SHREYA SRIRAM
195001106
CSE B

MAPPING LETTERS OVER EUROPE

Overview

This project analyzes the postal network in Europe for a selected time period. The network is made of 2 components - nodes (actors) and edges (relationships). The network is visualized in different layouts, analyzed the position of actors in the network using centrality measures and identified the communities using community detection algorithms.

Dataset

The dataset was downloaded from here as 2 CSV files, one for nodes and the other for edges.

Nodes.csv => Id; Label; Attribute1; City; Latitude; Longitude
Edges.csv => Source; Target; Type; Weight

Nodes.csv

Sl No	Attribute	Values
1	ID	Unique for each person
2	Label	Name of each person
3	Attribute1	Gender (0/1)
4	City	City name
5	Latitude	Corresponding to each city
6	Longitude	

Edges.csv

Sl No	Attribute	Values
1	Source	Sender ID
2	Target	Receiver ID
3	Type	Directed
4	Weight	Number of relationships

Dataset statistics

The Nodes.csv dataset contains details of 1000 people and 14116 sender receiver relationships in Edges.csv

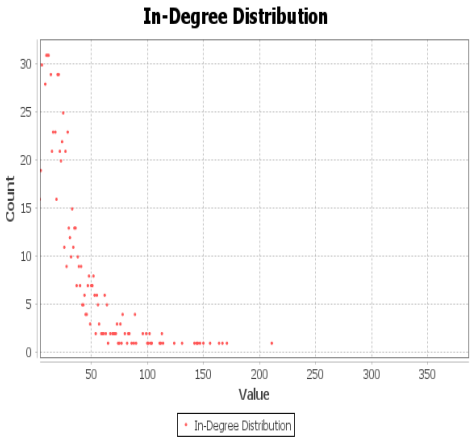
SHREYA SRIRAM
195001106
CSE B

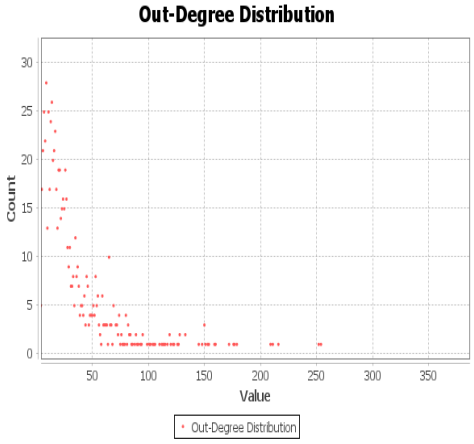
The following metrics are used to provide a holistic view of the network, in different layouts.

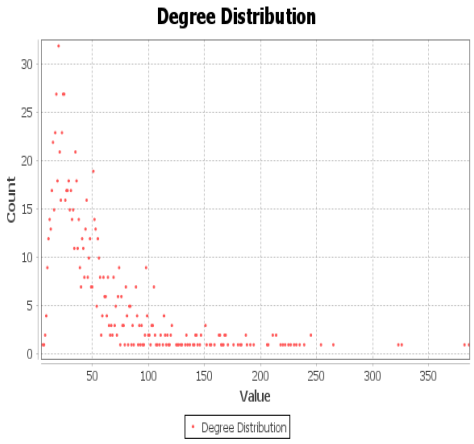
Network statistics:

1. **Average weighted degree**
   This measures the average sum of weights of the edges of nodes, where the weight of edges represents how many times the edge is traversed between the nodes.

