ADA Assignment 04 — Community Detection

Ayushman Anupam (MDS202411)

Project Link: https://drive.google.com/file/d/1Nv6e6psW2_9MfiH03eAhHW9w9b6FyEzO/view?usp=sharing

Contents

1 Pipeline Overview

This project performs community detection on the Enron Email Network dataset using multiple graph-based algorithms. The workflow includes:

1. Loading and constructing the Enron email communication graph.
2. Building a core subgraph of top-degree nodes (where-ever needed).
3. Applying five community detection algorithms.
4. Evaluating their efficiency, scalability, and interpretability.
5. Visualizing the detected communities for insight into real-world communication patterns.

2 About the Dataset

Dataset: Enron Email Network (Stanford SNAP Repository)

- **Nodes:** 36,692 (email addresses)
- **Edges:** 183,831 (email exchanges)
- **Type:** Undirected Graph
- **Format:** Text file with two columns — `FromNodeId, ToNodeId`
- **Source:** SNAP Enron Dataset

This dataset represents real-world email communication within the Enron Corporation, making it ideal for studying network-based community structures.

3 Problem Statement and Objective

The objective is to detect and analyze communities within the Enron email communication graph to reveal organizational divisions and interaction patterns.

Goals:

- Identify groups of employees with frequent internal communication.
- Compare algorithms based on accuracy, scalability, and interpretability.
- Evaluate real-world performance on a large network dataset.

4 Methods Employed

4.1 1. Girvan–Newman Algorithm

A modularity-based edge betweenness approach that iteratively removes the most central edges to separate communities.
Performance: Accurate for small subsets with clear modular boundaries.
Limitation: Computationally infeasible for large graphs due to $O(VE^2)$ complexity.
Improvement: Use parallelized or approximate versions for scalability.

4.2 2. Louvain Method

A hierarchical modularity optimization algorithm that efficiently detects communities in large networks.
Performance: Produced the strongest and most balanced community structures.
Captured: Hierarchical clusters resembling functional departments.
Limitation: Misses small communities due to modularity resolution limits.
Improvement: The Leiden algorithm enhances stability and captures smaller clusters.

4.3 3. Spectral Method

Based on the eigenvectors of the graph Laplacian matrix for clustering.
Performance: Detected 3–5 major groups with well-defined boundaries.
Captured: High-level partitions (e.g., managerial vs. operational units).
Limitation: Sensitive to sparsity; misses fine-grained clusters.
Improvement: Recursive partitioning or K-Means on spectral embeddings.

4.4 4. Label Propagation

A simple and fast approach where nodes adopt the most common label among neighbors.
Performance: Fastest algorithm, completed in seconds.
Captured: Local dense clusters effectively.
Limitation: Random initialization leads to inconsistent results.
Improvement: Consensus clustering can enhance stability.

4.5 5. Hierarchical Clustering

Agglomerative clustering builds nested communities based on similarity or linkage metrics.
Performance: Provided interpretable hierarchical relationships.
Captured: Multi-level structures and team dependencies.
Limitation: High computational cost for large graphs.
Improvement: Apply on low-dimensional spectral embeddings for scalability.

5 Comparison and Results

Algorithm	Type	Scalability	Detected Communities	Speed
Girvan–Newman	Edge betweenness	Low	Few large clusters	Slowest
Louvain	Modularity optimization	High	Many well-defined clusters	Fast
Spectral	Laplacian eigenvectors	Moderate	Clear subgroups	Moderate
Label Propagation	Heuristic local update	Very High	Variable	Fastest
Hierarchical	Linkage / distance-based	Low	Nested clusters	Slow

Table 1: Comparison of Community Detection Algorithms on Enron Dataset

Key Observations:

- Louvain performed best overall, balancing modularity and scalability.
- Girvan–Newman was accurate but extremely slow.
- Spectral and Hierarchical methods revealed interpretable structures.
- Label Propagation was fast but unstable due to random initialization.

6 Conclusion

This project demonstrated the effectiveness of multiple algorithms in revealing structural communities within the Enron email network.

Findings:

- Louvain provided the most meaningful and computationally efficient communities.
- Spectral and Hierarchical clustering offered deeper structural insights.
- Label Propagation is ideal for quick, large-scale exploration.
- Combining modularity-based and spectral methods can yield balanced results.