Toronto Metropolitan University Data Science & Analytics Graduate Studies

Comprehensive Literature Review for Network Resilience Testing

1. Introduction

Network resilience refers to the ability of a network to maintain its essential functions despite failures or attacks. Understanding and testing network resilience is crucial for ensuring the robustness and reliability of systems ranging from telecommunications and social networks to distributed control systems (DCS).

2. Definitions and Frameworks of Network Resilience

Network resilience encompasses the network's ability to absorb disturbances, adapt to changes, and recover from disruptions. Metrics such as robustness, redundancy, and adaptability are used to quantify resilience.

- **Robustness**: The network's inherent strength against failures.
- **Redundancy**: Availability of alternative paths or components.
- **Adaptability**: Ability to adjust to new conditions.

3. Graph-Theoretic Approaches to Network Resilience

Graph theory provides a robust framework for analyzing network resilience through various metrics and models.

- **Algebraic Connectivity**: Measures the network's overall connectivity and its ability to remain connected under node/edge failures.
- **Betweenness Centrality**: Identifies critical nodes/edges that, if removed, would significantly impact network connectivity.
- **Path Diversity**: Number of disjoint paths between nodes, indicating the network's ability to reroute traffic.

4.Literature Review

4.1 Introduction

The objective of this project is to evaluate the resilience of complex networks, particularly telecommunications backbone networks, against failures and attacks. Resilience is crucial as it ensures the network's ability to maintain functionality despite disruptions. This review examines various metrics and models used to measure and predict network resilience, drawing from 15 key research papers.

4.2 Key Research Papers and Findings

The evaluation of network resilience is a multifaceted challenge that has garnered significant attention in the research community. In their seminal work, Albert, Jeong, and Barabási (2000)[1] explored the resilience of scale-free networks, demonstrating robustness to random failures but high vulnerability to targeted attacks on high-degree nodes. Their findings laid the groundwork for understanding the criticality of node roles within network structures. Complementing this, Palmer et al. (2001)[2] examined the internet's router and AS-level topologies, highlighting the importance of path diversity for resilience. They found that the internet remains connected despite frequent router problems due to redundant paths.

Wang et al. (2015)[3] further extended this understanding by presenting resilience metrics for supply networks, using synthetic and real-world data, showing that networks with high redundancy exhibit greater resilience to disruptions. This was echoed by Alenazi and Sterbenz (2015)[4], who evaluated various graph metrics for predicting network resilience under targeted attacks, finding that metrics like betweenness centrality are strong indicators of robustness.

Cohen et al. (2000)[5] and Holme et al. (2002)[6] both contributed by examining the internet's robustness to random node failures and targeted attacks, emphasizing the need to protect critical nodes to maintain resilience. Newman (2003)[7] provided a comprehensive review of network structures and their robustness, discussing how different topologies affect resilience in varied ways.

Paul et al. (2004)[8] highlighted the trade-offs between robustness to random failures and targeted attacks, suggesting that network design should consider both types of threats. Motter and Lai (2002)[9] focused on cascading failures in networks, demonstrating that network structure plays a crucial role in mitigating cascading risks.

Buldyrev et al. (2010)[10] and Gao et al. (2012)[11] explored the resilience of interdependent networks, showing how failures in one network can cause cascading failures in another, highlighting the importance of coupling strength. Pagani and Aiello (2013)[12] reviewed the power grid's network structure and resilience, discussing methods to enhance robustness, which was further supported by Huang et al. (2019)[13], who proposed new topological metrics for analyzing network resilience.

Callaway et al. (2000)[14] studied network connectivity under random node removal, identifying critical thresholds for maintaining connectivity. Rohlin (2016)[15] attempted to predict the popularity of Reddit posts using various metrics, illustrating the significance of user interactions and post characteristics in predicting popularity.

Together, these studies provide a comprehensive understanding of network resilience, emphasizing the need to protect critical nodes and design networks with redundancy to withstand both random failures and targeted attacks. By leveraging insights from these studies, the project

aims to enhance the resilience of complex networks, ensuring robust performance despite disruptions.

4.3 Conclusion

The reviewed papers collectively offer a comprehensive understanding of network resilience, emphasizing the need to protect critical nodes and design networks with redundancy to withstand both random failures and targeted attacks. By leveraging insights from these studies, the project aims to enhance the resilience of complex networks, ensuring robust performance despite disruptions.

