Alma Mater Studiorum - University of Bologna

School of Science

Department of Physics and Astronomy “Augusto Righi” (DIFA)

Master’s Degree in Physics

Comparative Network Analysis of Five Real-World Graphs: Zachary’s Karate Club, Les Misérables, College Football, Email-Eu-core, Facebook Social Circles

Supervisor:

Prof. Daniel Remondini

Submitted by:

Wafa Mamaani Mamaan

Academic Year 2024/2025

Contents

1. Introduction

Complex networks have become an essential framework for understanding a wide range of systems in physics, biology, social sciences, and informatics. From the interactions between proteins in a cell to the structure of social communities, complex networks offer a versatile way to capture and analyze the relationships between entities.

This project focuses on analyzing and comparing five real-world networks from the Stanford SNAP database and other well-known datasets:

1. **Zachary's Karate Club network** – a classic social network of 34 nodes.
2. **Les Misérables co-occurrence network** – a literary character network with 77 nodes.
3. **College Football network** – a network of American college football games with 115 nodes.
4. **Email-Eu-core network** – a large email communication network in a European research institution with over 1000 nodes.
5. **Facebook Social Circles network –** a large-scale online social network with 4,039 nodes and 88,234 edges, capturing friend connections within Facebook.

By choosing networks of different sizes and complexity, we can highlight how network size and density impact structural features and dynamic processes. This allows us to compare how centrality, clustering, and community structures evolve from small social groups to large-scale organizational communication.

The aim of this project is to apply the fundamental concepts and methods of complex networks to these datasets. We compute core topological measures, including centrality metrics (degree, closeness, betweenness, eigenvector) and clustering coefficients, to understand the structural properties of each network. Furthermore, we explore community detection using both the Girvan–Newman and Louvain algorithms, which reveal modular organization within the networks. Finally, we implement the Susceptible-Infected (SI) spreading model to simulate dynamic processes on these networks.

This comprehensive analysis not only deepens our understanding of these real-world networks but also demonstrates how the key concepts of complex networks can be practically applied to various domains. The findings highlight the similarities and differences in network structure and dynamics, reflecting how these abstract mathematical models align with real-world systems.

2. Theoretical Background

2.1 Networks and Graph Representations

A network, or graph, consists of a set of nodes (vertices) and the connections between them, called edges (or links). Depending on the type of system, graphs can be directed or undirected, weighted or unweighted, and may also be bipartite. In directed graphs, edges have a specific direction, as in the case of hyperlinks on the internet, while undirected graphs represent mutual connections, such as friendships in a social network. Weighted graphs assign numerical values to edges, reflecting the strength or frequency of interactions. Bipartite graphs, on the other hand, divide nodes into two disjoint sets where connections only occur between the sets.

2.2 Key Network Measures

To analyze the importance of nodes and the overall structure of a network, several fundamental measures are used like:

- **Degree:** The number of connections a node has. It’s a basic measure of importance within a network.
- **Closeness Centrality:** Measures how easily a node can reach others. Nodes with high closeness centrality can quickly spread information.
- **Betweenness Centrality:** Captures how often a node appears on the shortest paths between other nodes. High betweenness nodes often act as bridges between communities.
- **Eigenvector Centrality:** Considers not just how many connections a node has, but also how important its neighbors are.
- **Clustering Coefficient:** Measures how interconnected a node’s neighbors are, reflecting the tendency of nodes to form tightly knit groups.

2.3 Community Structure and Detection

Many networks naturally exhibit a community structure: groups of nodes that are more densely connected among themselves than with the rest of the network. Identifying these communities helps to uncover functional modules or social groups. Two widely used community detection algorithms are the **Girvan–Newman algorithm**, which progressively removes edges with high betweenness centrality to reveal modular structure, and the **Louvain algorithm**, which seeks to maximize a quantity known as modularity by merging and optimizing communities in a hierarchical manner.

2.4 Models of Network Growth

While our project did not directly simulate the Erdos-Renyi, Watts–Strogatz, or Barabási–Albert models, these models provide important theoretical benchmarks for understanding and interpreting our real networks. For example, networks like Zachary's Karate Club and Les Misérables show clustering and short average path lengths, properties reminiscent of small-world networks described by the **Watts–Strogatz model**. In contrast, larger real-world networks such as Email-Eu-core often contain hubs and exhibit broad degree distributions, similar to the **scale-free** structure predicted by the **Barabási–Albert model**. Recognizing these theoretical patterns allows us to better contextualize our empirical findings.

2.5 Dynamics on Networks: The SI Model

Dynamic processes such as information diffusion or disease spreading are strongly influenced by network structure. In this project, we applied the **Susceptible-Infected (SI) model** to each network to simulate how an infection (or information) propagates over time. In the SI model, nodes are either susceptible (S) or infected (I). At each step, infected nodes can infect their susceptible neighbors with a fixed probability β, leading to a gradual increase in the fraction of infected nodes. The simulation results provide insight into how the topological properties of each network impact the speed and extent of spreading processes.

