MGMT 590 – Multivariate Analysis and Network Data

Final Project by Abhishek, Kavya, Timothy, and Sohan

Final Project Report

Background of the dataset:

The dataset comprises anonymized email interactions within a large European research institution, focusing solely on internal communication. Represented as a directed graph, nodes represent individual members, and edges denote emails sent between them. Each node is associated with one of 42 departments, aiding in intra- and inter-departmental analysis and community detection.

The core network, a subset of the larger "email-EuAll" network, consists of 1005 nodes and 25571 edges. The largest weakly connected component (WCC) encompasses 986 nodes and 25552 edges, while the largest strongly connected component (SCC) includes 803 nodes and 24729 edges. The average clustering coefficient is 0.3994, suggesting moderate clustering. There are 105461 triangles, with a closed triangle fraction of 0.1085, indicating prevalent triangular communication loops. The network's diameter is seven, with a 90-percentile effective diameter of 2.9, highlighting its small-world nature. More details about the dataset can be found in our project proposal document.

Problem Statement:

The primary objective of analysing the email communication network within the research institution is to derive actionable insights into the patterns of interactions among members and assess the overall structure and efficiency of internal communications. This analysis aims to identify key influencers within the network, understand the nature of departmental interactions, and evaluate the robustness of communication channels.

Significance and Impact of Analysis:

The proposed analysis of the email communication network within the research institution offers substantial value to both its strategic direction and operational efficiencies. Constructing a network model of email communications establishes a foundational visualization and quantitative framework, encapsulating the flow of information among individuals. This model enables the identification of complex interaction patterns and anomalies, enhancing our understanding of how communications are distributed across the institution. Analysing betweenness centrality illuminates crucial bridges in the network, facilitating information flow between different parts of the organization. Understanding closeness centrality provides insights into how quickly information can disseminate, aiding in the design of effective communication strategies and preparation for potential disruptions in the information flow.

Exploring inter-departmental communication patterns reveals insights into the collaborative structure of the institution, highlighting whether existing organizational structures support optimal collaboration and information exchange. Addressing these issues can lead to more integrated and effective collaborative efforts, enhancing research outputs and institutional cohesion. Focusing on in-degree and out-degree centrality metrics informs about active recipients and senders of information, aiding in managing information overload and leveraging individuals for spreading important messages or driving change initiatives. Employing probabilistic models for network analysis enables predictions about future communication patterns and potential changes in network structure, empowering leadership to make informed decisions based on modelled outcomes of different organizational changes or external impacts on the communication network.

Network Analysis:

Network Construction:

As a first step, we try to visualize the network. For this project, we are using packages available in Python to visualize and analyse the network. The packages "NetworkX" and the module "Pyplot" from the "Matplotlib" package are used to visualize the network.

The visual representation of the network is given in the figure 1 (next page). It is observed that there is a central cluster with a lot of internal communication and there are nodes with sparse communication. The nodes with sparse communication are located mostly on the outside of the network plot, while the densely linked network is in the centre of the visual.