4.4 references

1. [Error and Attack Tolerance of Complex Networks](https://arxiv.org/abs/cond-mat/0008064)

2. [The Connectivity and Fault-Tolerance of the Internet Topology](https://dl.acm.org/doi/10.1145/505200.505202)

3. [Topological Resilience Analysis of Supply Networks under Random Disruptions and Targeted Attacks](https://ieeexplore.ieee.org/document/7474108)

4. [Comprehensive Comparison and Accuracy of Graph Metrics in Predicting Network Resilience](https://ieeexplore.ieee.org/document/7069364)

5. [Resilience of the Internet to Random Breakdowns](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.66.036113)

6. [Attack Vulnerability of Complex Networks](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.65.056109)

7. [The Structure and Function of Complex Networks](https://www.santafe.edu/research/results/working-papers/the-structure-and-function-of-complex-networks)

8. [Resilience of Complex Networks to Random Breakdowns and Targeted Attacks](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.70.056107)

9. [Cascade-based Attacks on Complex Networks](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.66.065102)

10. [Catastrophic Cascade of Failures in Interdependent Networks](https://www.nature.com/articles/nature08932)

11. [Robustness of a Network of Networks](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.85.066134)

12. [The Power Grid as a Complex Network: A Survey](https://journals.aps.org/pre/abstract/10.1103/PhysRevE.87.062809)

13. [Analysis and Improvement of Network Resilience Using Topological Metrics](https://www.sciencedirect.com/science/article/pii/S0140366418307677)

14. [Network Robustness and Fragility: Percolation on Random Graphs](https://link.springer.com/article/10.1007/s100510070173)

15. [Popularity Prediction of Reddit Texts](https://arxiv.org/abs/1609.08347)

5. Exploratory Data Analysis

5.1 Exploratory Data Analysis (EDA) Report for the Email-Eu-core Dataset

5.1.1 Introduction

This report provides a comprehensive exploratory data analysis (EDA) of the email-Eu-core dataset. The dataset includes email communications between individuals within a European research institution, along with department labels for each individual. Our analysis aims to understand the structure and properties of the email network

5.1.2 Data Overview

The dataset consists of two files:

1. **email-Eu-core.txt**: Contains edges representing email communications between nodes.
2. **email-Eu-core-department-labels.txt**: Contains labels representing the department each node belongs to.

5.1.3 Basic Information

Edges Dataset:

- Number of edges: 25,571
- Columns: source, target

Labels Dataset:

- Number of nodes: 1,005
- Columns: node, department

Both datasets contain no missing values.

5.1.4 Statistical Summary

Degree Distribution

The degree distribution shows the number of connections (emails) each node has. This distribution is often skewed, indicating that most nodes have a few connections, while a few nodes have many connections.

