Assignment 1: Find the communities of your graph

Network Analysis

Delaram Doroudgarian
5881909

Seyedeh Aida Atarzade Hosseini
6936800

Academic Year 2024/2025

1 Graph Characteristics of Email-Eu-core Network

The Email-Eu-core dataset models email communication between members of a large European research institution. Each node represents an individual and each directed edge indicates that one person sent at least one email to another. For the purpose of structural analysis, we considered the undirected version of the graph.

1.1 Original Network View and Directional Analysis

Before performing any cleaning or transformation on the dataset, we visualized the original directed graph representing the email interactions. Figure 1 shows the raw structure of the graph.

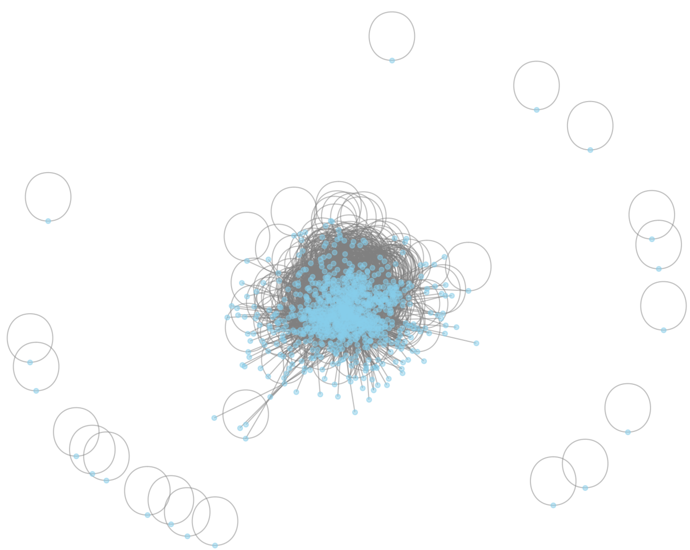

Figure 1: Visualization of the original directed Email-Eu-core graph before any preprocessing.

In this version of the network:

- **Average In-Degree:** 25.28
- **Average Out-Degree:** 25.28

These values are equal because, in any directed graph, each edge contributes exactly once to an in-degree and once to an out-degree (Figure 2). This confirms the consistency of the raw dataset before cleaning.

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

Figure 2: in-degree and out-degree formula

1.2 Basic Properties

The network has the following properties:

- Number of nodes before cleaning: **1005**
- Number of nodes after cleaning: **986**
- Number of edges before cleaning: **25571**
- Number of edges after cleaning: **16064**
- Density: **0.0331**
- Diameter (of the largest connected component): **7**
- Average clustering coefficient: **0.4**
- Degree assortativity coefficient: **-0.0257**

Interpretation: The density of 0.033 indicates a sparse network, which is typical for real-world communication networks, such as email exchanges. In this case, while each individual doesn't communicate with many others directly, the overall network remains well connected. The diameter of 7 means that even the most distant individuals in the largest connected component are separated by at most 7 steps (i.e., 7 email connections). This reflects the "small-world" nature of organizational email systems, where any two members can usually reach each other through a short chain of colleagues. The average clustering coefficient of 0.4 shows a strong tendency for people who email the same person to also email each other — typical in workgroups or departments where email communication forms closed triads. Finally, the assortativity coefficient of -0.0257 suggests that highly active users (e.g., managers or admins with high degree) tend to communicate with less active users (e.g., regular staff), rather than forming elite-only communication clusters.

1.3 Degree Distribution

Figure 3 shows the degree distribution of the Email-Eu-core network. Most nodes have low degree (less than 10), and only a few nodes exhibit very high connectivity (hubs). This long-tailed distribution suggests the presence of a scale-free structure.