3. Methods

This section describes the methods, algorithms, and tools we used to analyze the five real-world networks: Zachary's Karate Club, Les Misérables, College Football, Email-Eu-core, and Facebook Social Circles.

3.1 Tools and Environment

All analyses were conducted using **Python** in **Google Colab**, which provides an efficient environment for large-scale data processing and visualization. We relied on several libraries to implement the various network algorithms:

- **NetworkX** for core network analysis functions (centrality measures, clustering, community detection).
- **python-louvain** for Louvain community detection.
- **Matplotlib** for visualizing graphs and metric distributions.
- **NumPy** for numerical operations.

3.2 Data Preparation and Loading

For each network, we retrieved the dataset either from the Stanford SNAP database or built-in NetworkX datasets. The **Zachary's Karate Club**, **Les Misérables**, and **College Football** networks were loaded directly using NetworkX functions or GML files. The **Email-Eu-core** dataset was obtained from the SNAP repository and loaded as an edge list. The Facebook Social Circles network was also obtained from the SNAP dataset and loaded as an edge list, followed by basic cleaning and visualization to ensure consistency with the other datasets.

For the Email-Eu-core dataset, which includes many disconnected nodes, we extracted the **largest connected component (LCC)** to focus the analysis on the most relevant and structurally connected part of the network. This ensures that centrality and community detection metrics reflect the main functional core of the network.

3.3 Network Measures and Visualizations

For each network (or its LCC), we computed the following key measures to characterize its structure:

- **Degree centrality**: Identifying the most connected nodes.
- **Closeness centrality**: Capturing nodes' efficiency in spreading information.
- **Betweenness centrality**: Highlighting nodes acting as bridges.
- **Eigenvector centrality**: Measuring the influence of nodes based on their connections' importance.

- **Clustering coefficient**: Both local (for each node) and global (average across the network).

We visualized the distribution of these metrics using histograms to identify structural patterns in each network.

3.4 Community Detection

We employed two distinct algorithms to identify community structure:

- The **Girvan–Newman algorithm**, which iteratively removes edges with high betweenness centrality to expose hierarchical modularity.
- The **Louvain algorithm**, which optimizes modularity by merging nodes and communities in a hierarchical manner.

Visualizations of the communities were created by coloring nodes according to their community membership, providing clear insight into the modular organization of each network.

3.5 SI Model Simulation

To understand how the structure of each network influences dynamic processes, we implemented the **Susceptible-Infected (SI) model**. Starting with a randomly chosen infected node, at each step infected nodes could transmit the infection to susceptible neighbors with a probability β. We tracked and plotted the fraction of infected nodes over 20 time steps to observe the spread dynamics.

4. Results

In this section, we present the results of our analysis for each of the five real-world networks: Zachary's Karate Club, Les Misérables, College Football, Email-Eu-core, and Facebook Social Circles. For each network, we computed centrality and clustering metrics, identified communities using the Girvan–Newman and Louvain algorithms, and simulated the SI model to study dynamic spreading behavior. Relevant figures and tables provide a visual representation of these findings.

4.1 Zachary's Karate Club

The Karate Club network consists of 34 nodes and 78 edges. Centrality analysis revealed key individuals in the network, with node 0 (the instructor) and node 33 (the club administrator) exhibiting the highest degree, betweenness, and eigenvector centralities see **Table 1** for a summary of these metrics.

*Table 1: **Zachary's Karate Club Network Metrics***

Metric	Value / Note
Number of Nodes	34
Number of Edges	78
Global Clustering Coefficient	0.57
Highest Degree Centrality	Node 0: 0.48
Highest Closeness Centrality	Node 0: 0.57
Highest Betweenness Centrality	Node 0: 0.44
Highest Eigenvector Centrality	Node 0: 0.35
Number of Communities (Girvan–Newman)	2
Number of Communities (Louvain)	4

The degree centrality distribution (Figure 1) and other centrality histograms highlight the concentration of connections in these two influential nodes. The global average clustering coefficient was approximately 0.57, consistent with the presence of local clusters within the social network.