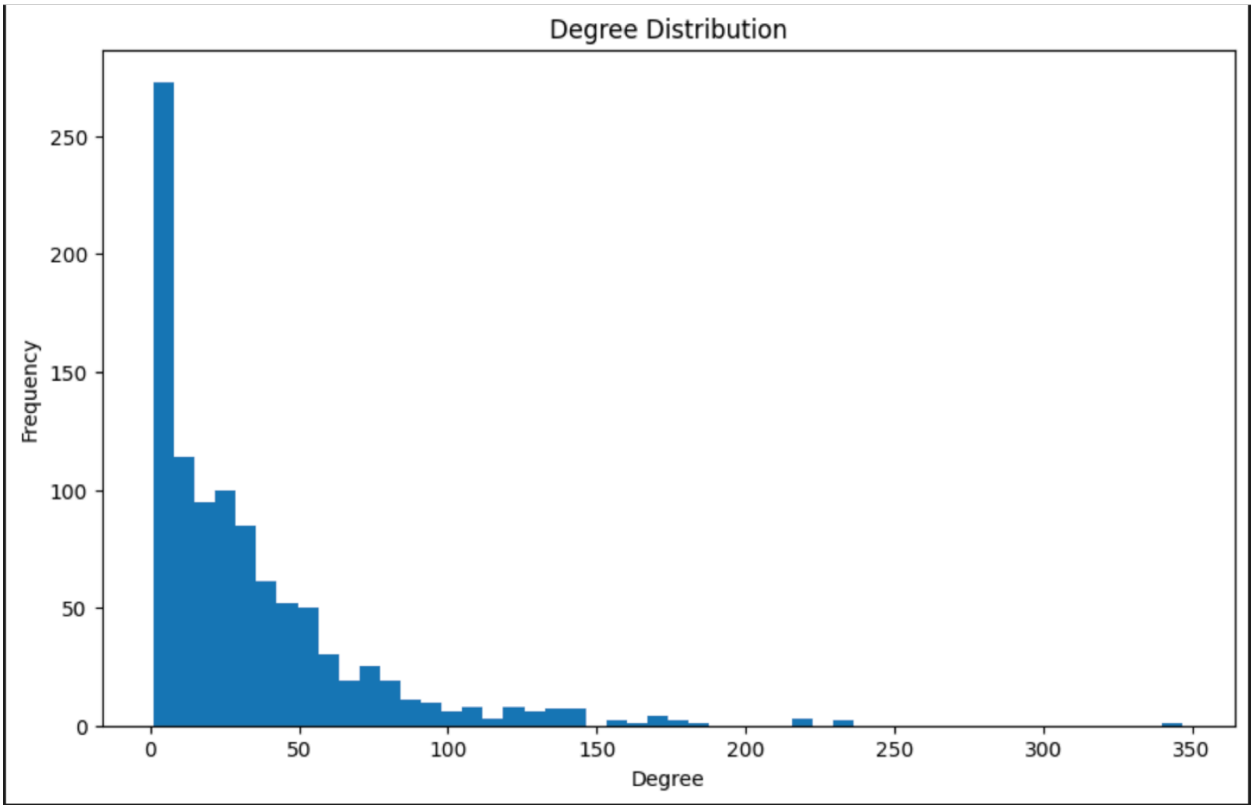


Fig.1.1

Department Labels Distribution

The department labels distribution visualizes the number of individuals in each department. This helps identify the departments with the highest and lowest representation in the dataset.

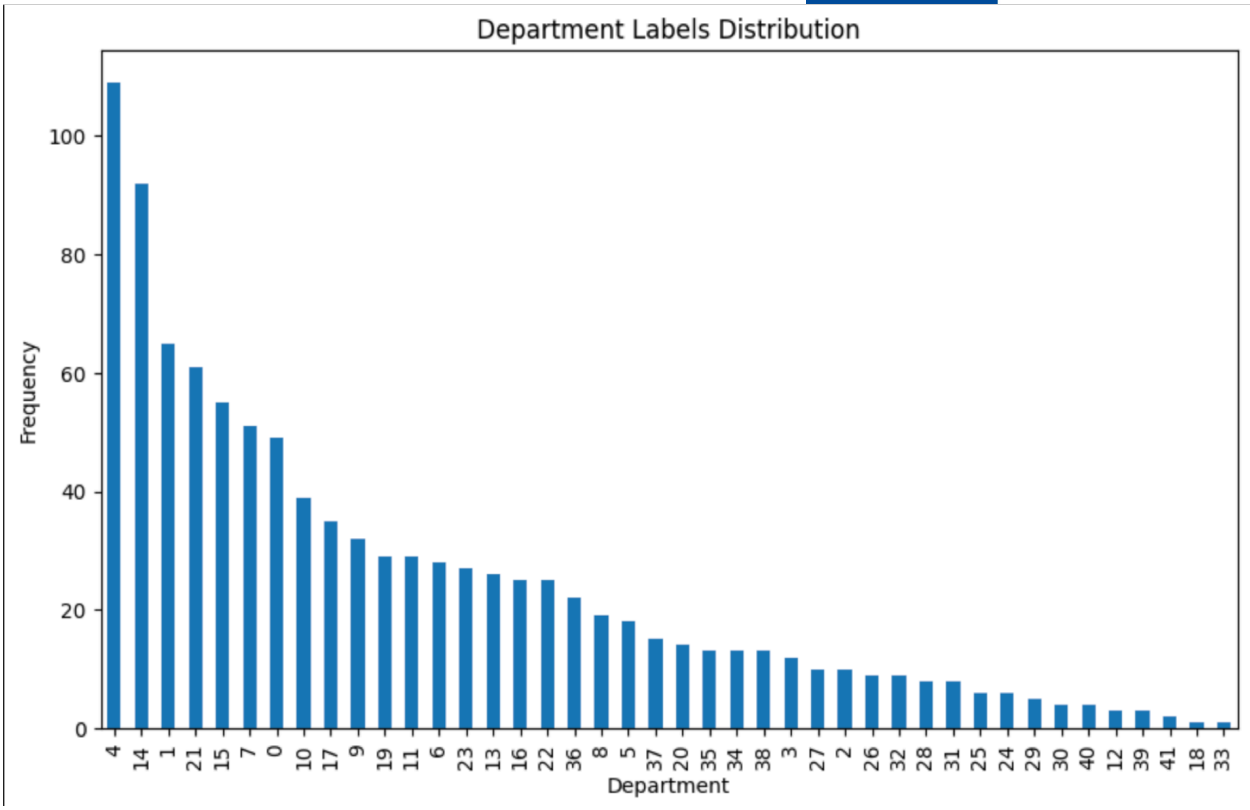


Fig.1.2

Network Properties

These properties indicate a moderately dense network with a single connected component, suggesting strong interconnectivity among the nodes.

