1. Introduction: This is about diving into a fascinating dataset specifically, the communication network within a European research institution. The goal is to unravel the hidden story behind this complex web of interactions. By understanding this network, it'll be possible to identify key players, pivotal departments, and how communication flows between them. The impact of this analysis can be significant, helping department heads and administrators make informed decisions. They could streamline resource allocation, pinpoint potential communication roadblocks, and build bridges for better collaboration between departments.

The proposed approach is thorough: it involves using various techniques such as centrality measures to spot influential individuals, community detection algorithms to uncover the network's structure, and assessing how tightly-knit groups form. Moreover, they'll dive into the degree distribution to understand how individuals are connected and even employ epidemic modeling to predict information spread. The plan also includes visually representing the network, using colors to differentiate departments and node sizes to highlight influential figures.

2. Methodology

2.1 Preliminary Data Analysis

2.1.1 Data Collection

Approach: Our journey began with the acquisition of a rich dataset from the Stanford Network Analysis Platform, featuring a complex web of email communications within a European research institution. This dataset is not just a collection of numbers and connections; it's a window into the intricate fabric of institutional communication.

Anticipated Insights: By unpacking this dataset, we aim to not only map out the network of interactions but also to attach real-world significance to these connections, revealing how individuals and departments interact within the institutional ecosystem.

2.1.2 Dataset Loading and Network Initialization

Approach: Using NetworkX, a powerful tool in our analytical domain, we transformed raw data into a structured network graph. Each node in this graph represents an individual, and each edge symbolizes an email interaction. To add a layer of complexity, we incorporated departmental data, assigning labels to each node to reflect departmental affiliations.

Anticipated Insights: This step sets the stage for our analysis, equipping us with a graph that is not just a network but a representation of the institutional social connection. It allows us to see beyond mere numbers and understand how departmental affiliations might influence communication patterns.

2.1.3 Exploratory Data Analysis (EDA):

Approach: Here, we dive into the network's anatomy, examining its scale and structure. We calculated basic statistics like the number of nodes and edges and the average degree. But it's more than just numbers; it's about understanding the heartbeat of the network, how connected it is, how dense or sparse, and the balance or imbalance in departmental representation.

Anticipated Insights: This initial approach will give us a snapshot of the network's health and composition. It's like taking the pulse of the institution's communication network, setting the stage for a deeper diagnostic analysis.

2.2 Methods Applied

2.2.1 Centrality Measures

Approach: We used degree centrality and betweenness centrality as our guiding stars to navigate the network. These measures aren't just statistical tools; they help us spotlight the individuals who stand as pivotal points or bridges in the communication landscape.

Anticipated Insights: By identifying these key players, we can start to understand who drives communication within the institution and how information flows through these central nodes.

2.2.2 Community Detection:

Approach: Employing sophisticated community detection algorithms, we peeled back the layers of the network to reveal the underlying community structures. This step is alike to

uncovering the hidden tribes within the organization, groups that might be defined by more than just departmental lines.
Anticipated Insights: This analysis will shed light on the fabric of inter-departmental interactions, showing us how the institution organically clusters and where the lines of collaboration and division lie.

2.2.3 Clustering Coefficient
Approach: By calculating the average clustering coefficient, we will understand the tendency of individuals to form tight-knit groups. This coefficient will tell us about the network's social circles, how close-knit they are and how this closeness might impact information flow and collaboration.
Anticipated Insights: We expect to uncover insights into the network's social dynamics, such as which departments or groups form close alliances and how these groupings affect the broader communication network.

2.2.4 Degree Distribution Analysis:
Approach: Analyzing the degree distribution allowed us to delve into the connectivity patterns of the network. This analysis is more than a mere count of connections; it's a probe into the network's backbone, revealing how evenly or unevenly communication channels are spread out across individuals.
Anticipated Insights: From this, we can discern the presence of hubs or influential nodes and understand how these pivotal points shape the network's overall connectivity.

2.2.5 Epidemic Modeling (SIR Model)
Approach: Using the SIR model, a technique borrowed from epidemiology, we simulated how information or trends might ripple through the network. This modeling is not just about predictions; it's a way to visualize the pathways of influence and the speed at which information travels across the network.
Anticipated Insights: The model will help us identify potential communication bottlenecks or super-spreaders of information, offering insights into how to streamline or enhance information flow.

2.2.6 Visualization Techniques
Approach: We are going to bring the network to life with visual representations, employing color-coding for departments and varying node sizes based on centrality measures. This step is not just about creating a visual aid; it's about translating complex data into a form that is both insightful and accessible.
Anticipated Insights: Through these visualizations, we aim to provide a clear and intuitive view of the network's structure. It's about seeing the forest for the trees , understanding how individual pieces fit into the larger puzzle of institutional communication.