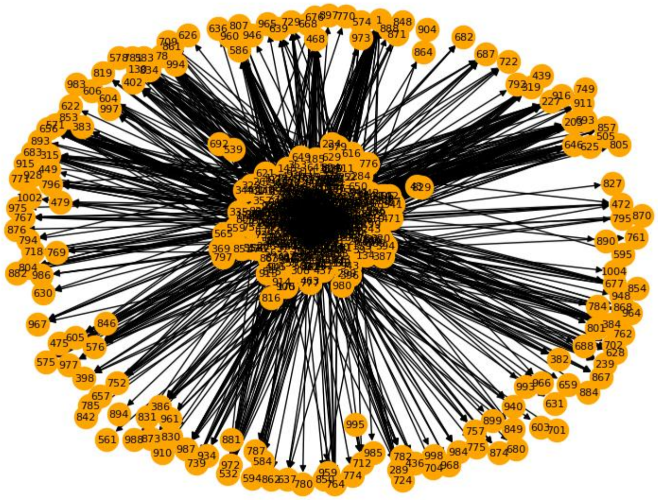


Figure 1 - Visualization of the network

As an additional checkpoint, we check if the network is a simple graph or a multi graph. A simple graph is a graph in which there is a at most one edge between any pair of nodes, and Githere are no self-loops (reference: *Kolaczyk and Csardi 2020, Chapter 2*). To perform this check, we use the 'isinstance' function the classes 'nx.Graph' and 'nx.MultiGraph'. Our output suggests that the graph is a simple graph and not a multi graph.

Descriptive Statistics:

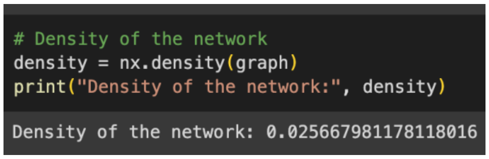
Now, we look at some of the basic stats of the network and interpret the values. The metrics are either at a node-level, an edge-level, or a network-level.

MGMT 590 – Multivariate Analysis and Network Data

Final Project by Abhishek, Kavya, Timothy, and Sohan

1. **Density:**

We start with computing the density of the network, which is defined as the ratio of the number of edges in the network to the total number of possible edges in the network.



```
# Density of the network
density = nx.density(graph)
print("Density of the network:", density)

Density of the network: 0.025667981178118016
```

Figure 2 - Density of the network

The density of our network is 0.025 which indicates a weak network. This indicates that there are limited interactions among the nodes. This also indicates a potential to improve the collaborations among these entities.

2. **Degree distribution:**

Since the network is directional, we compute the basic descriptive stats for the nodes. The following table summarizes the same:

Parameter	Value
Minimum In-degree	0
Maximum In-degree	211
Minimum Out-degree	0
Maximum Out-degree	333
Average In-degree	25.28
Average Out-degree	25.28

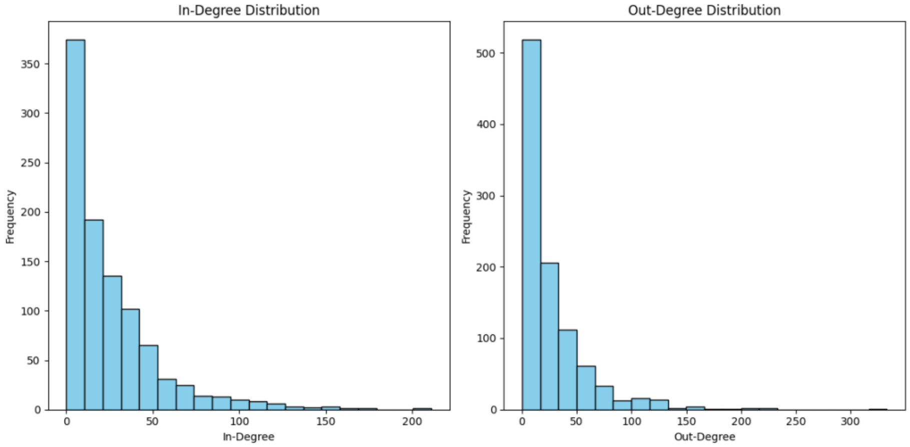


Figure 3 - Degree Distribution in the network

The above figure (figure 4) also displays the histograms for the in-degree and out-degree distributions. We used 20 bins to study the network distribution. Given the sparsely connected graph, as expected, we see positively skewed graphs that indicate that a higher number of nodes have low in and out degrees. Based on the above analysis, we also look at the nodes with the highest in-degrees and out-degrees. This analysis would give us an idea about the most active senders / receivers in the network. We ignore the least active senders and receivers as they are a huge number based on the degree distribution.