Average Weighted Degree: 25.486







2. **Graph density**
   This metric measures how close the network is to completion, in a complete graph all possible edges are present and the value is 1.
   Graph density = 0.014

3. **Average graph distance**
   This is the distance between all pairs of nodes, connected nodes have a graph distance of 1.
   Average graph distance = 3.542446586244327

4. **Diameter of the graph**
   This measures the longest distance between any two nodes in the network.
   Diameter of the graph = 9

5. **Connected components**
   This measure defines the number of strongly and weakly connected components of the network.
   Number of Weakly Connected Components: 1

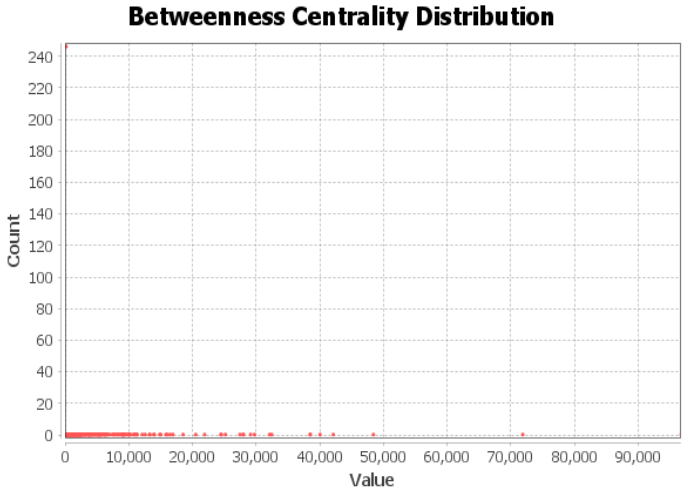
SHREYA SRIRAM
195001106
CSE B

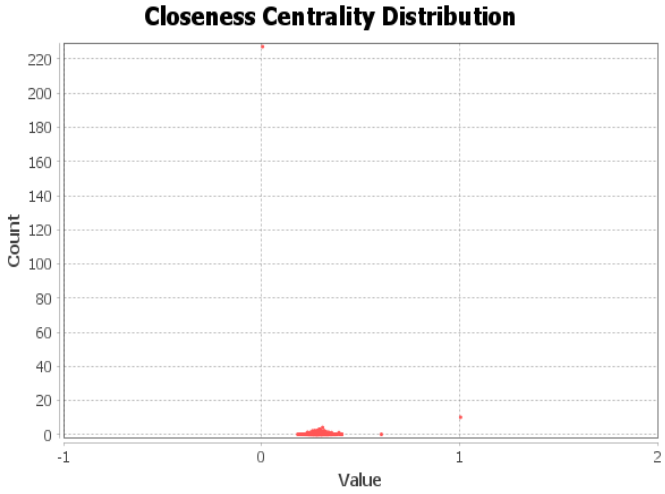
Number of Strongly Connected Components: 256

Network centrality metrics:

1. **Betweenness centrality**

   This measures how often a node appears on the shortest paths between nodes in the network.



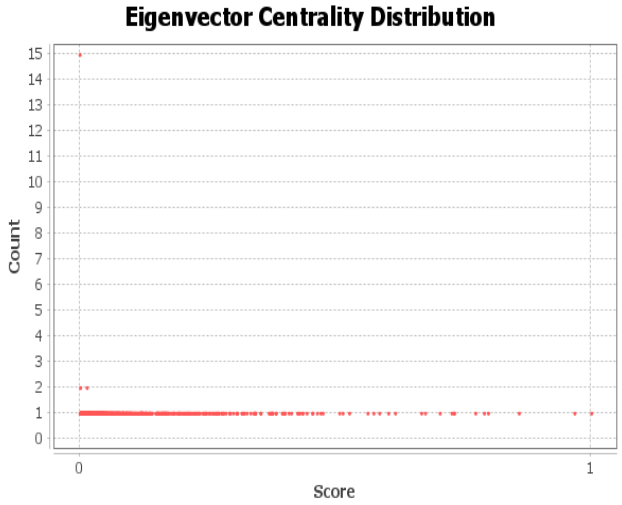


2. **Closeness centrality**

   This measures the average distance from the starting node to all the other nodes in the network.