3. Result (Observation with findings)

3.1 Extracting preliminary data information:
Number of Nodes (Individuals): 1005
Number of Edges (Email Interactions): 16,706
Average Degree (Average Number of Connections per Individual): Approximately 33.25
42 Departments: Diverse representation within the network.
Largest Department: 109 members, suggesting a significant and potentially influential group.
Smallest Departments: One member each, likely with specialized or unique roles.
Varying Membership: Notably uneven distribution across departments.
Potential Influence Imbalance: Larger departments may hold more sway, indicating potential hierarchical structures.

3.2 Network visualization:
3.2.1 Static Network Visualization (Fig 1):

- Represents the email network with nodes colored by department memberships.
- Highlights clusters indicating close internal communication within departments.
- Shows inter-department links, signifying communication between different departments.
- Reveals diverse interactions: some departments have widespread connections, while others are more isolated.

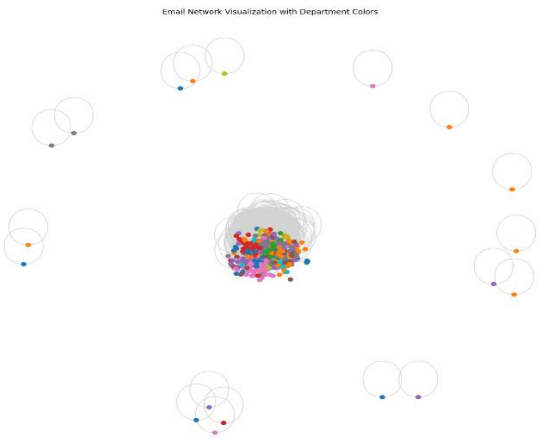

Fig 1: Static Network Visualization

3.2.2 Zoomable Interactive Network Visualization (Fig 2,3):

- Allows interactive exploration of the network.
- Hover over nodes to view details and explore connections visually.
- Offers a closer examination of department structures and interactions.
- Enables a detailed view of the network's dynamics and inter-departmental connections.

Fig 2: Interactive Network Visualization (Before zoom)

Fig 3: Interactive Network Visualization (After Zoom)

3.3 Centrality measures and community detection (Fig 4): The visualization below displays the email network with nodes sized according to their degree centrality (indicating how connected they are) and colored based on the communities detected algorithmically. This provides insights into:

Central Nodes: Larger nodes represent individuals with many connections, potentially playing pivotal roles in communication.

Community Structure: Different colors represent different communities as detected by the algorithm, showing how the network is divided into clusters of closely interacting individuals.

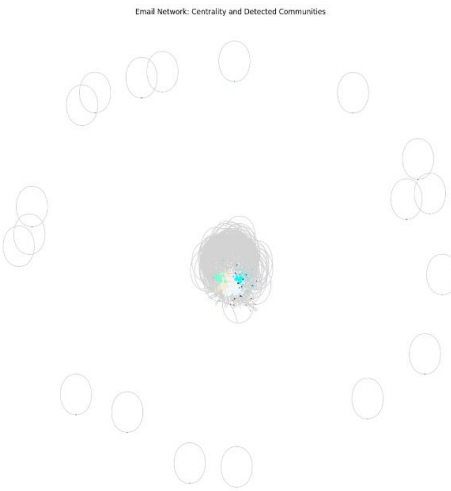

Fig 4: Centrality measures and community detection

3.4 Clustering Coefficient: The average clustering coefficient for the network is approximately 0.40. This suggests a moderate level of cliquishness, where individuals tend to form localized clusters or groups.

3.5 Degree Distribution (Fig 5): The degree distribution histogram illustrates how connections are distributed among individuals in the network. It shows a pattern where most people have only a few connections, while a small number have a substantial amount. This kind of distribution, known as right-skewed or "long-tail," is quite common in real-world networks. It's a way to visualize that most folks are modestly connected, but a select few are highly connected within the network.

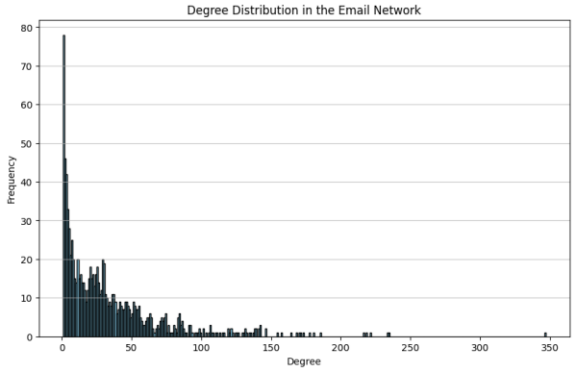

Fig 5: Degree Distribution

3.6 Betweenness Centrality Visualization (Fig 6): The visualization showcases the email network, where node sizes reflect their betweenness centrality. Larger nodes represent individuals crucial in linking various network segments. These pivotal figures facilitate communication by serving as bridges, allowing numerous shortest communication paths to pass through them. They play a pivotal role in connecting diverse parts of the network, promoting efficient information flow and connectivity among different sections.

