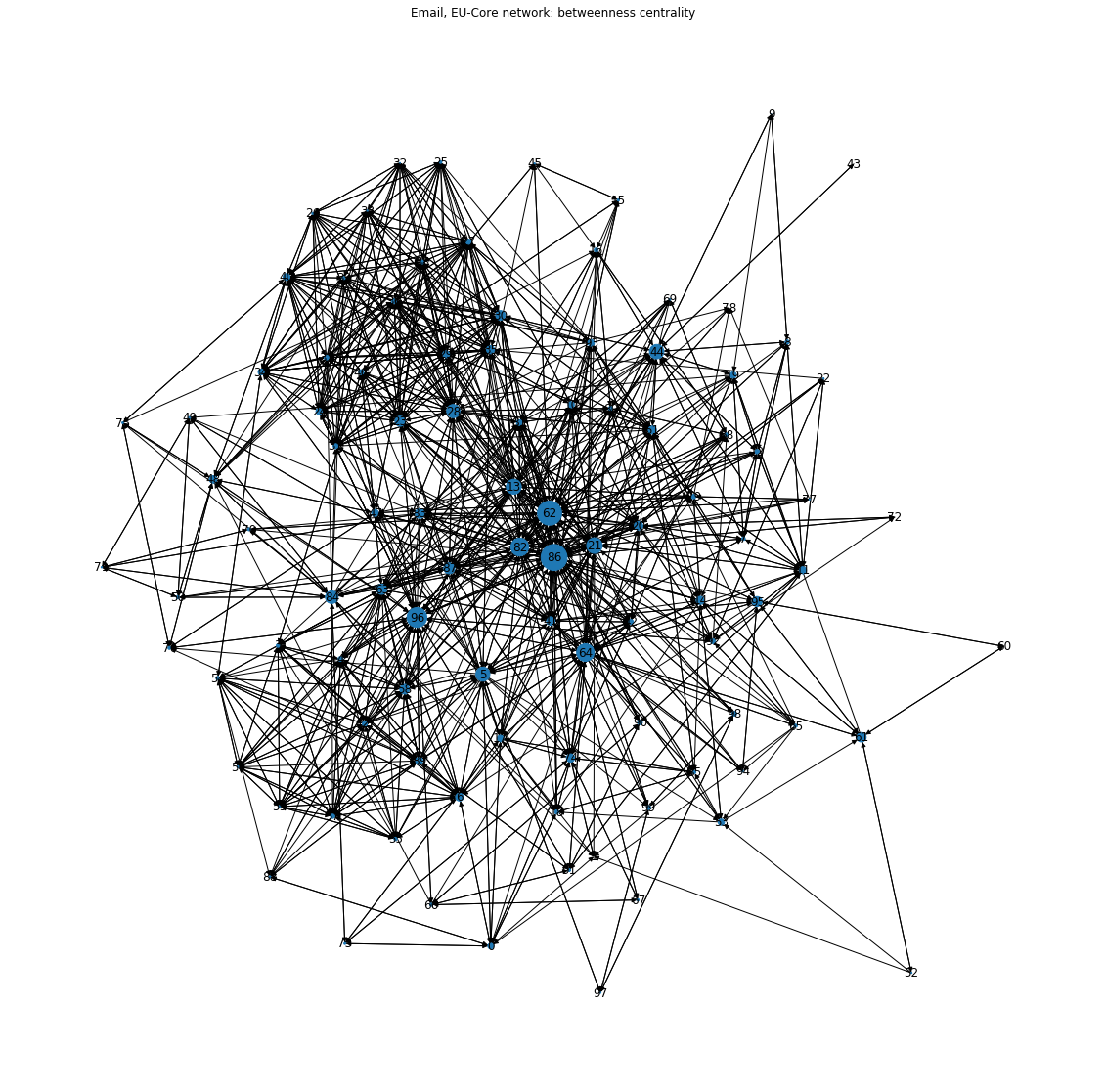
The Betweenness Centrality is the centrality of control.

It represents the frequency at which a point occurs on the shortest paths that connected pair of points. It quantifies how many times a particular node comes in the shortest chosen path between two other nodes.

The nodes with high betweenness centrality play a significant role in the communication/information flow within the network.

Although person 86 has sent mails to the maximum number of people, he has still sent mails to only 45 out of the 100 people. This means if we want to send a mail to the whole class, that there might be other important people through whom the remaining people receive mails. We should identify people who are part of different closely knit groups and who are essential for passing information between groups. This is given by betweenness centrality. The most important people according to this metric in descending order are:



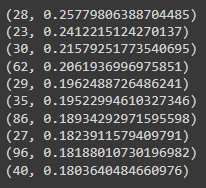


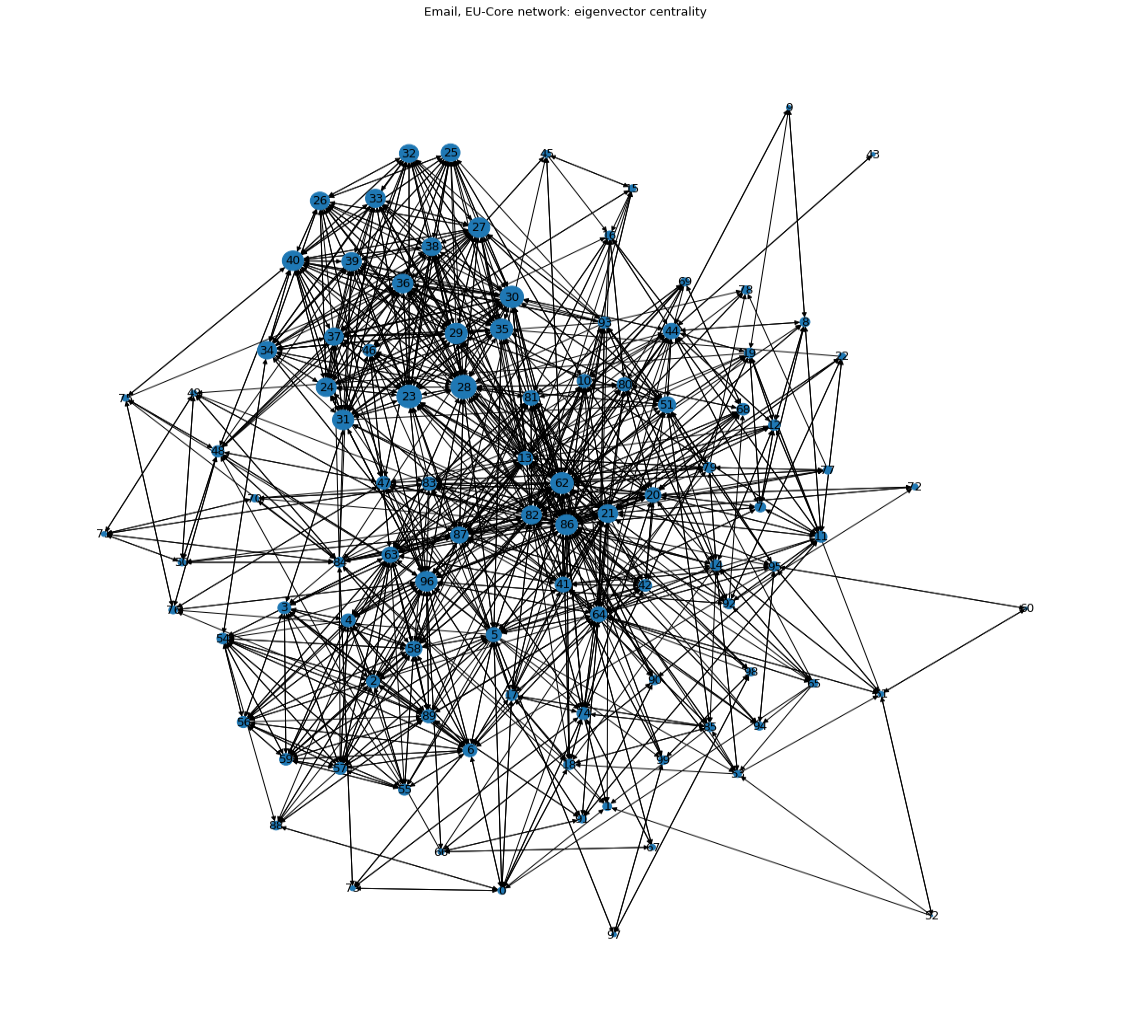
We get the same people at the top, nut this time including a few additional people.

**PAGE RANK (EIGEN VECTOR CENTRALITY):**

Eigenvector centrality is a measure of how import a node is by accounting for the fact of how well it is connected to other important nodes.

Another way to identify the important people in a network is to not only look at the number of people who are connected to the person but also the importance of the people whom they are connected to. The importance of the person can be defined as proportional to the sum of importance of the immediate people connected to him/her. This is defined by eigenvector centrality. Below is the top ten people who are important based on who they know and whom they are connected to:

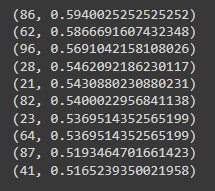




**CLOSENESS CENTRALITY:**

Closeness Centrality is a measure where each node’s importance is determined by closeness to all other nodes.

Closeness centrality is a metric that assesses how quickly data flows from one node to another. It tells you how close a node is to the rest of the network's nodes. Closeness centrality can help us identify the people who are most likely to have an impact the entire network in the shortest amount of time .



**CLUSTERING COEFFICIENT:**

Any graph in general can be densely connected or sparsely connected. The behaviour of the graph, and therefore the actions that have to be taken, are different for sparse graphs vs densely connected graphs. The betweenness measures above should be looked at along with clustering coefficient. The clustering coefficient is a way of measuring the degree to which the nodes in a graph cluster together. Networks with high number of clustering coefficient are more social. The average clustering coefficient for the graph is (assuming an undirected graph):



From the above analysis, we can find the most important people in the network. The people who have the most connections, the people who are part of multiple groups and are important to send message across and the people who are important based on whom they know immediately.