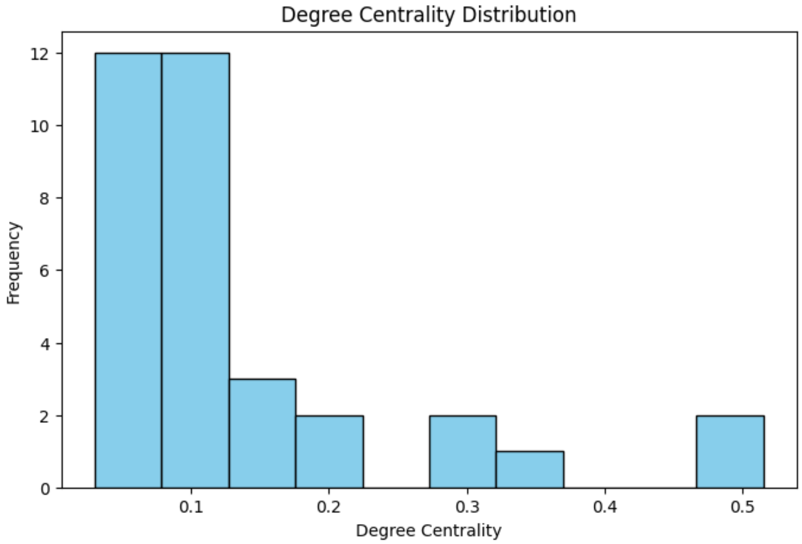

Figure 1: The degree centrality distribution of Zachary's Karate Club

Community detection with the Girvan–Newman algorithm uncovered two major communities (see Figure 2), aligning with the known split of the club. The Louvain algorithm detected four smaller communities (Figure 3), offering a finer breakdown of social subgroups.

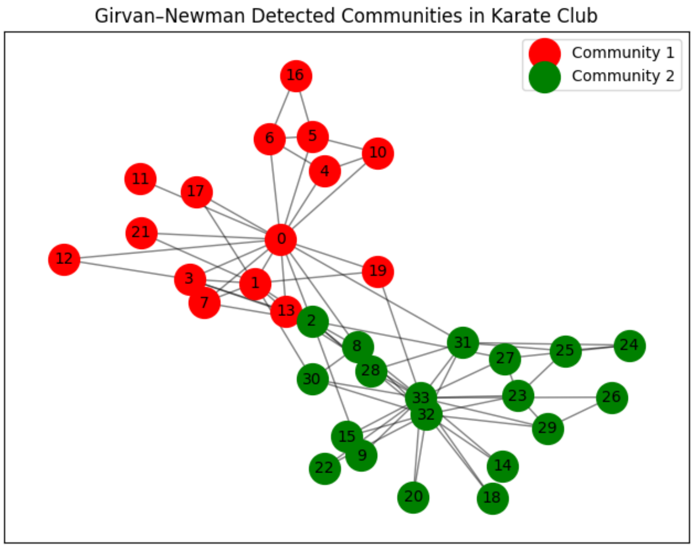

Figure 2: Community detection with the Girvan–Newman of Zachary's Karate Club

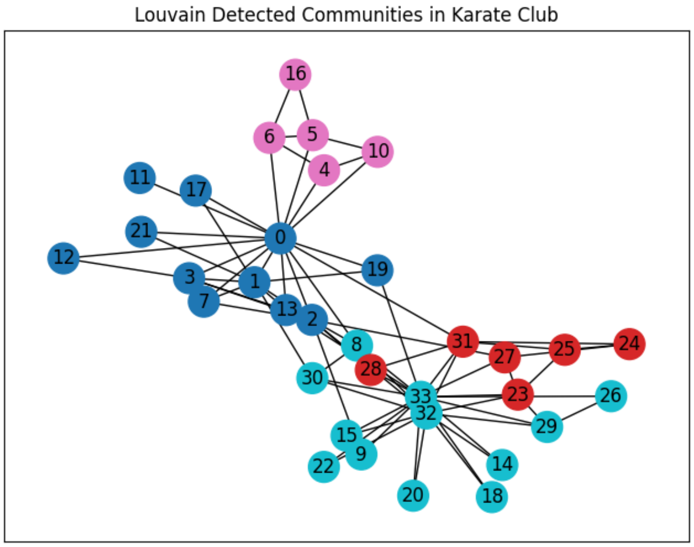

Figure 3: Community detection with the Louvain of Zachary's Karate Club

The SI model simulation (Figure 4) demonstrated a rapid initial infection spread, stabilizing as most nodes became infected. This behavior reflects the network's small-world structure and the prominent role of central individuals in facilitating rapid diffusion.

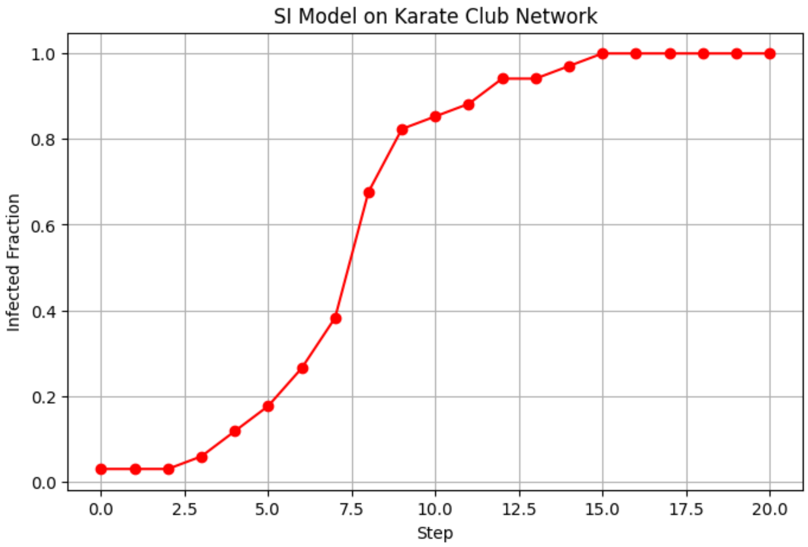

Figure 4: the SI model simulation on Karate Club Network

4.2 Les Misérables

The Les Misérables co-occurrence network includes 77 nodes and 254 edges, representing characters and their interactions. Centrality metrics identified key characters such as Valjean and Javert, who exhibit high betweenness and eigenvector centralities (Table 2).