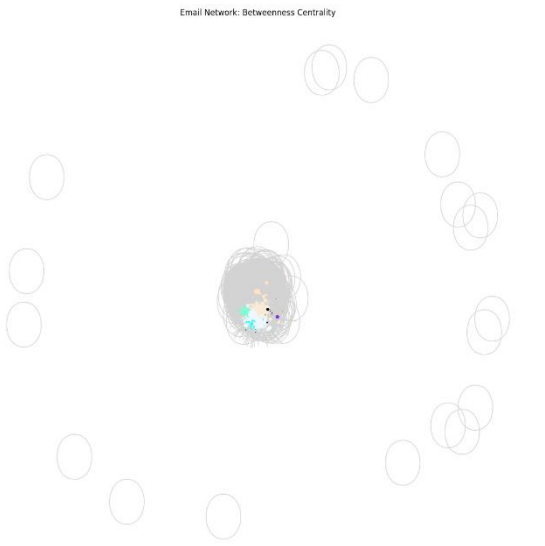

Fig 6: Betweenness Centrality

3.7 Community Overlap with Departments (Fig 7): The side-by-side visualization compares actual department memberships (left) with algorithmically detected communities (right) in the email network.

On the left, colors represent the "ground-truth" department memberships, unveiling intra-department connections and inter-department interactions.

On the right, colors signify algorithm-detected communities. While there are similarities with departments, differences suggest communities might span multiple departments or departments could be fragmented into smaller communities.

This comparison delves into the relationship between formal organizational structures (departments) and informal communication patterns (detected communities), shedding light on their alignment and divergence.

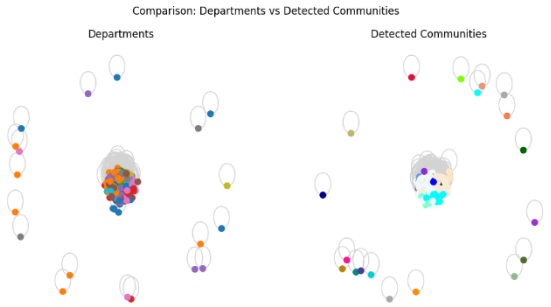

Fig 7: Community Overlap with Departments

3.8 Subnetwork of a Specific Department (Fig 8): The visualization zooms into Department 4's network, showcasing its internal communication (blue nodes) and external connections (red nodes) with other departments. The blue nodes expose how members within the department are interconnected, while the red nodes reveal connections to individuals from other departments. Gray nodes represent other network parts not directly linked to Department 4. This focused view unveils communication dynamics within and around the specific department, providing insights into its internal structure and external interactions.

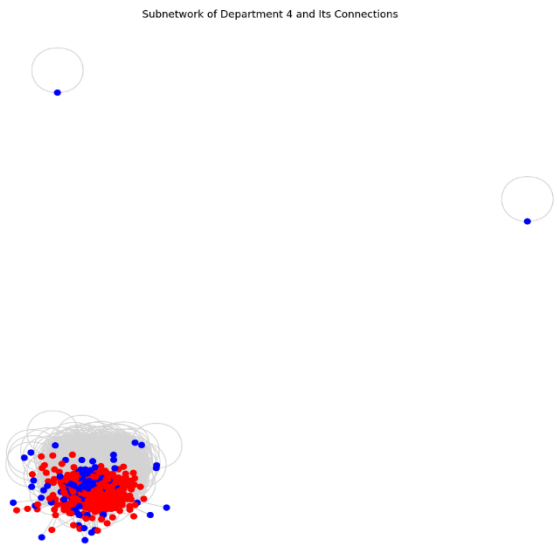

Fig 8: Subnetwork of Department 4

3.9 Eigenvector Centrality (Fig 9): The visualization sizes nodes based on eigenvector centrality, considering a node's connections and the centrality of those it connects to. Larger nodes indicate individuals not only well-connected but linked to other influential figures. This metric spotlight influential clusters where certain individuals hold sway due to their connections to other influential members, unveiling critical nodes in the network.

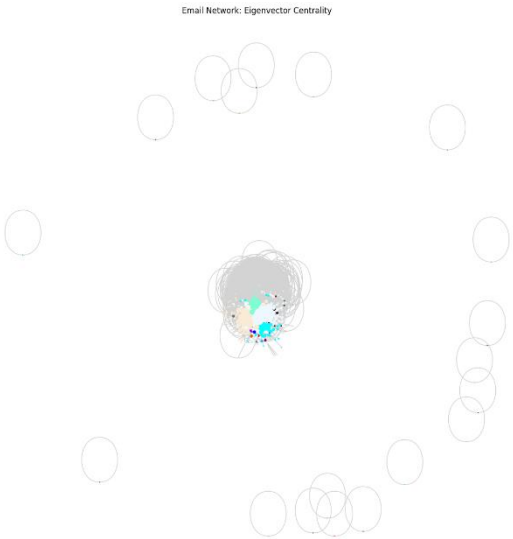

Fig 9: Eigenvector Centrality

3.9 Network Density and Transitivity: The network's density is around 3.3%, indicating a moderately sparse network, typical for large networks where not everyone is directly connected. The transitivity at approximately 26.7% signifies a moderate tendency for nodes to form triangles, suggesting a reasonable likelihood of mutual connections between connected nodes. These metrics offer an overview of the network's structure, revealing moderate local clustering and relatively sparse connections considering the network's size.