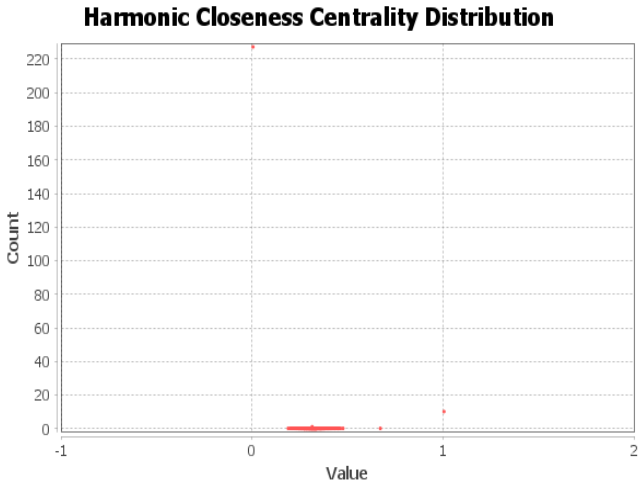
3. **Eccentricity**

   This measures the distance from a given starting node to the farthest node from it in a network.



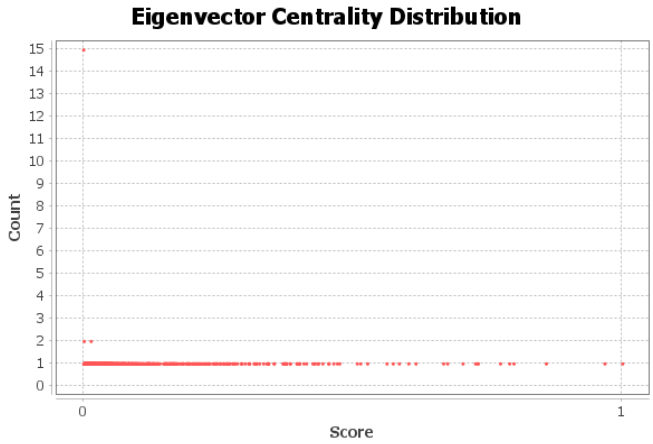
SHREYA SRIRAM
195001106
CSE B

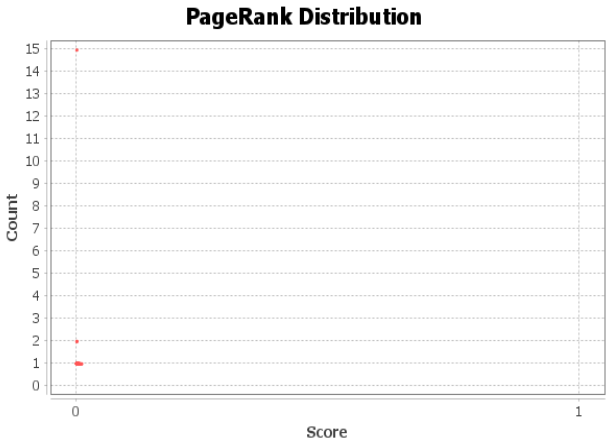
4. **Harmonic closeness centrality**
   This measures the closeness centrality for unconnected graphs by calculating the inverse distance from node to every node excluding itself.



5. **Average clustering coefficient**
   This indicates the number of nodes embedded in the neighborhood, thus giving an idea about the clustering in the network.
   Average clustering coefficient = 0.228

6. **Eigenvector centrality**
   This measures the importance of a node in a network based on its connections.