	Property	Value
0	Number of Nodes	1005.000000
1	Number of Edges	16706.000000
2	Density	0.033113
3	Number of Connected Components	20.000000

Fig1.3

Degree Centrality

Degree centrality measures the importance of nodes based on the number of connections they have. Higher degree centrality indicates nodes with many connections, often playing a key role in communication.

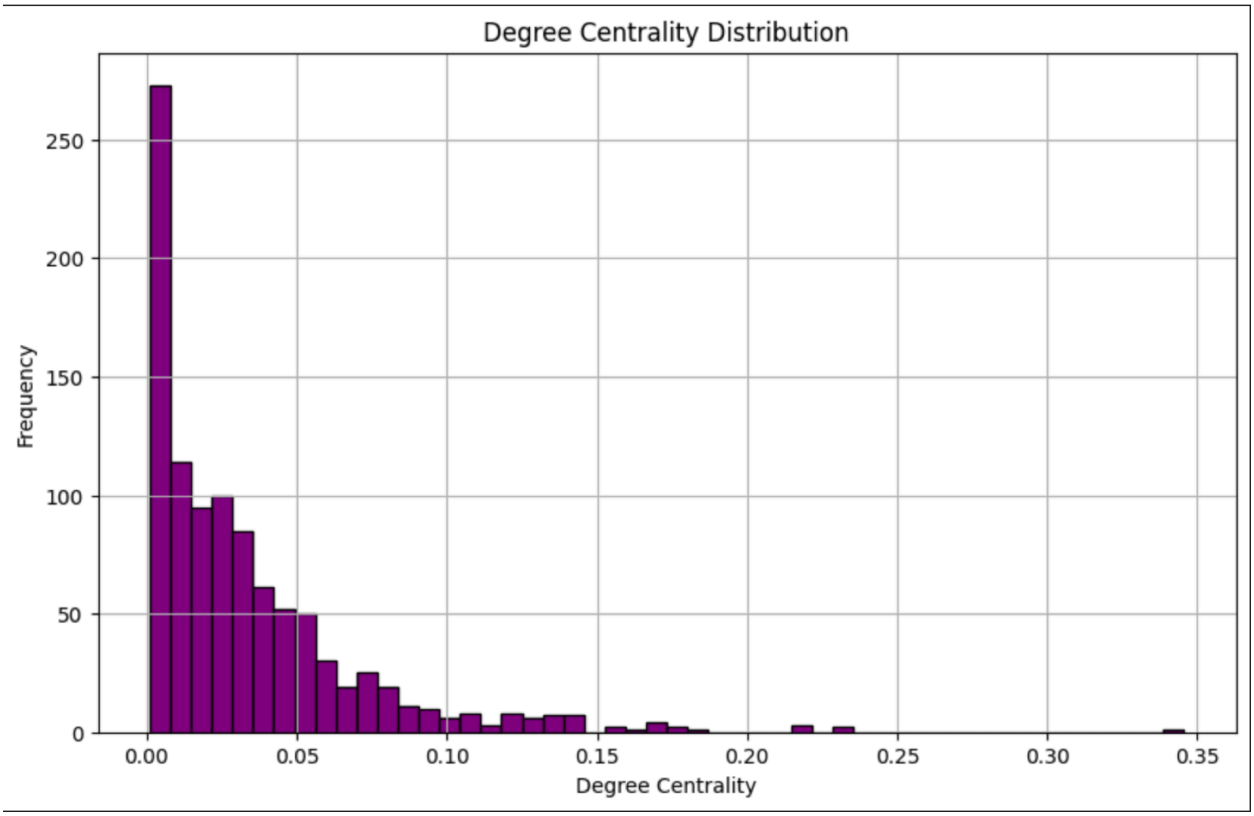


Fig 1.4

Clustering Coefficient

The clustering coefficient measures the degree to which nodes tend to cluster together. A higher clustering coefficient indicates a greater tendency for nodes to form tightly knit groups.(Fig1.5)

Betweenness Centrality

Betweenness centrality identifies nodes that act as bridges within the network. Nodes with high betweenness centrality are critical for maintaining communication paths and network integrity.(Fig1.6)