3.11 Assortativity by Department: The assortativity coefficient, around 0.316, indicates a moderate inclination for individuals to connect within their own department. This suggests a

tendency for intra-departmental connections over inter-departmental ones, emphasizing a moderate preference for communication and connections within one's own department in the network.

3.12 Closeness and Harmonic Centrality: Node 160 holds significant centrality in the network, marked by both the highest closeness centrality (approximately 0.574) and harmonic centrality (about 655.5). These metrics signify that this individual is notably close to others and serves as a vital link for rapid information dissemination. Their pivotal position suggests they likely play a crucial role in network connectivity, potentially acting as a key information broker within the network.

3.13 Network Robustness (Fig 10): The plot demonstrates the impact of removing nodes, starting with highly central ones, on the network's largest connected component size. Initially, as central nodes are removed, there's a gradual decline in the component size. However, beyond a certain point, the network experiences rapid fragmentation, signifying a loss of overall connectivity. This showcases the network's moderate resilience to the removal of a few central individuals but highlights a substantial decline in robustness as more key nodes are removed.

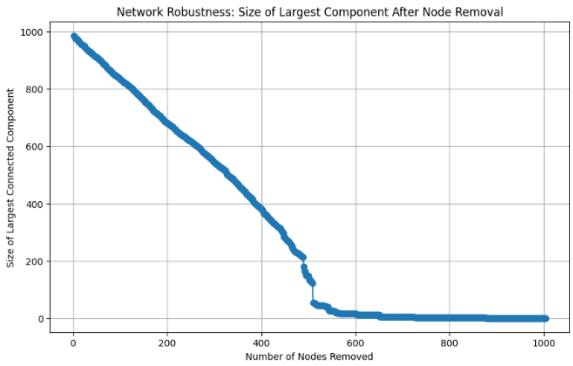

Fig 10: Network Robustness

3.14 Degree Correlation: The degree correlation in the network measures around -0.011, hinting at a slight disassortative mixing by degree. This suggests a minor inclination for well-connected nodes to link with less well-connected nodes rather than with other highly connected nodes. However, the value being close to zero indicates that this tendency is not notably strong within this network.

3.15 Small-World Property (Largest Connected Component): The largest connected component of the network displays small-world characteristics. With an average shortest path length of approximately 2.59, it suggests that individuals can be connected within a few steps. The relatively high clustering coefficient, around 0.41, indicates a tendency for nodes to form closely-knit groups. These traits align with small-world networks, showcasing efficient connectivity with short paths between individuals and a propensity for clustering despite the network's scale.

3.16 Influence of Node Removal on Centrality Measures: The simulation involved removing the top 5 central nodes based on eigenvector centrality. Following this, the most affected individual (Node 6) experienced a minute increase of approximately 0.0573% in degree centrality. This subtle change suggests a minor redistribution of connections without significantly altering the network's structure. The analysis indicates that while the removal of highly central nodes does impact the network, it doesn't notably reshape the centrality landscape for the remaining nodes, signifying a level of resilience in the network's structure against the loss of key individuals.

3.17 Visualization of Shortest Paths: The first visualization (Fig 11) showcases the shortest paths originating from Node 100 to all other nodes within its connected component. Blue edges represent these paths, illustrating how information might disseminate from Node 100 throughout the network. Node 100, depicted in red, acts as the source from which these shortest paths emerge. Gray nodes and edges represent the rest of the nodes and connections in the largest connected component. This visualization offers a clear insight into Node 100's network reach, emphasizing its potential influence and the

pathways through which it can connect with other nodes.

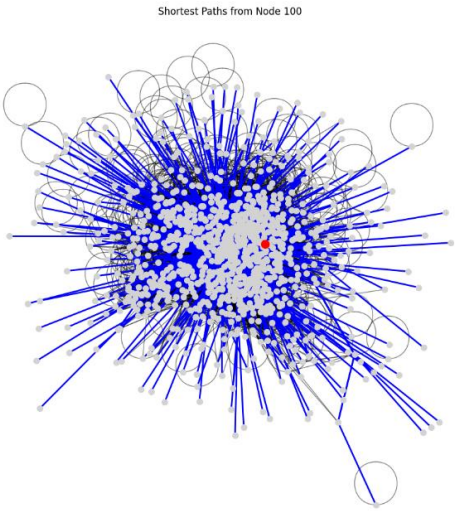

Fig 11: Visualization of Shortest Paths from node 100

The second visualization (Fig 12) depicts the shortest paths originating from Nodes 100, 200, and 300 in the network's largest connected component. Colored edges—red for Node 100, green for Node 200, and blue for Node 300—represent these paths, demonstrating how information might spread from each node.