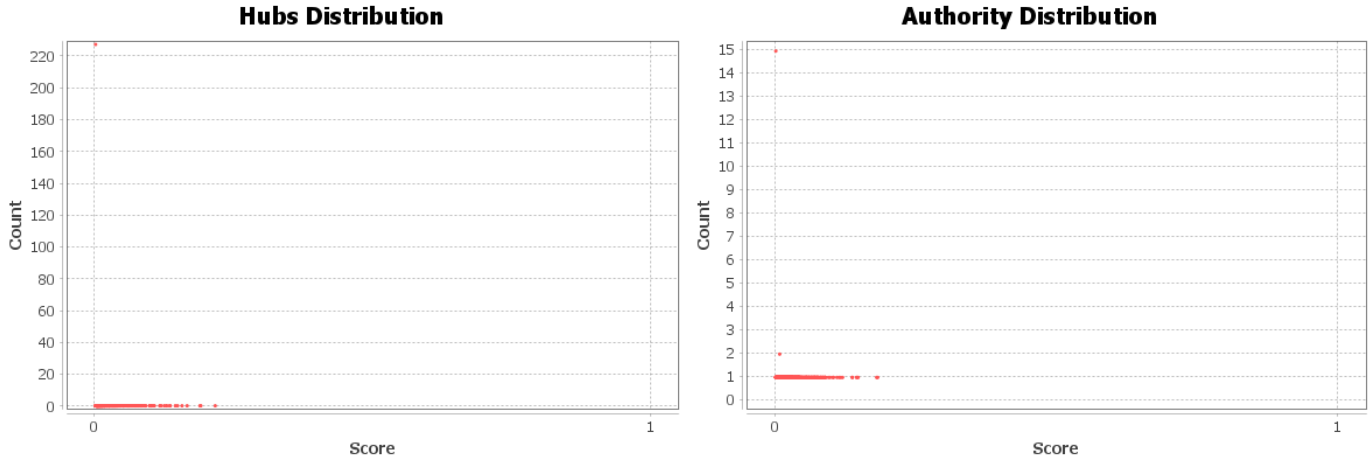


Link analysis ranking results:

1. **PageRank**
   This measure ranks the nodes according to how often a user following the links will non-randomly reach the same node.

2. **HITS**
   This measure computes two different values for each node. The authority measures how valuable the information stored in the node is, the Hub value measures the quality of the node links.

SHREYA SRIRAM
195001106
CSE B

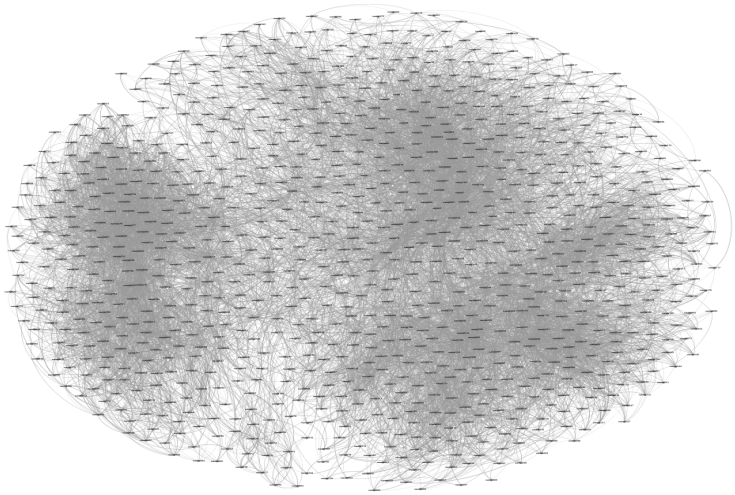


Layouts for visualization:

Gephi supports different kinds of layouts to visualize the dataset. Three such layouts have been used to analyze the dataset in this project.