*Table 2: **Les Misérables Network Metrics***

Metric	Value / Note
Number of Nodes	77
Number of Edges	254
Global Clustering Coefficient	0.57
Highest Degree Centrality	Valjean: 0.22
Highest Closeness Centrality	Valjean: 0.46
Highest Betweenness Centrality	Valjean: 0.15
Highest Eigenvector Centrality	Valjean: 0.10
Number of Communities (Girvan–Newman)	2
Number of Communities (Louvain)	6

The network's structure shows a small-world-like property, with high clustering and the presence of key bridging nodes. Girvan–Newman detected two main communities (Figure 5), while Louvain identified six smaller communities (Figure 6), capturing the complex social dynamics of the story.

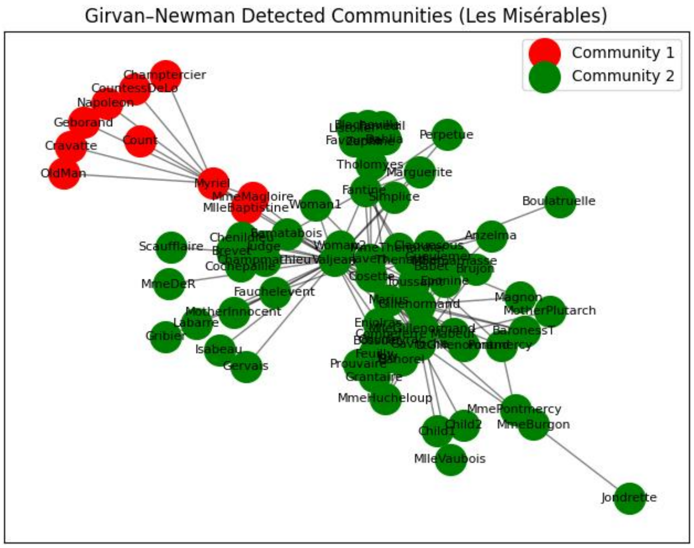

Figure 5: Community detection with the Girvan–Newman of Les Misérables

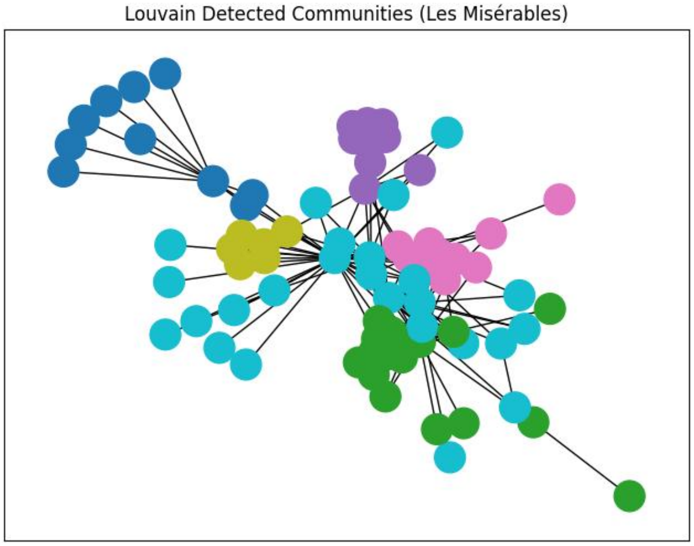

Figure 6: Community detection with the Louvian of Les Misérables

The SI model infection curve (Figure 7) showed rapid propagation of infection, consistent with the network's high clustering and the presence of hubs that accelerate spread.

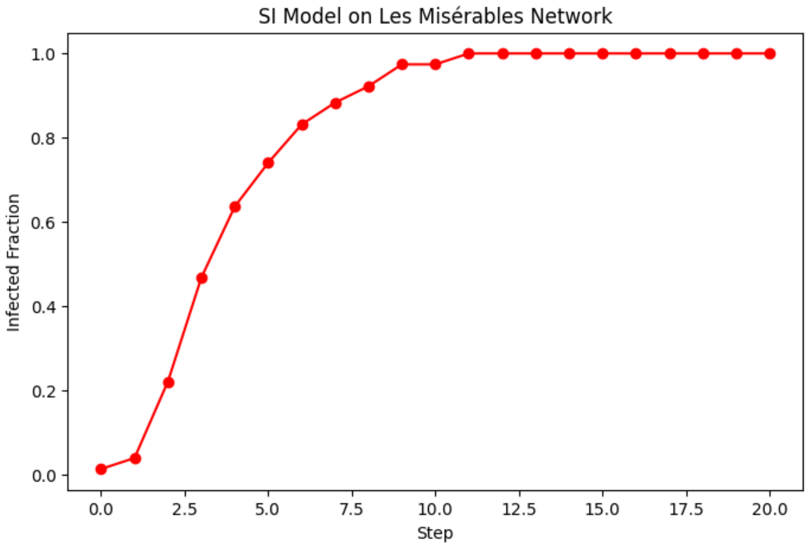

Figure 7: The SI Model simulation on Les Misérables

4.3 College Football

The College Football network comprises 115 nodes and 613 edges. Centrality analysis revealed several high-degree nodes, representing major college football teams with central roles in the schedule (Table 3).