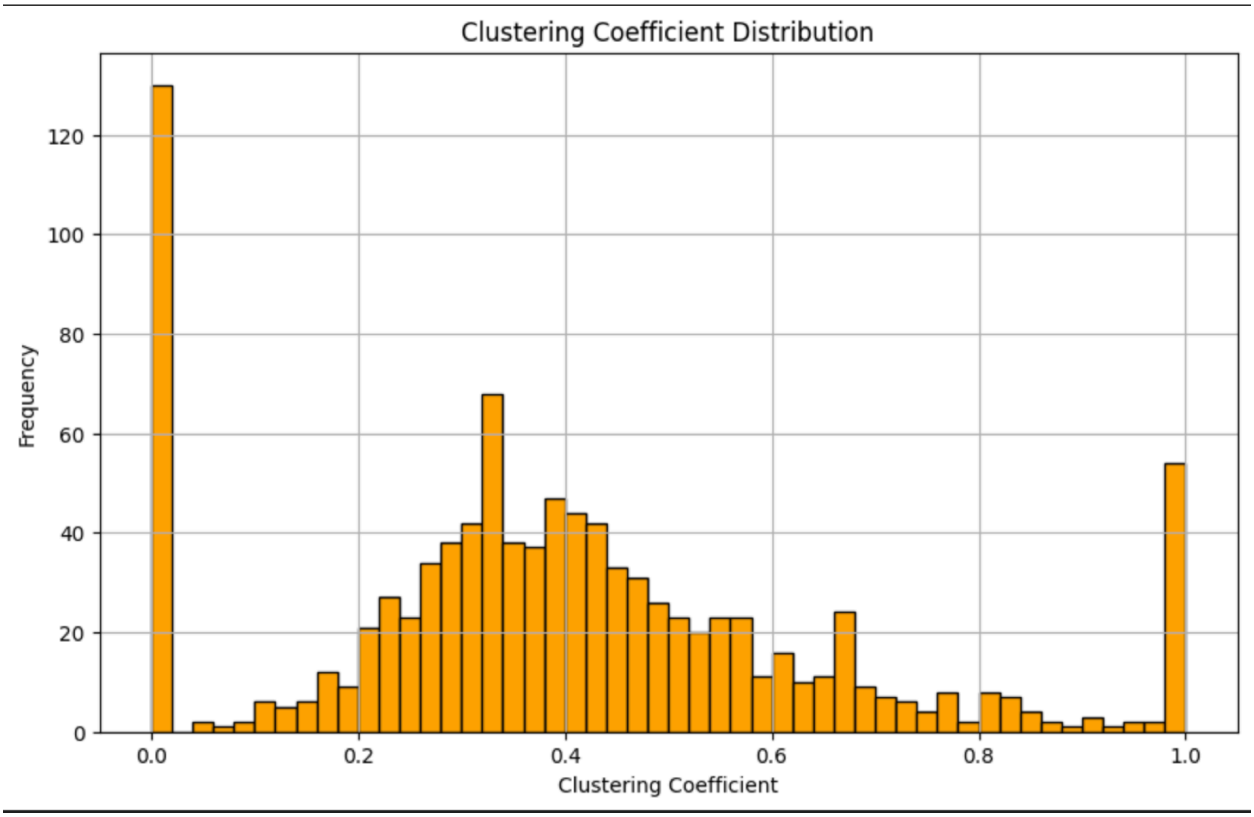

Clustering Coefficient Distribution
Frequency
Clustering Coefficient
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0