Top 5 nodes with highest incoming emails	
Node Number	*In-degree*
160	211
62	178
107	168
121	156
86	153

Top 5 nodes with highest outgoing emails	
Node Number	*Out-degree*
160	333
82	226
121	221
107	203
86	201

It is noteworthy that some of the top receivers of emails are also senders of emails. Those nodes have been highlighted in both the tables. These are clearly the most active players in the network.

3. **Geodesics**:

The geodesic distributions for the in-degree and out-degree have been shown in Figure 5. It is interesting to observe that the in-degree and out-degree geodesic distributions have similar shapes. It is also interesting to note that there are no nodes with infinite geodesic distance. This means that it is possible for a node to reach another node indirectly in the given network. This is a strong signal indicating the potential for building a denser network with more connections if the institution takes an initiative.

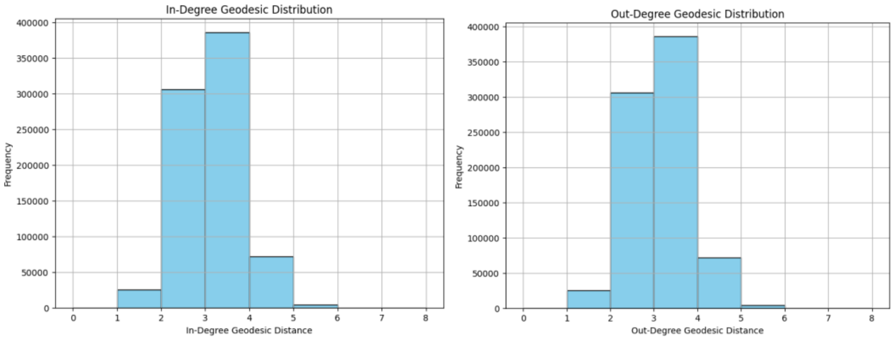


Figure 4 - Geodesic Distributions

4. **Clustering Co-efficient:**
The clustering co-efficient of the network as given in the data description is 0.3994, while we observed in our analysis that the clustering co-efficient is 0.3727.

5. **Closeness Centrality:**
Closeness centrality measures how close a node is to all other nodes in the network. Given that we computed the geodesic distributions, let's compute the top 5 nodes with the highest closeness centrality based on in-degree and out-degrees:

Top 5 nodes with highest CC – in degree	
Node Number	*In-degree*
160	0.568
82	0.530
121	0.524
107	0.513
86	0.512

Top 5 nodes with highest CC – out degree	
Node Number	*Out-degree*
160	0.458
62	0.445
107	0.441
434	0.436
121	0.435

In this metric too, there are some common nodes with leading closeness centrality for in- and out-degrees.

6. **Directed Network Reciprocity:**
Directed network reciprocity measures the extent to which edges in a directed network tend to form reciprocated pairs. In this context, it implies the number of nodes which have interacted both ways, i.e. sending and receiving emails. The value of this metric is observed to be 0.711. This indicates that 71% of the nodes often engage with each other through email communications.

With the basic descriptive statistics of the network in place, we move on to identifying the top 10 hubs and authorities in the network.

Kleinberg's Hub and Authority Measures:

In the context of an email communication network, a node with a high hub score would represent an individual who sends emails to many other individuals within the institution. These individuals act as hubs of communication, facilitating the dissemination of information or messages to various recipients. On the other hand, a node with a high authority score would represent an individual who receives emails from many other individuals within the institution. These individuals are seen as authoritative sources of information, as they are frequently included in email conversations or are recipients of important messages.

Top 10 Hubs	Hub Score	Top 10 Authorities	Authority Score
160	0.0107	160	0.0072
82	0.0097	107	0.0069
121	0.0096	62	0.0067
107	0.0088	434	0.0065
62	0.0082	121	0.0064
249	0.0080	183	0.0060
434	0.0075	128	0.0059
183	0.0072	256	0.0057
86	0.0070	249	0.0057
114	0.0064	129	0.0056

Given that our network is a directional network, we analysed and visualized the hubs and authorities in the network. The same python package 'NetworkX' is utilized for computing the hubs and authorities. The top 10 hubs and authorities are summarized in the table above.