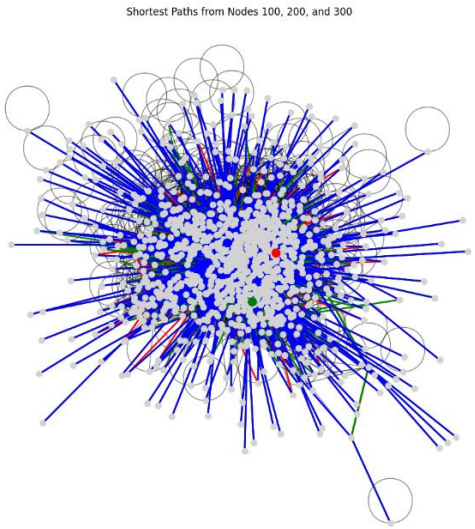

Fig 12: Visualization of Shortest Paths from node 100, 200 and 300.

3.18 Shortest Path Length Distribution (Fig 13): The histogram illustrates the distribution of shortest path lengths within the network's largest connected component. The most common shortest path length of 2 indicates that numerous pairs of individuals are just two steps away. As the path length increases, the frequency of those lengths decreases, with very few paths exceeding a length of 4. This distribution emphasizes the network's small-world characteristics, depicting how most individuals are closely connected by a relatively small number of steps. These visualizations offer a comprehensive view of network connectivity, shedding light on information spread and typical distances between individuals.

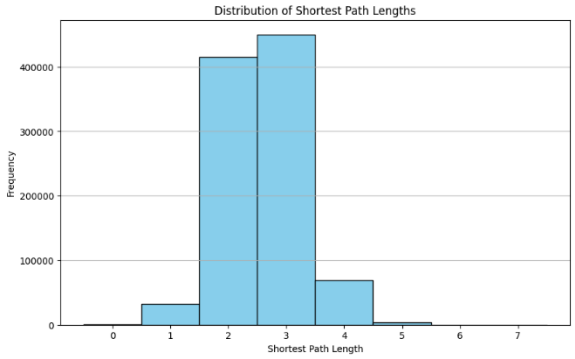

Fig 13: Shortest Path Length Distribution

3.19 Bipartite analysis: In this case, attempting to create a bipartite graph projection based on individuals and departments isn't feasible due to the nature of the original network. The network's edges represent email communication between individuals rather than direct connections between individuals and departments. Hence, conducting a bipartite analysis might not be suitable for this network, as it lacks the necessary structure for a clear separation into distinct sets of nodes.

3.20 Independent Cascade Model (Fig 14): The Independent Cascade Model simulation, initiated with Nodes 100, 200, and 300, resulted in the activation of 715 nodes (It varies) in the network. The visualization highlights these activated nodes in green, representing those influenced or informed during the simulation. The spread

reached a significant portion of the network, displaying the potential for rapid information propagation and revealing clusters and pathways through which the influence traveled. This analysis offers insights into the network's dynamics, showcasing how information or influence spreads and identifying influential nodes and connections. The starting nodes significantly determined the reach and pace of the spread.

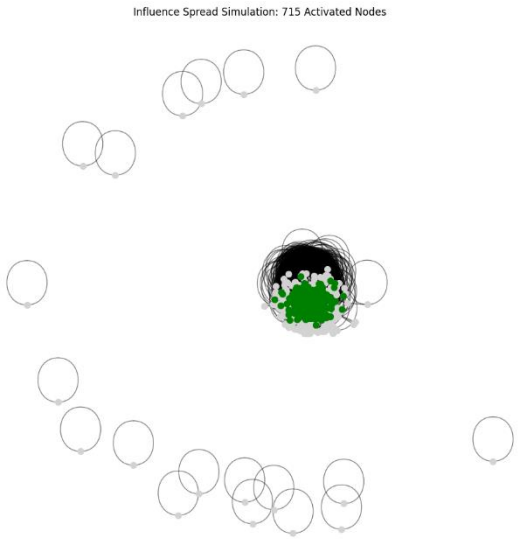

Fig 14: Independent Cascade Model (ICM)

3.21 Influence Maximization: Based on the network's degree centrality, nodes 160, 121, 82, 107, and 86 have been identified as potential top influencers. Their high degree centrality suggests significant connections, indicating their potential to maximize the spread of influence within the network.

3.22 Epidemic Modeling (SIR Model) (Fig 15): The SIR model simulation with Node 160 as the initial infected node resulted in 937 nodes (It varies) recovering and 68 remaining susceptible. The recovered nodes are shown in blue, while the initially infected node (Node 160) is highlighted in red. The significant infection and subsequent recovery within the network demonstrate the potential for rapid epidemic spread. Certain nodes and connections played pivotal roles in this spread, while the existence of susceptible nodes at the simulation's end suggests varying susceptibility or isolation in different network parts. This analysis underscores the network's vulnerability to epidemic spread and the influence of specific nodes and connections on this propagation.