Table 3: ***College Football Network Metrics***

Metric	Value / Note
Number of Nodes	115
Number of Edges	613
Global Clustering Coefficient	0.40
Highest Degree Centrality	Sample team: 0.10
Highest Closeness Centrality	Sample team: 0.42
Highest Betweenness Centrality	Sample team: 0.03
Highest Eigenvector Centrality	Sample team: 0.11
Number of Communities (Girvan–Newman)	2
Number of Communities (Louvain)	10

The global clustering coefficient was around 0.40, highlighting the mixture of tightly knit clusters (conferences) and inter-conference games. The Girvan–Newman algorithm revealed two large communities (Figure 8), while Louvain detected 10 smaller communities (Figure 9), capturing the rich modular structure of the competition.

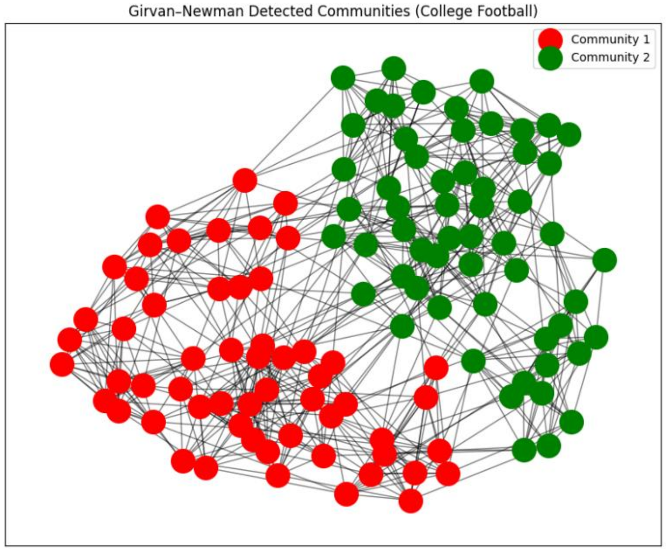

Figure 8: Community detection with the Girvan–Newman of College Football

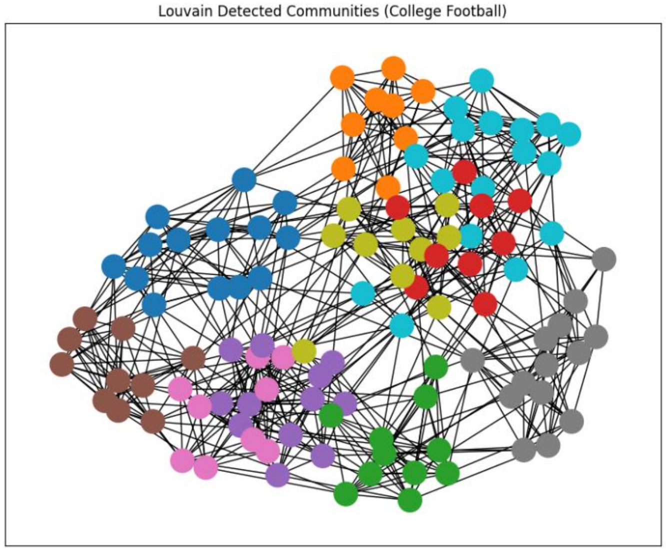

Figure 9: Community detection with the Louvian of College Football

The SI model simulation (Figure 10) showed a gradual but complete infection spread, driven by well-connected teams acting as hubs.