MGMT 590 – Multivariate Analysis and Network Data

Final Project by Abhishek, Kavya, Timothy, and Sohan

Departmental Communications and Clustering:

As a first step, we visualize the network based on the department data that we are given. The visual is shown below:

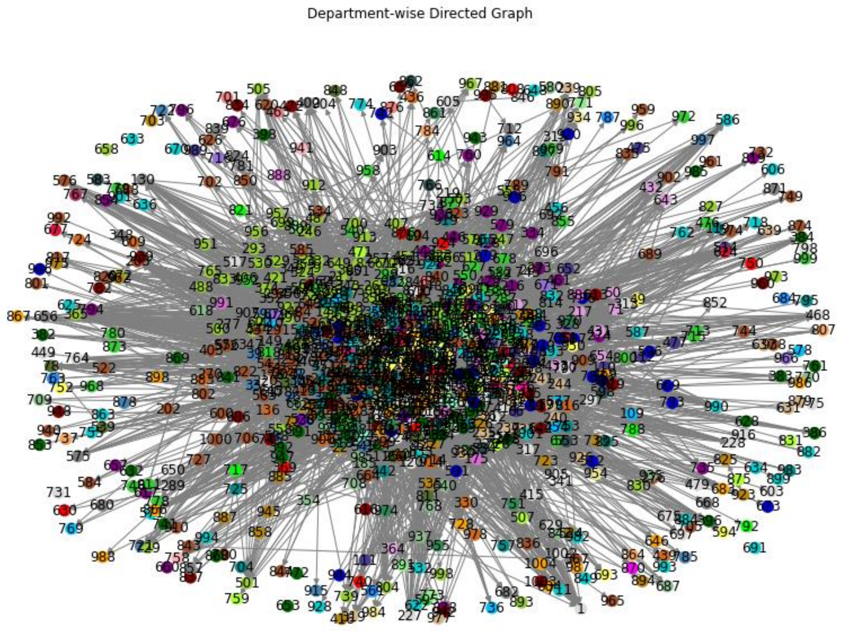


Figure 5 - Department Color-coded network visual

Since we have 1005 nodes, it becomes difficult to visualize the entire network in one diagram. Hence, we try to visualize only the strongly connected components in the network. On some additional analysis, we find out that there are 803 nodes in the strongly connected components' list.

The network diagram of the SCC is shown in Figure 7 (next page). The top 5 departments with nodes in the SCC have been summarized in the below table. The table has been sorted by the number of nodes in the SCC.

Department ID	No. of Nodes in SCC	Total number of nodes in department	Percentage of nodes in SCC
4	88	109	80.73%
14	80	92	86.95%
15	48	55	87.27%
1	48	65	73.85%
7	38	51	74.51%