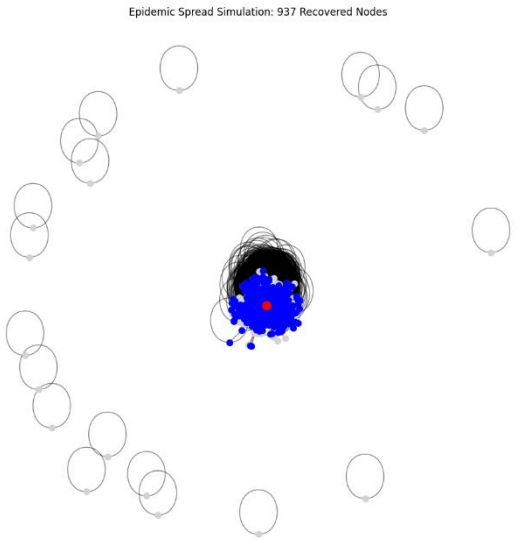

Fig 15: Epidemic Modeling (SIR Model)

3.23 Departmental hubs visualization (Fig 16): The visualization depicts departmental hubs within the network, highlighting nodes with significant connectivity in their respective departments. Each color represents a distinct department, showcasing hubs that likely hold central roles, possibly serving as key points for information exchange within their departments. Additionally, the presence of connections outside their departments suggests their involvement in inter-departmental communication, emphasizing their potential influence beyond their own departmental boundaries.

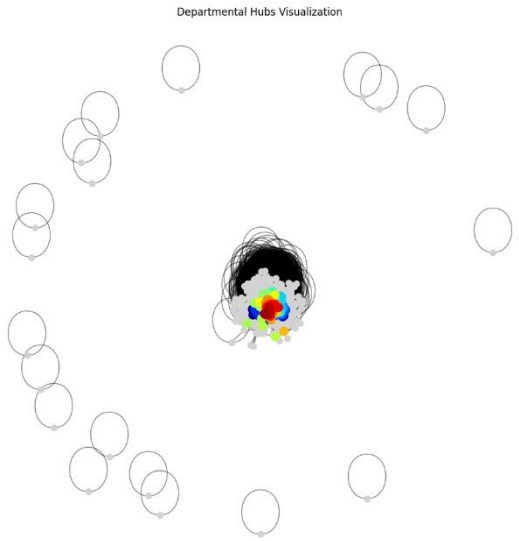

Fig 16: Departmental hubs visualization

3.24 Degree Distribution Visualization (Fig 17): The histogram illustrates the degree distribution in the network. A right-skewed distribution indicates that most nodes have relatively low degrees, while a few nodes exhibit exceptionally high degrees. The presence of these highly connected nodes suggests pivotal roles in the network's communication structure. Additionally, the skewness hints at a scale-free property, where a handful of highly connected nodes (hubs) coexist with many nodes having fewer connections, influencing the network's overall connectivity and structure.

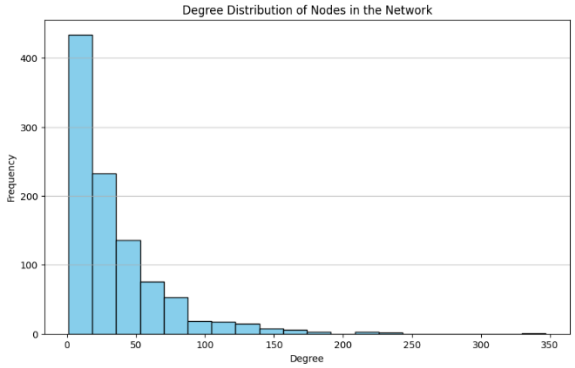

Fig 17: Degree Distribution Visualization

3.25 Department wise insights

3.25.1 Department size (Fig 18, 19): The visualization presents department sizes in the email network, sorted by the number of individuals in each. Departments 4 and 14 emerge as the largest, housing 109 (10.8%) and 92 (9.2%) individuals, respectively, suggesting their potential significant influence within the network due to their size. This variation, ranging from large departments with over 100 members to smaller ones, hints at varied internal communication capacities and potential connections to other departments, portraying a diverse landscape within the network.