1. **Fruchterman-Reingold Layout**

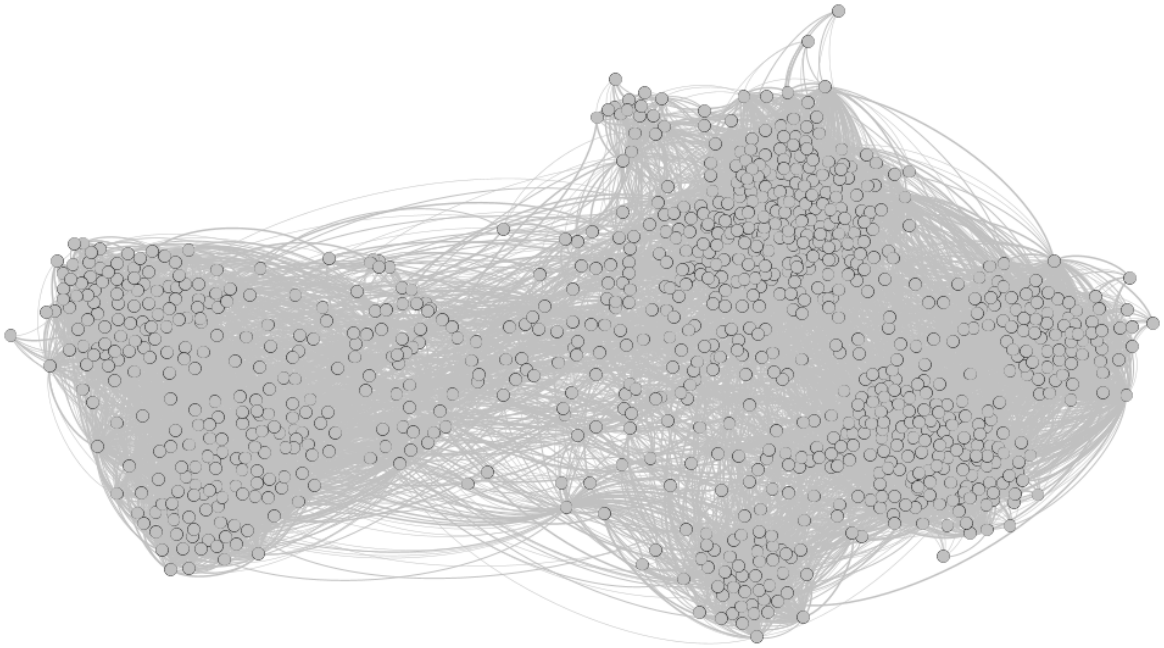
The Fruchterman-Reingold Layout is an example of a force-directed algorithm, where physical springs act as edges that attract connected vertices toward each other and a competing repulsive force that pushes all vertices away from one another, whether they are connected or not, it typically results in edges that are relatively similar in length. The algorithm uses an iterative process to adjust the relative position of each vertex in order to minimize the "energy" of the system.