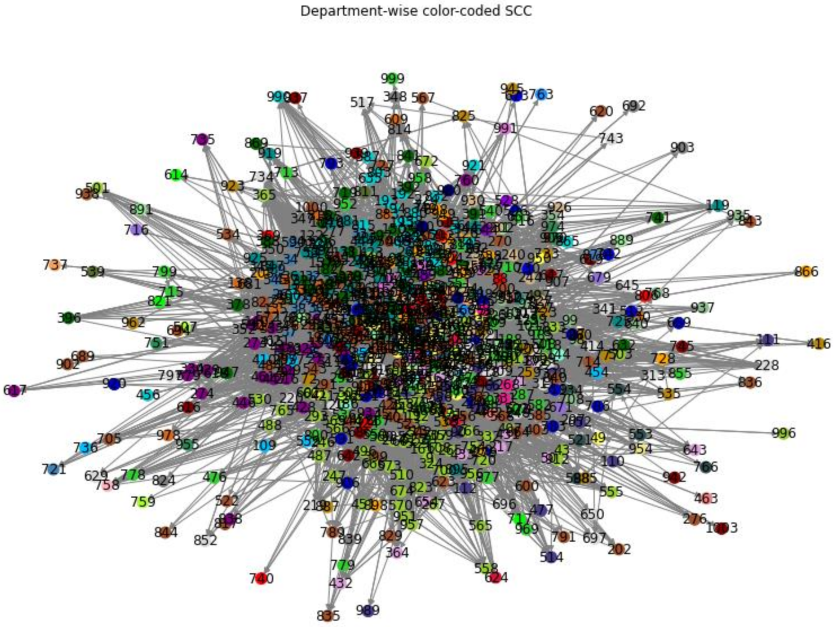
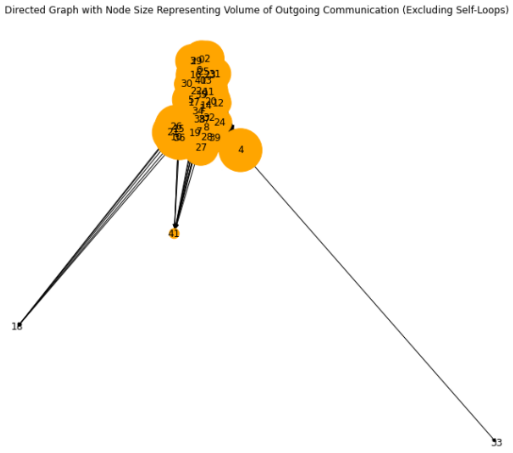


Figure 6 - Network diagram of the SCC

It is clear to infer from the above diagram and table that the more the number of nodes in a department, the more likely it is to be a part of the strongly connected components network. This is a very positive indicator that there is a lot of communication internally in a department.

In addition to the above, we can also look at the inter-departmental communication. The following figure is a visual representation with the node size representing the busiest departments (i.e., larger the node size, more communication between those departments).


Directed Graph with Node Size Representing Volume of Outgoing Communication (Excluding Self-Loops)
4
41
18
33

Top 10 Inter-departmental mails		
From Dept.	To Dept.	No. of mails
36	4	229
4	5	170
21	22	159
4	36	155
22	21	154
5	4	151
36	15	151
4	0	150
36	14	136
36	1	128

Figure 7 - Inter-departmental Communication and top 10 inter-departmental mails

From the above figure, we can infer that 33, 18, and 41 are the least busy departments, and 36, 4, 5, 21, and 22 are the busiest departments.

As a next step, we create clusters based on communication frequency between the nodes. We use k-means clustering and divide the network into 10 clusters. The figure given below displays the clustered network:

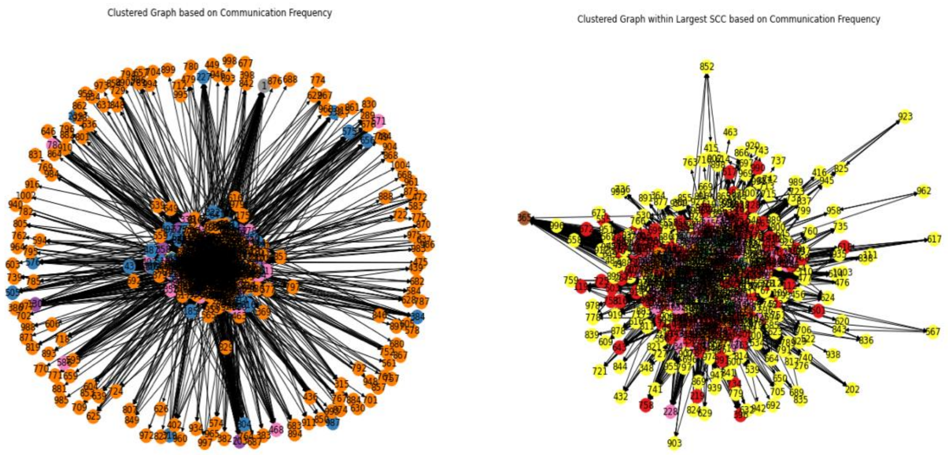

Clustered Graph based on Communication Frequency
Clustered Graph within Largest SCC based on Communication Frequency

Figure 8 - Network divided into clusters. Second diagram is for SCC.