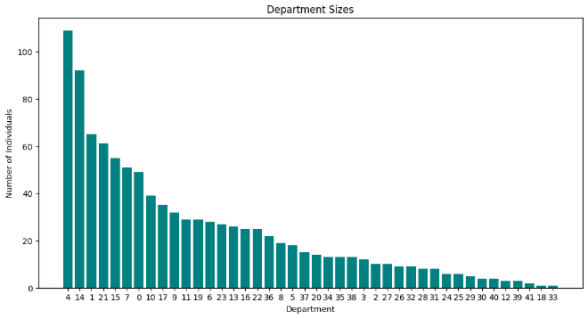

Fig 18: Department size (Bar chart)

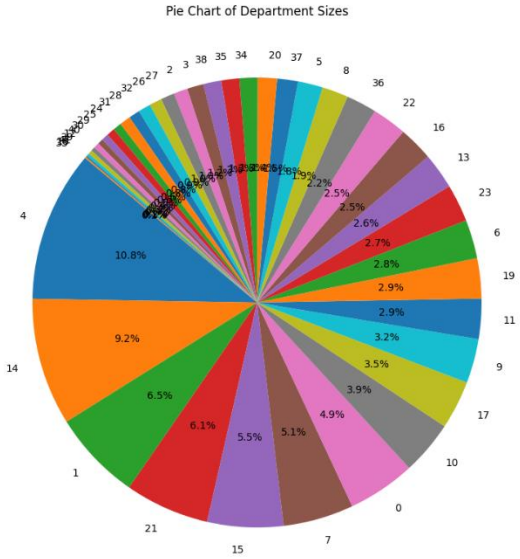

Fig 19: Department size (Pie chart)

3.25.2 Intra-Departmental Connectivity (Fig 20): The analysis of intra-departmental connectivity reveals diverse densities of connections within departments. Departments like 12, 25, and 40 exhibit a density of 1.0, signifying complete connectivity among their members, possibly indicating robust intra-departmental collaboration or communication. However, there's considerable variation across departments, with some displaying high internal connectivity while others have less dense connections. Departments with higher densities likely foster more cohesive internal communication and collaboration, while those with lower densities may have more fragmented internal structures.

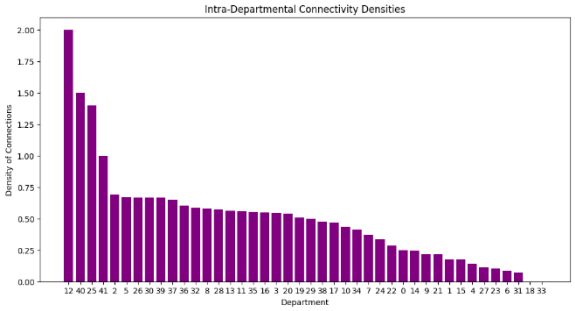

Fig 20: Intra-Departmental Connectivity Densities

3.25.3 Inter-Departmental Links (Fig 21): The visualization represents the inter-departmental communication network, showcasing connections between departments. Department pairs like (4, 36) and (4, 5) exhibit the highest number of connections, indicating robust inter-departmental communication or collaboration. Departments such as 4 and 36 appear as hubs, having multiple strong connections with other departments, signifying their central role in inter-departmental communication. The network structure illustrates departmental interconnections, revealing clusters and key bridges between departments, offering valuable insights into collaboration and communication dynamics within the institution.

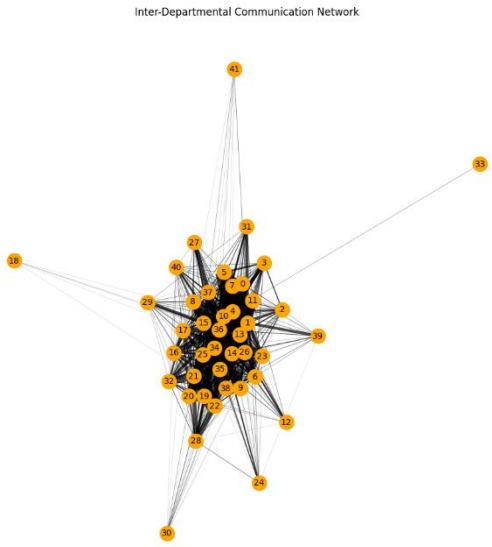

Fig 21: Inter-Departmental communication network

3.25.4 Heatmap of Departmental Interactions (Fig 22): The heatmap visualizes the frequency of interactions between departments within the email network. Darker shades signify higher interaction frequencies, showcasing hotspots where departments are closely connected, potentially due to shared interests or collaborative projects. Lighter areas indicate departments with fewer interactions, suggesting potential isolation or specialized functions. Overall, the heatmap offers an overview of departmental interconnectivity, emphasizing both active communication channels and opportunities for enhanced collaboration.

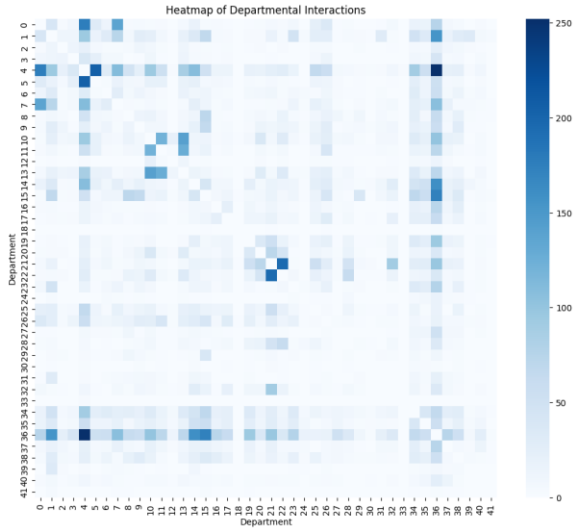

Fig 22: Heatmap of Departmental Interactions

3.25.5 Bar Chart of Departmental Email Activity (Fig 23): The bar chart depicts the total number of connections (inbound and outbound) for each department, reflecting their email activity within the institution. Departments 4, 14, and 36 stand out as the most active, having the highest number of connections, implying their heavy involvement in email communication. There's a notable variation in email activity across departments, ranging from highly active to those with comparatively fewer connections.