SHREYA SRIRAM
195001106
CSE B

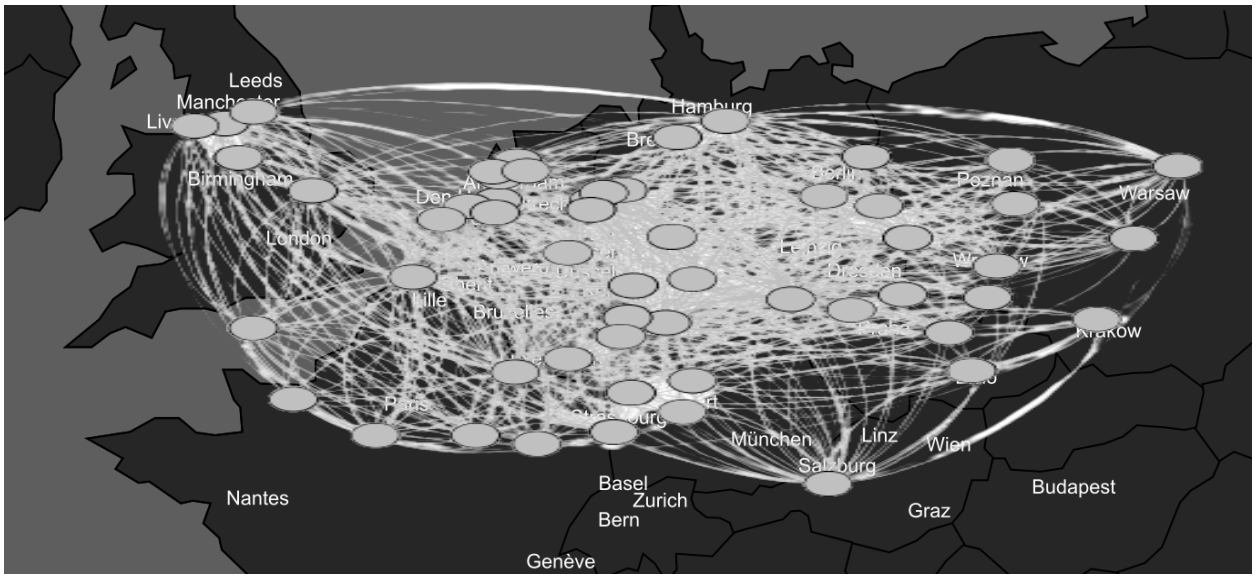
2. Force Atlas 2 Layout

This Force Atlas 2 layout is a continuous, classic force-vector algorithm. A linear-linear model where the attraction and repulsion is proportional to distance between nodes. A unique adaptive convergence speed to allow the graph to converge more efficiently is adopted.



3. Geolayout

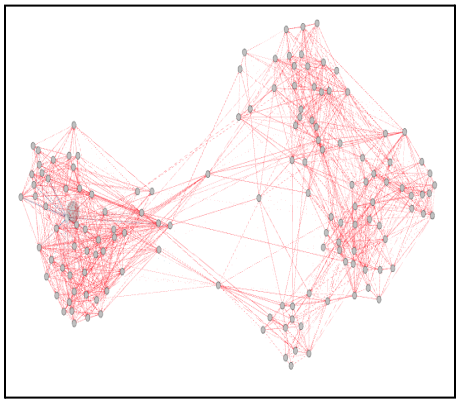
The GeoLayout uses latitude/longitude coordinates to set nodes position on the network. Several projections are available, including Mercator which is used by Google Maps and other online services.

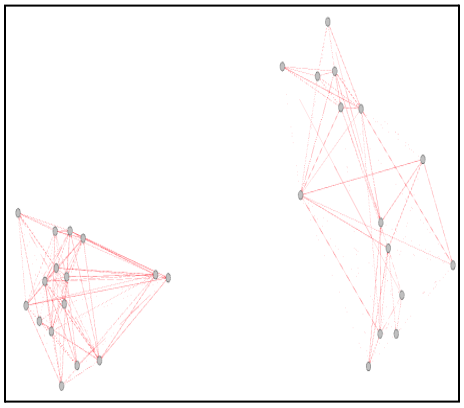


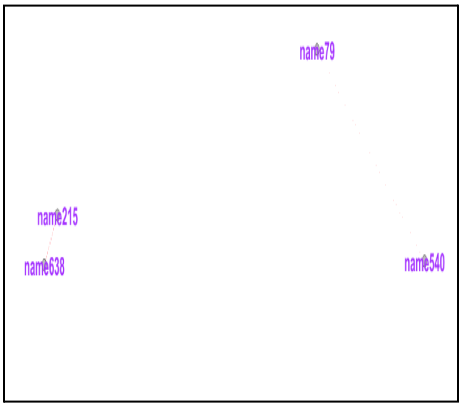
SHREYA SRIRAM
195001106
CSE B

Analyzing the network with respect to nodes

The filtered network after filtering out the top 10%, 2%, 1% of the mailers from the degree values is shown below. This shows the most influential persons in the network.