We also perform a cluster analysis on the SCC and divide it into 10 clusters. The output for the same is available in Figure 8 above. The clusters have also been ranked based on the communication strength within the cluster. There are two ways to look at the strength of communication in each cluster: 1. We can look at the sum of the no. of communications within each cluster, and 2. We can look at the average of the communication frequency within each cluster (sum of communications in each cluster / no. of nodes in the cluster). We think that using the sum is a better option as it checks for the total number of communications in the cluster. Based on this assessment metric, Cluster 3 is the busiest cluster followed by Cluster 8 as the second busiest cluster.

The summary of the same is given in the table below (sorted based on averages for users who prefer to go with the second metric):

Cluster Details	Avg. Communication Score	No. of nodes in cluster	Sum of communications
Cluster 9	185.66	3	557
Cluster 6	143.57	7	1005
Cluster 2	113.22	18	2038
Cluster 5	89.68	25	2242
Cluster 0	66.51	55	3658
Cluster 8	47.37	83	3932
Cluster 3	35.54	111	3945
Cluster 7	24.21	150	3632
Cluster 1	14.63	179	2619
Cluster 4	3.66	355	1301

<u>Note</u>: We also created 42 clusters to compare with the departments and analyse if it has any patterns like the inter-departmental patterns. But we did not find any. The data is available in Appendix table 1.

Probabilistic Network Models:

<u>Erdos – Renyi Random Network:</u>

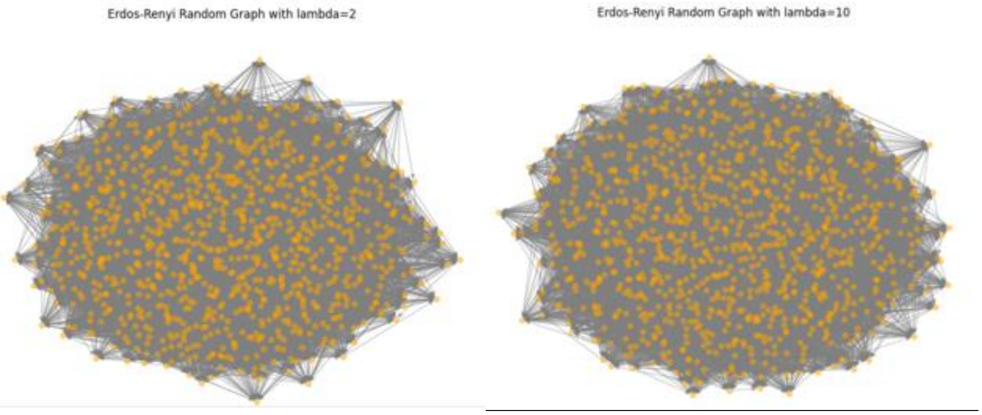
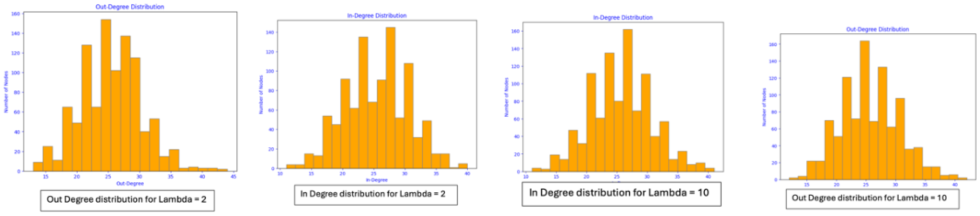