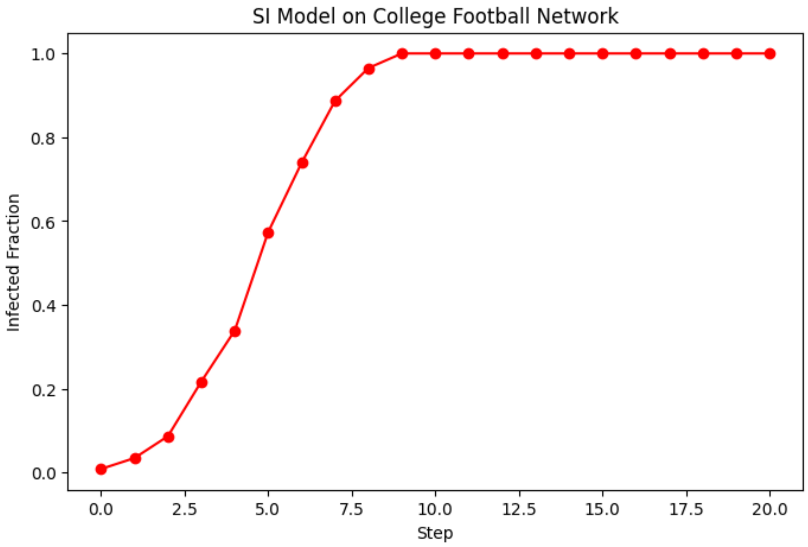

Figure 10: The SI Model simulation on College Football

4.4 E-mail-Eu-core

The Email-Eu-core network originally included over 1000 nodes and 16,706 edges. To focus the analysis on the core structure, we extracted the largest connected component (986 nodes, 16,687 edges). Centrality measures identified key accounts with high degree and eigenvector centralities (Table 4).

*Table 4: **Email-Eu-core Network (LCC) Metrics***

Metric	**Value / Note**
Number of Nodes	986
Number of Edges	16,687
Global Clustering Coefficient	0.41
Highest Degree Centrality	Node 0: 0.10
Highest Closeness Centrality	Node 0: 0.47
Highest Betweenness Centrality	Node 0: 0.01
Highest Eigenvector Centrality	Node 0: 0.06
Number of Communities (Girvan–Newman)	2
Number of Communities (Louvain)	7

The centrality and clustering metrics revealed a wide range of connectivity, with a handful of highly central nodes serving as key hubs. Community detection identified a large central community and smaller peripheral groups. Girvan–Newman detected two major communities (Figure 11), while Louvain revealed seven more nuanced communities (Figure 12).

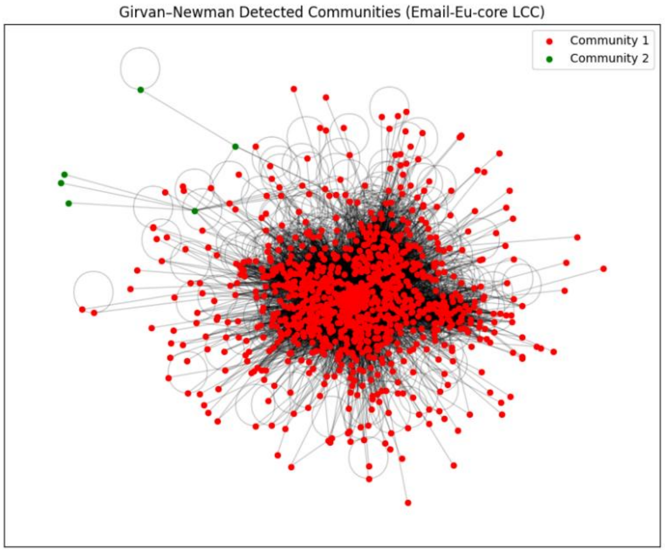

Figure 11: Community detection with the Girvan–Newman of E-mail-Eu-core

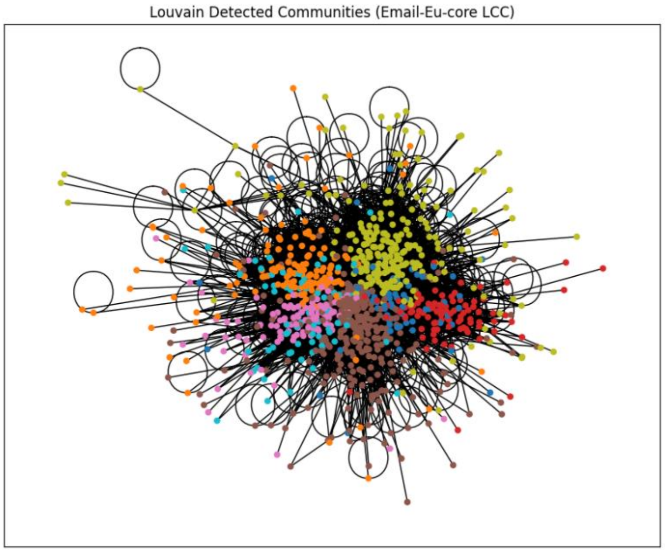

Figure 12: Community detection with the Louvian of E-mail-Eu-core

The SI model infection curve (Figure 13) demonstrated a very rapid spread due to the network's dense connectivity and core hubs.