Fig1.5

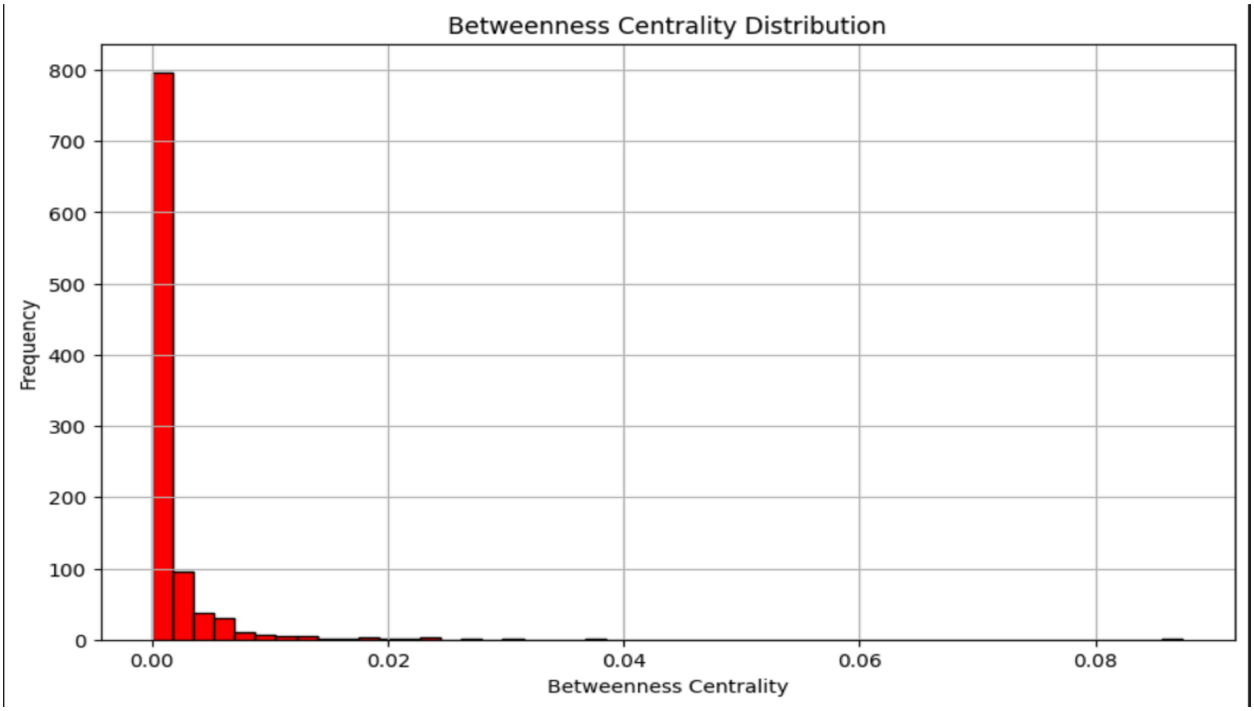

Betweenness Centrality Distribution
Frequency
Betweenness Centrality
0
100
200
300
400
500
600
700
800
0.00
0.02
0.04
0.06
0.08

Fig1.6

5.2 Exploratory Data Analysis (EDA) Report on LastFM Dataset

5.2.1 Introduction

The LastFM dataset consists of two primary components: edges_df and target_df.
The edges_df represents the connections between users, while target_df provides additional information about each user, including their country of origin. This EDA aims to explore the structure and properties of the LastFM network, understand the distribution of node degrees, centralities, and clustering coefficients, and visualize the country distribution of users.

5.2.2 Data Loading and Initial Inspection

We begin by loading the data and inspecting the first few rows of each dataset to understand its structure. The `edges_df`contains two columns, `node_1` and `node_2`, representing connections between users. The `target_df` includes the `user_id`and `country` columns.

5.2.3 Basic Information and Statistical Summary

We use the `info()` method to get a concise summary of the datasets, including the number of entries, column names, and data types. This is followed by a statistical summary using `describe()` to understand the distribution of numerical data within the datasets.

	Property	Value
0	Number of Nodes	7624.000000
1	Number of Edges	27806.000000
2	Density	0.000957
3	Number of Connected Components	1.000000