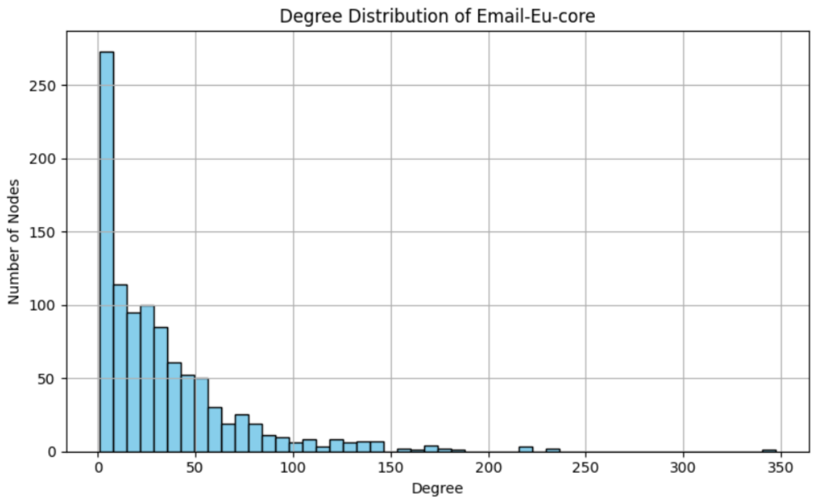

Figure 3: Degree distribution of the Email-Eu-core network

1.4 Normalized Degree Distribution

To visualize the global degree behavior, we show the nirmalized frequency in Figure 4. The curve decays gradually, confirming that while most nodes have low degrees, a significant portion of the connectivity is concentrated in a few hubs.

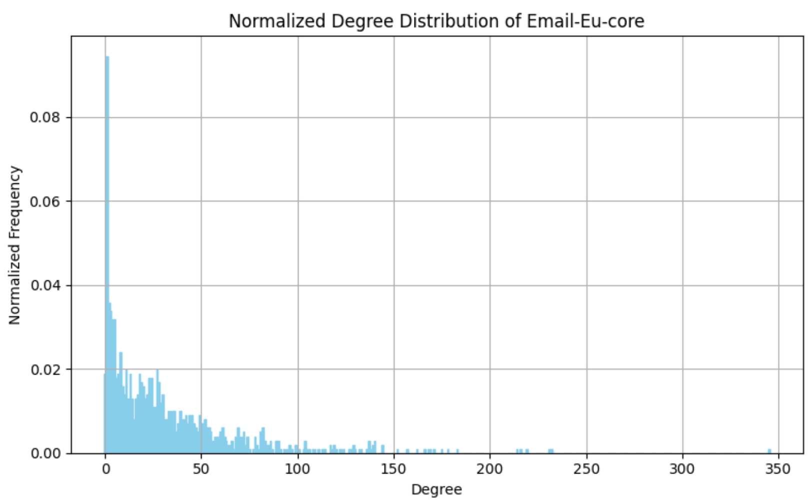

Figure 4: Normalized degree distribution

Conclusion: The Email-Eu-core network exhibits typical features of real-world communication graphs: sparse but connected, with local clustering and disassortative hub connectivity. The degree distribution strongly suggests a scale-free topology. The scale-free topology of the Email-Eu-core network is evident from its long-tailed degree distribution. Most users exchange emails with only a few colleagues, while a small number of individuals act as hubs, communicating with hundreds.

2 Community Analysis

2.1 Community Detection using Louvain and Leiden Algorithms

we applied two popular community detection algorithms on the network: Louvain and Leiden. Both algorithms aim to find communities by maximizing modularity but differ in their optimization techniques.

The Louvain algorithm detected a total of **8** communities in the Email-Eu-core network. Each community was checked for internal connectivity and confirmed to be a connected subgraph. This ensures that the community structure found by the algorithm is consistent with actual communication groups in the network. The execution time for the Louvain algorithm on this dataset was approximately **0.27 seconds**. This reflects the high performance and practicality of the method when dealing with real-world social or communication networks of moderate size. The structure and size of the communities suggest meaningful divisions in the organization, possibly corresponding to departments or tightly collaborating workgroups. The largest community includes over 200 nodes.

The Leiden algorithm also detected **8 communities**, but executed faster, taking only **0.16 seconds**. This performance difference is expected, as Leiden includes improvements over Louvain by refining partitions and avoiding disconnected communities through more stable optimization.

Community Structure and Performance Comparison

In order to evaluate the performance of the Louvain and Leiden community detection algorithms, we compared them based on the number of communities identified, modularity values, and execution time.