Erdos-Renyi Random Graph with lambda=2
Erdos-Renyi Random Graph with lambda=10

Figure 9 - Erdos Renyi Random Graphs for lambda values 2 and 10



Out Degree distribution for Lambda = 2

In Degree distribution for Lambda = 2

In Degree distribution for Lambda = 10

Out Degree distribution for Lambda = 10

Figure 11: ER Model degree distributions

For the Erdos-Renyi random network model, we tried to construct networks with a higher lambda value (2, 10). The in/out-degree distributions have been provided in the figure 11 and can be compared to the actual in/out-degree distributions of our network. The degree distributions of the Erdos-Renyi model suggest a normal distribution-like degree distribution while our network has a positively skewed distribution. Nevertheless, the Erdos-Renyi models suggest that the network can be connected more densely with the same number of nodes and edges.

The Erdős-Rényi model may not perfectly represent the nuances of real-world communication networks, as it assumes equal probability of connection between any two nodes, ignoring potential community structures or preferential attachment (where popular nodes get more links). However, it helps in understanding the baseline properties of random connectivity, indicating the non-random patterns of communication that could be due to departmental structure, workflow processes, or communication hierarchical policies.

<u>Scale-Free Network:</u>

We also checked our network to see if it has any resemblance with a scale-free network.

MGMT 590 – Multivariate Analysis and Network Data

Final Project by Abhishek, Kavya, Timothy, and Sohan

Based on the degree distribution, many nodes have been observed to have a low degree. This is an indicator of a scale-free network. Additionally, there are a lot of hubs observed in the network. More than 20% of the nodes have a higher degree than the mean and median degrees of the network, indicating a higher presence of hubs. This adds to the validation that our network exhibits some properties of a scale-free network.

Small World Network:

Small-world networks typically exhibit a high clustering coefficient, meaning that nodes tend to cluster together. Based on our analysis, we observed that with a rewiring probability of 0.01, 0.05, 0.1, the clustering coefficients for the small world networks generated were 0.71, 0.628, 0.54 respectively. Compared to 0.37 of the email network, we infer that if we sufficiently increase the rewiring probability, we can achieve more similarity between both networks. The global efficiency of the network measures how efficiently information can be transmitted across the entire network and the email network had a global efficiency of 0.33 which was slightly lower than the average global efficiency of the small world networks generated – 0.403.

Practical Implications and Conclusions:

In summary, we noticed that the nodes 160, 107, 121, and 86 are the busiest nodes. Interestingly, they all belong to the same department (dept. number 36). To make broad guesses about the department, it could be the IT support team to which users are sending emails about issues and receiving resolutions. In a broad sense, this team is somehow connected to the entire research institute. It is also observed that these nodes act as hubs and authorities, which reinforce our conclusion that this team is a centre for operations for the entire institution. Department 36 (which has the busiest nodes) is also the busiest department with the highest number of inter-department communication as seen in the analysis. The degree distribution suggests that there is an immense potential to make this a more connected network and change the degree distribution to a more homogenous distribution. However, the current network is robust enough to be able to each every node through some path, as seen in the geodesic distribution.

The Erdos Renyi random network analysis suggests that the current network can be more clustered with the same number of nodes and edges. In a way, one of the highly clustered networks can show the way forward on how the network can be reshaped to be more connected and collaborative. Keeping the real-world nuances in mind, this random network can be modified to still give a network with a more homogenous degree distribution. Given that our network has properties similar to that of a scale free network, it can be inferred that while the network is resilient to node failures and there is smooth information flow, it is also vulnerable to failure since it has a high number of hubs.

Overall small-world networks combine the properties of high clustering and short average path lengths, enabling efficient communication between nodes at the global scale. Although, this email network is not completely representative of a small network, it exhibited some characteristics such as a relatively high global efficiency, relatively high clustering coefficient as well as existence of shortcuts. The information flow showed that individuals who send or receive many emails) are more likely to be connected to nodes with low degrees (e.g., individuals who send or receive fewer emails), and vice versa.