Fig2.1

5.2.4 Degree Distribution

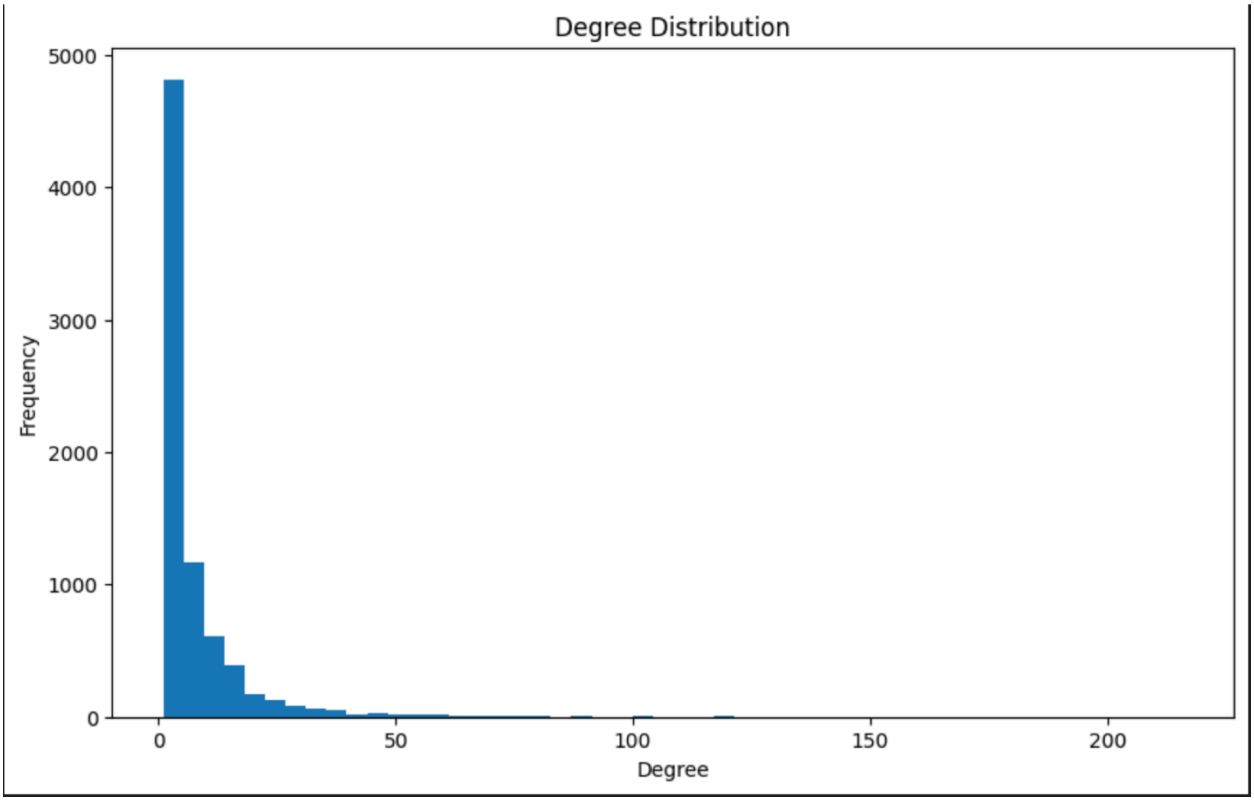


Fig 2.2

5.2.5 Labels Distribution

We analyze the distribution of users across different countries. The `country` column in the `target_df` is used to plot a bar chart, showing the number of users from each country. This helps in understanding the geographical diversity of the user base. (fig2.3)

5.2.6 Degree Centrality

Degree centrality measures the importance of a node based on its number of connections. We calculate and plot the distribution of degree centrality for all nodes in the network. This helps identify highly connected nodes that play a critical role in the network.(fig2.4)