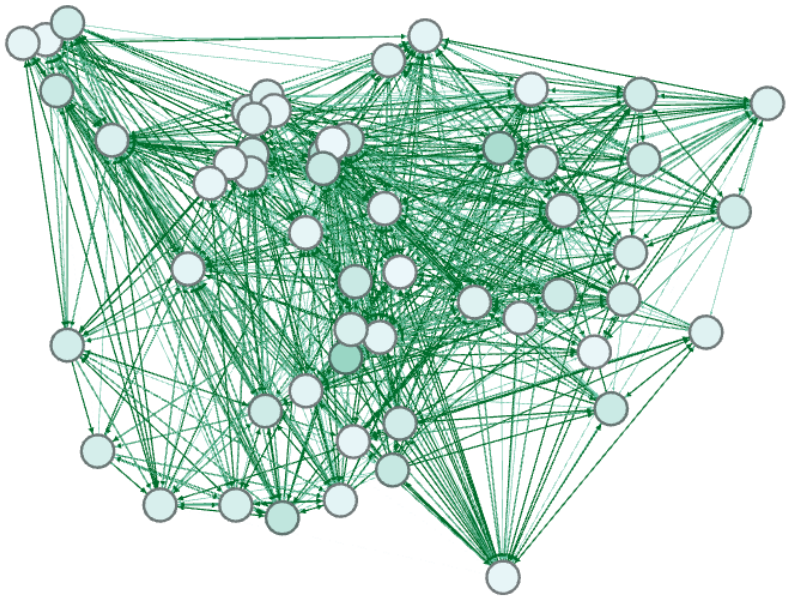






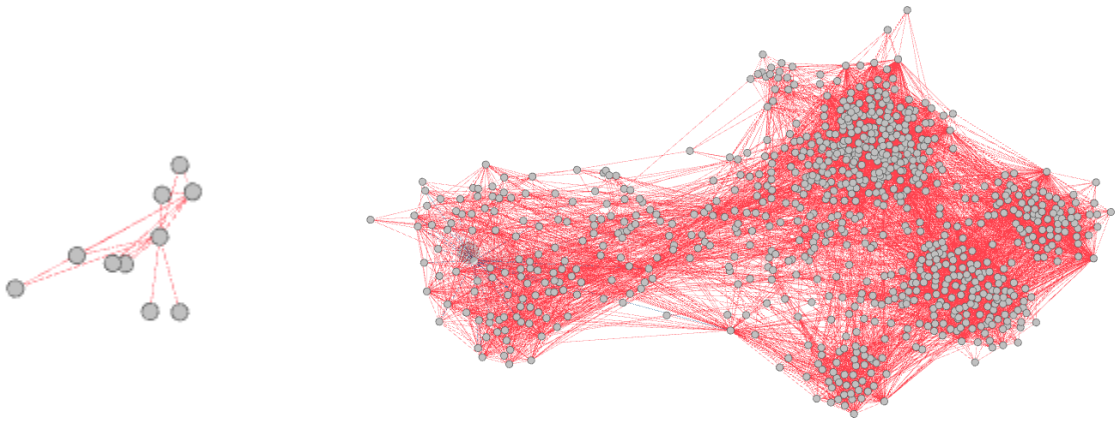
The mailing network has degrees in the range of 3-209, which means all the persons have sent/received at least 3 mails. Name79 is the person who has maximum communication (209) and Name897 has the least communication (3).

The overall distribution of the network degrees is shown as below, the nodes shaded in darker shades of green indicate stronger communication.