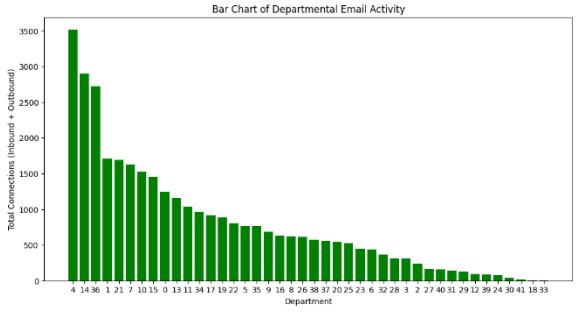

Fig 23: Bar Chart of Departmental Email Activity

3.25.6 Stacked Bar Chart of Intra- vs Inter-Departmental Connections (Fig 24): The stacked bar chart showcases the intra-departmental (blue) and inter-departmental (red) connections for each department, providing insights into their communication patterns. Departments like Department 4 exhibit a balanced number of connections within and outside the department, implying active communication both internally and externally. Conversely, some departments show a higher proportion of intra-departmental connections, signaling a stronger focus on internal communication. Departments with a notable count of inter-departmental connections potentially act as bridges or hubs in the network, facilitating communication between different areas of the institution.

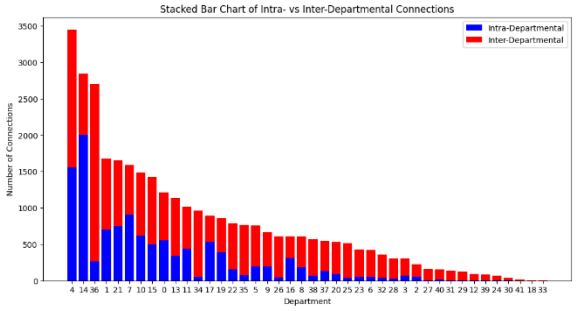

Fig 24: Stacked Bar Chart

3.25.7 Treemap of Departmental Connectivity (Fig 25): The treemap-style bar chart offers a visual representation of departmental connectivity, with each bar's length corresponding to the total number of connections (inbound and outbound) for the respective department. Larger bars indicate higher connectivity, suggesting potential influence and centrality within the network. Departments with extensive connectivity might serve as central communication hubs or bridges within the institution. This visualization allows for a straightforward comparison between departments, highlighting their relative scale of connectivity.

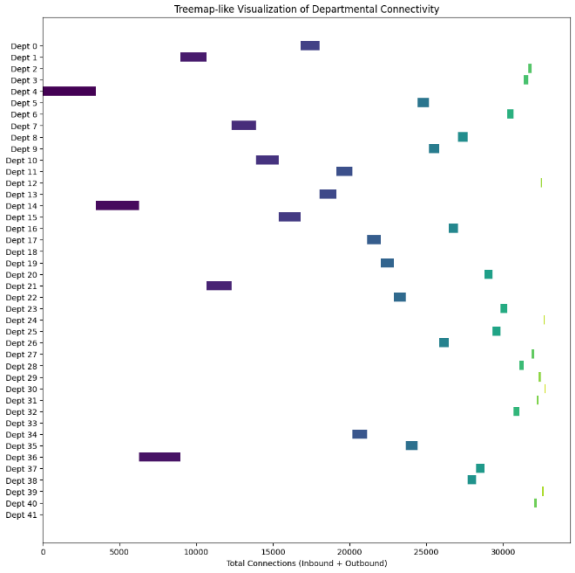

Fig 25 Tree map

4. Conclusion

Analyzing the communication network in this European research institution unveiled key insights. Firstly, larger departments like 4 and 14 emerged as potential hubs due to their substantial sizes, potentially influencing communication dynamics. The intricate web of connections, revealed through various metrics, highlighted the presence of highly connected individuals and some isolated departments.

The visualizations exposed clusters of internal communication and strong inter-departmental connections, shedding light on potential communication bridges between different areas of the institution. Moreover, the examination of network metrics like degree distribution emphasized the presence of both highly and modestly connected nodes, reflecting a diverse communication landscape.

Moving forward, the analysis could explore deeper departmental roles and individual influences, adding granularity to the understanding of communication patterns. Furthermore, observing the network's evolution over time and predicting future trends could be a promising avenue for enhancing collaborative efficiency within the institution.