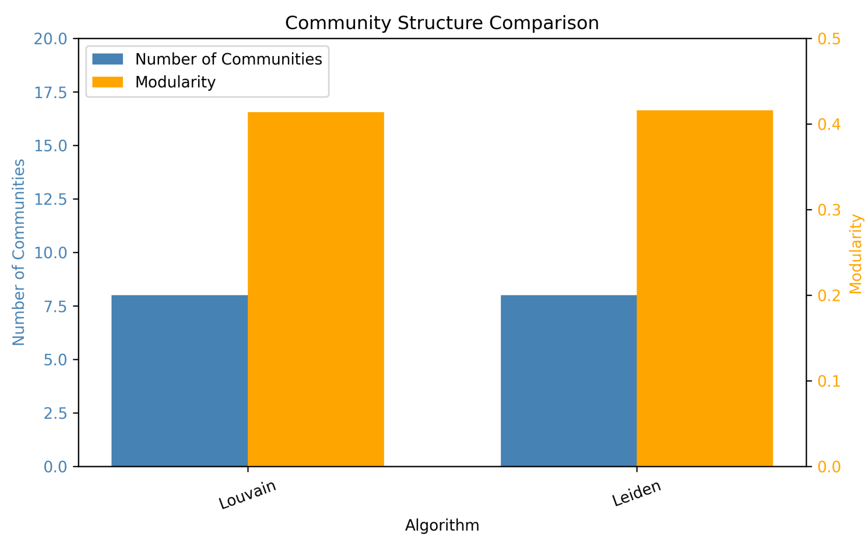

Figure 5: Comparison of Louvain and Leiden algorithms in terms of the number of detected communities and modularity.

As seen in Figure 5, both algorithms produced 8 communities, but Leiden yielded slightly higher modularity(=0.416) than Louvain (=0.413), indicating better internal cohesion.

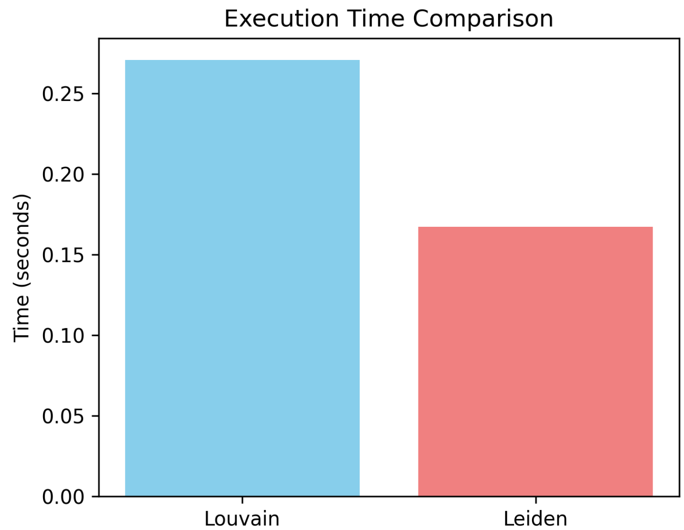

Figure 6: Execution time of Louvain and Leiden algorithms.

Figure 6 shows that the Leiden algorithm outperformed Louvain in terms of execution speed, confirming its better scalability.

3 Comparison of Community Detection Algorithms

In order to evaluate the consistency between the community detection algorithms applied to the Email-Eu-core network, we computed the Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) between the partitions obtained from the Louvain and Leiden algorithms.

These metrics allow us to assess how similar the two detected community structures are, even in the absence of ground truth labels.

3.1 Metrics Explained

Normalized Mutual Information (NMI) measures the amount of information shared between two partitions. It ranges from 0 (completely independent) to 1 (identical), and it is particularly robust when the number of clusters differs between the two partitions. NMI is based on entropy and mutual information and reflects the structural similarity of the clustering assignments.

Adjusted Rand Index (ARI) evaluates the agreement between two clusterings by comparing all pairs of nodes and checking whether they are assigned to the same or different communities in both partitions. ARI corrects for chance by adjusting the Rand Index, making it a more conservative and precise metric in the presence of many small groups.