SHREYA SRIRAM
195001106
CSE B

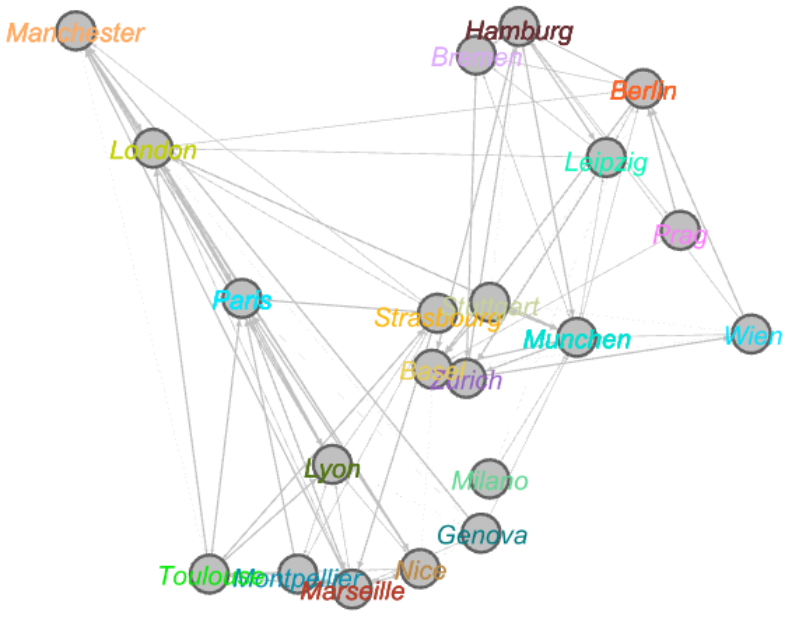
For Name897, the ego network for depth 1 and 3 is shown as below, implying that it's densely connected to many other nodes.



Analyzing the network with respect to geographic locations

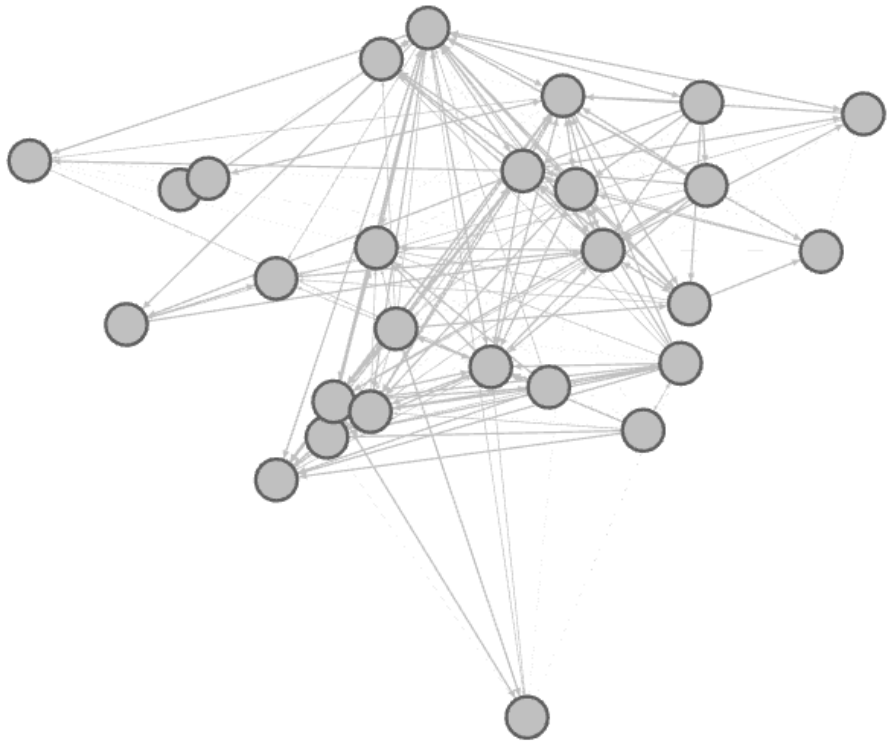
From the geolayout, we can observe that Hamburg has the highest number of communications and Koln has the least. This could be due to the effective mail communication system in Hamburg.

Top 25% of the cities with best mailing facilities are shown as follows



SHREYA SRIRAM
195001106
CSE B

The ego network for Hamburg, this shows the best connected cities in the network.



CONCLUSION

The mailing system in cities across Europe is studied using a social network and visualized using both node-centered and geography centered layout. Also, the most connected actors and cities are identified, densely connected networks form clusters and centrality and clustering metrics are computed and plotted.