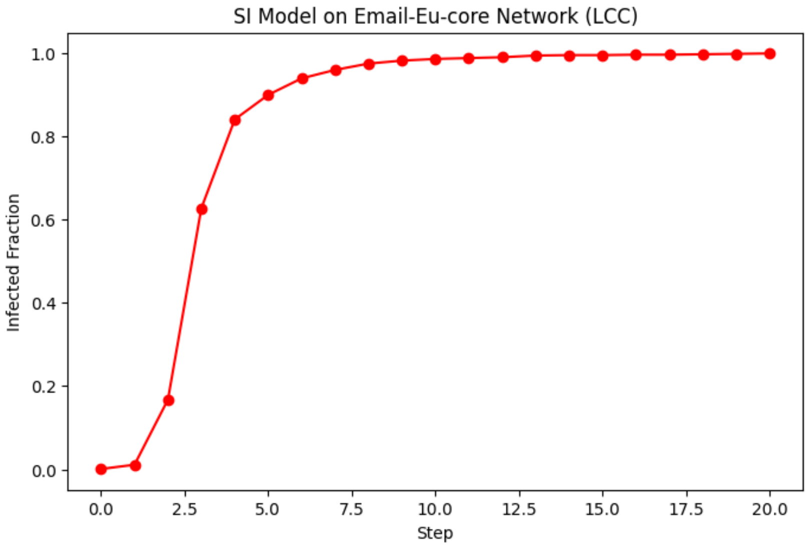

Figure 13: The SI Model simulation on E-mail-Eu-core

4.5 Facebook Social Circles

The Facebook Social Circles network includes 4,039 nodes and 88,234 edges, representing social connections within Facebook. Centrality measures reveal that node 0 has the highest degree, closeness, and betweenness centralities, while the global clustering coefficient is notably high at 0.61 (Table 5).

*Table 5: **Facebook Social Circles Network Metrics***

Metric	Value / Note
Number of Nodes	4,039
Number of Edges	88,234
Global Clustering Coefficient	0.61
Highest Degree Centrality	Node 0: 0.086
Highest Closeness Centrality	Node 0: 0.353
Highest Betweenness Centrality	Node 0: 0.146
Highest Eigenvector Centrality	Node 0: 0.000
Number of Communities (Louvain)	16

The centrality and clustering metrics reveal a few highly connected hub nodes and many peripheral nodes, typical of large-scale social networks. Louvain community detection identified 16 communities (Figure 14), capturing the rich modular structure of Facebook social circles.

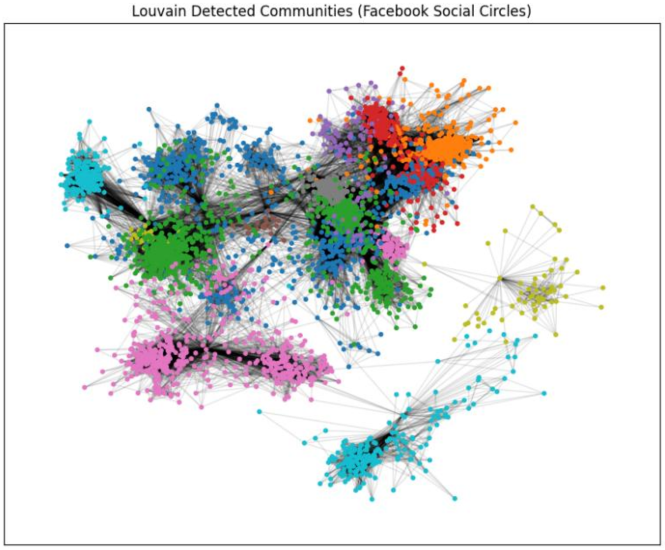

Figure 14: Community detection with the Louvian of Facebook Social Circles

Due to the large size of the network, the Girvan–Newman algorithm was not applied, as it is computationally intensive for networks of this scale.
The SI model simulation (Figure 15) demonstrates a steady initial infection spread that accelerates once key hub nodes become involved, illustrating the importance of central individuals in social networks.

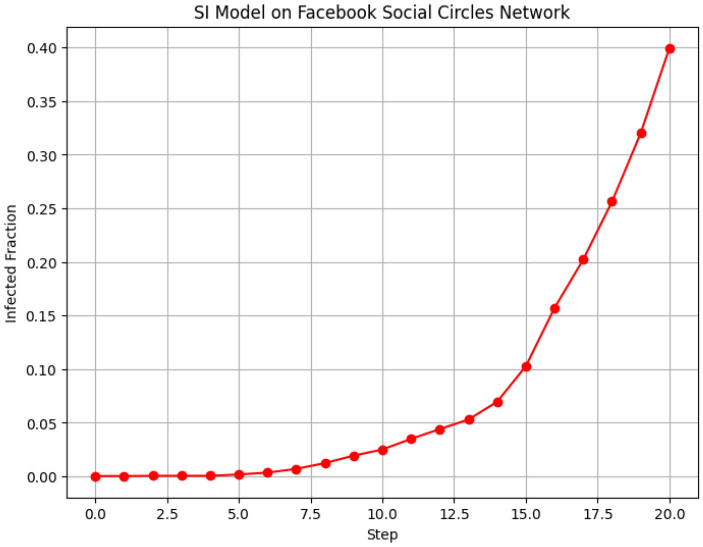

Figure 15: The SI Model simulation on Facebook Social Circles

5. Discussion

The analyses performed on these five real-world networks—Zachary's Karate Club, Les Misérables, College Football, Email-Eu-core, and Facebook Social Circles—reveal both common structural patterns and unique features related to their different contexts and sizes.