3.2 Results

The partitions obtained from the Louvain and Leiden algorithms were aligned to the same node order, and the following results were obtained:

- **NMI (Normalized Mutual Information)**: 0.7900
- **ARI (Adjusted Rand Index)**: 0.6983

These values indicate a strong, though not perfect, agreement between the two methods. Despite differences in their internal mechanisms — Louvain performs hierarchical modularity optimization, while Leiden incorporates refinements to avoid badly connected communities — both algorithms identified largely overlapping community structures within the Email-Eu-core network.

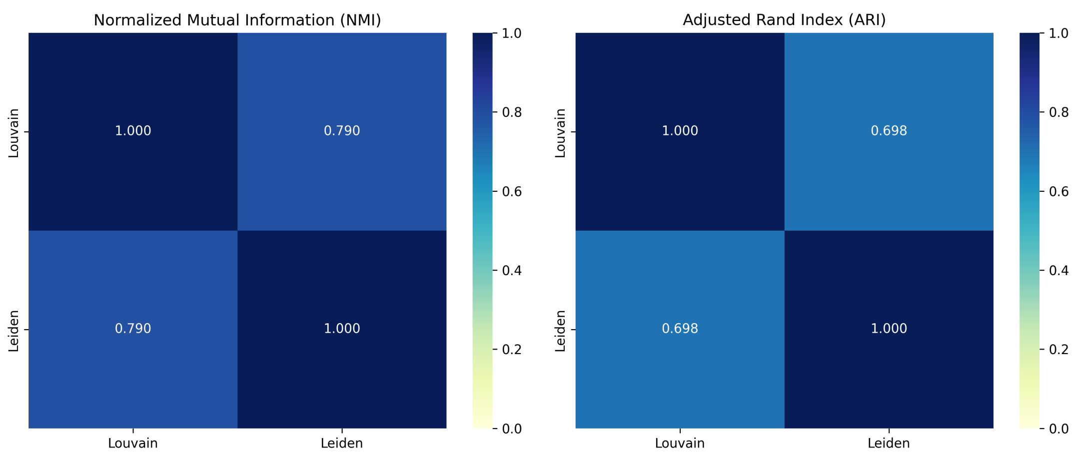

Figure 7: Algorithm similarity heatmaps between Louvain and Leiden community detection results using NMI (left) and ARI (right).

3.3 Interpretation

The NMI value of 0.7900 indicates that the community structures discovered by Louvain and Leiden share a considerable amount of mutual information, suggesting that their overall partitioning is globally similar. The ARI value of 0.6983 supports this finding by confirming that a large portion of node pairs were grouped similarly by both algorithms, although some discrepancies remain.

Together, these metrics suggest that the detected community structures are generally robust and consistent, while allowing for minor differences in finer group assignments. The observed agreement between Louvain and Leiden indicates that both methods are capturing meaningful structural patterns within the Email-Eu-core network.

In summary, Louvain and Leiden produced largely compatible community structures, reinforcing the applicability of modularity-based community detection approaches in analyzing real-world communication networks.

4 Community Visualization

To better interpret the structure of the detected communities, we visualized the Email-Eu-core network using the community partitions obtained from both the Louvain and Leiden algorithms. Each node is colored based on the community it was assigned to.

4.1 Louvain Community Structure

The following graph shows the visualization of communities detected by the Louvain algorithm:

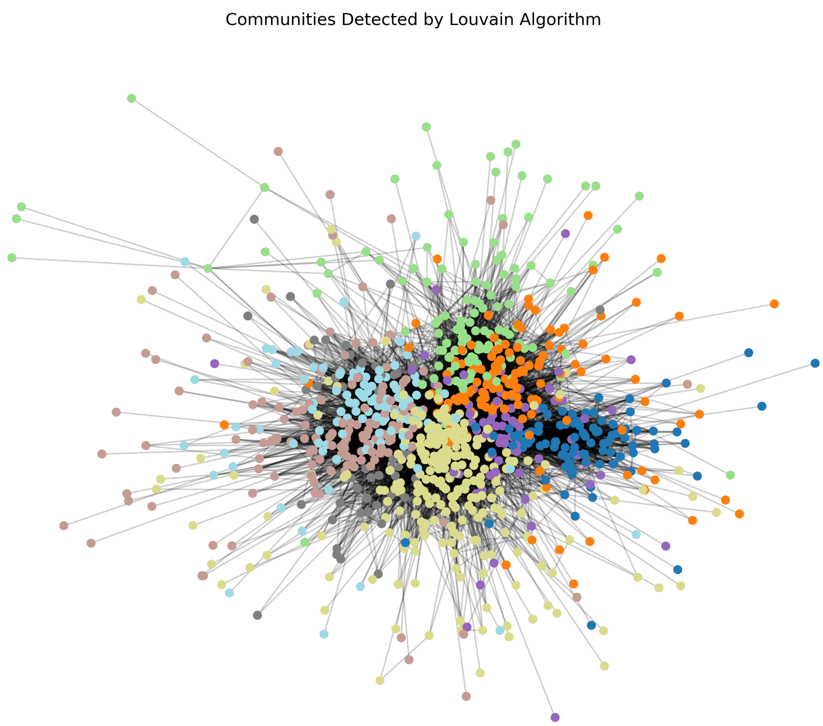

Figure 8: Communities Detected by Louvain Algorithm

The Louvain algorithm identifies a moderately granular community structure, with several well-defined groups clustered around a dense core. While the communities are reasonably balanced in size, some overlapping and loosely connected nodes can be observed at the periphery. The visualization reflects Louvain's tendency to favor modularity optimization, sometimes at the expense of finer internal cohesiveness.

4.2 Leiden Community Structure

The following figure presents the visualization of the communities detected by the Leiden algorithm:

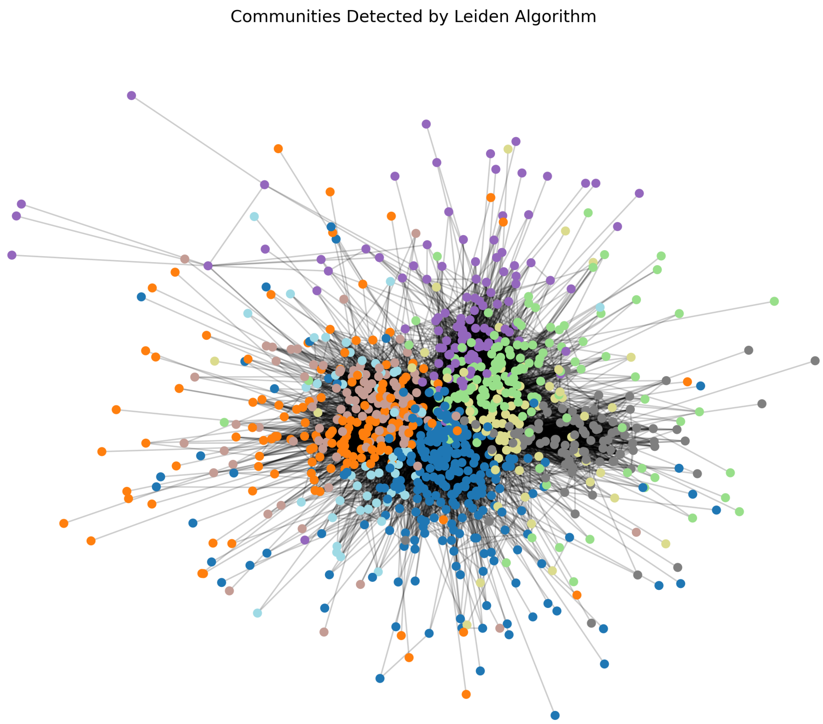

Figure 9: Communities Detected by Leiden Algorithm

The Leiden algorithm reveals a more refined and internally coherent community structure. Compared to Louvain, it produces smaller, more consistent groups, especially at the network's boundaries, while preserving the dense core. This improvement stems from Leiden's guarantee of intra-community connectivity, leading to more reliable community partitions.

In general, both visualizations reinforce the results obtained in previous sections, confirming the reliability and interpretability of the detected community structures. The high overlap between Louvain and Leiden outputs—supported by NMI and ARI metrics—can also be observed visually through the similar yet refined partitioning of the graph.