One clear finding across all networks is the presence of highly central nodes. In smaller networks like Karate Club and Les Misérables, these central nodes often represent key individuals (e.g., the instructor in the Karate Club or Valjean in Les Misérables). In larger networks such as Email-Eu-core and Facebook Social Circles, these hubs are key individuals or accounts that play a crucial role in maintaining communication across the entire system. Such hubs have high degree and eigenvector centralities, reflecting their prominent positions and importance in the overall structure.

Clustering coefficients also reveal important patterns. The Karate Club and Les Misérables networks have high global clustering coefficients (around 0.57), typical of social networks where nodes tend to form tightly knit groups. In contrast, the College Football and Email-Eu-core networks show moderately lower clustering (~0.4), indicating the presence of local groupings (conferences and collaborative teams) alongside broader connectivity. The Facebook Social Circles network stands out with the highest clustering coefficient (0.61), reflecting the densely interconnected local communities that are characteristic of online social platforms.

Community detection results illustrate how different algorithms can reveal varying modular structures. The Girvan–Newman algorithm consistently identifies two large communities, reflecting the most significant divisions in smaller or moderately sized networks. The Louvain algorithm, with its modularity optimization, often finds a richer, finer community structure (four in the Karate Club, six in Les Misérables, 10 in College Football, seven in Email-Eu-core, and 16 in Facebook Social Circles), uncovering the more nuanced substructures within these systems. Due to computational constraints, the Girvan–Newman algorithm was not applied to the Facebook Social Circles network, highlighting the scalability challenges of certain algorithms for large-scale graphs.

Dynamic processes on these networks, as modeled by the SI model, also highlight the impact of network structure on spreading processes. In smaller networks like Karate Club and Les Misérables, the infection spreads rapidly and stabilizes quickly due to the small-world property and the role of central nodes as bridges. In larger networks, especially

Email-Eu-core and Facebook Social Circles, the SI model demonstrates an even faster spread once hub nodes begin participating, driven by the high density of connections and the presence of influential hubs that rapidly connect distant parts of the network.

By comparing these results, we see how the size and density of networks shape both their structural properties and their dynamic behavior. Smaller networks exhibit clear, well-defined communities and slower but still efficient spreading. Larger networks display a richer, more hierarchical modular structure and more explosive dynamic processes, driven by their complex connectivity and the presence of influential hubs.

Overall, this comparative analysis underscores the importance of key network metrics—centrality, clustering, and community structure—in understanding how real-world networks function and evolve. It also demonstrates the power of combining static structural analysis with dynamic models like SI to provide a comprehensive picture of network behavior.

6. Conclusion

In this project, we conducted a comprehensive analysis of five real-world networks: Zachary's Karate Club, Les Misérables, College Football, Email-Eu-core, and Facebook Social Circles. By computing fundamental structural metrics—degree, closeness, betweenness, and eigenvector centralities, as well as clustering coefficients—we identified key nodes and uncovered the underlying connectivity patterns of each network. Community detection algorithms (Girvan–Newman and Louvain) revealed the modular organization within these systems, offering insights into their hierarchical and functional divisions.

Due to the size and computational demands of the Facebook Social Circles network, only the Louvain algorithm was applied for community detection, showcasing how scalability can influence methodological choices in large-scale network analysis.

Simulating the Susceptible-Infected (SI) model provided a dynamic perspective on how the structure of each network influences spreading processes. In all networks, central hubs and clustering played a critical role in accelerating or constraining the spread. The comparative results showed that smaller networks with clear community structures exhibit rapid yet locally contained spreading, while larger networks with more complex, hierarchical modularity—like Email-Eu-core and Facebook Social Circles—enable faster and more global diffusion once hubs become involved.

By examining networks of different sizes and domains—ranging from small social networks to large organizational and online social networks—we demonstrated how core principles of complex network theory apply across diverse contexts. The analysis confirmed the importance of centrality measures and modularity in shaping the behavior of real-world networks and highlighted the dynamic interplay between structure and processes.

This project underscores the value of integrating theoretical foundations with empirical data to better understand the complexity of networks in the real world. The approaches and tools used here can be extended to other datasets and questions in network science, offering a robust framework for future exploration and practical applications.

References

M. E. J. Newman, **Networks: An Introduction**, Oxford University Press, 2010.

G. Caldarelli and A. Vespignani (eds.), **Large Scale Structure and Dynamics of Complex Networks – Vol. 2**, World Press.

A. Barrat, M. Barthélemy, and A. Vespignani, **Dynamical Processes on Complex Networks**, Cambridge University Press, 2008.

Stanford Large Network Dataset Collection (SNAP), https://snap.stanford.edu/data/

NetworkX: Network Analysis in Python, https://networkx.org/

T. Aynaud, **Louvain Community Detection for Python** (python-louvain), https://github.com/taynaud/python-louvain