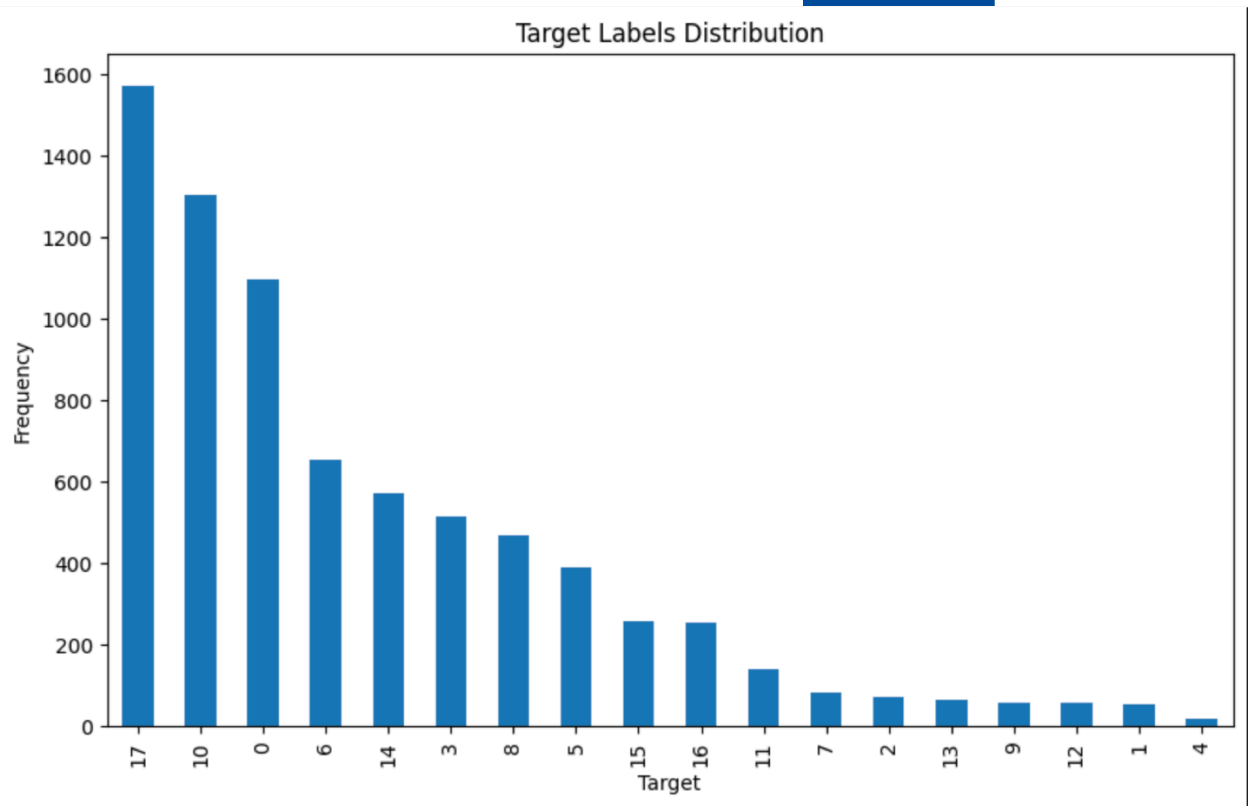

Target Labels Distribution
Frequency
Target
0
200
400
600
800
1000
1200
1400
1600
17
10
0
6
14
3
8
5
15
16
11
7
2
13
9
12
1
4

Fig 2.3

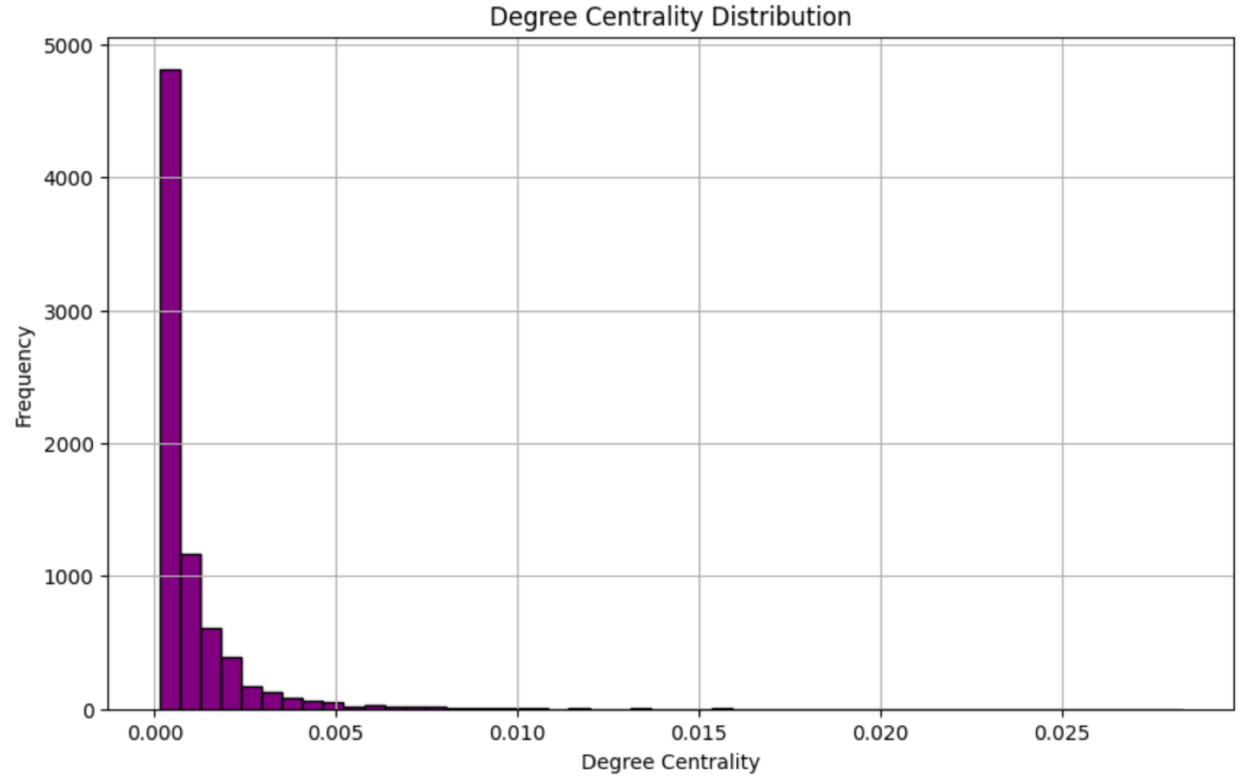

Degree Centrality Distribution
Frequency
Degree Centrality
0
1000
2000
3000
4000
5000
0.000
0.005
0.010
0.015
0.020
0.025

Fig 2.4

5.2.7 Network Properties

Key network properties such as the number of nodes, number of edges, network density, and the number of connected components are calculated. These metrics provide insights into the overall structure and connectivity of the network.

5.2.8 Clustering Coefficient

The clustering coefficient measures the tendency of nodes to form tightly knit groups. We calculate the clustering coefficient for each node and plot its distribution. High clustering coefficients indicate a high level of local cohesion in the network.

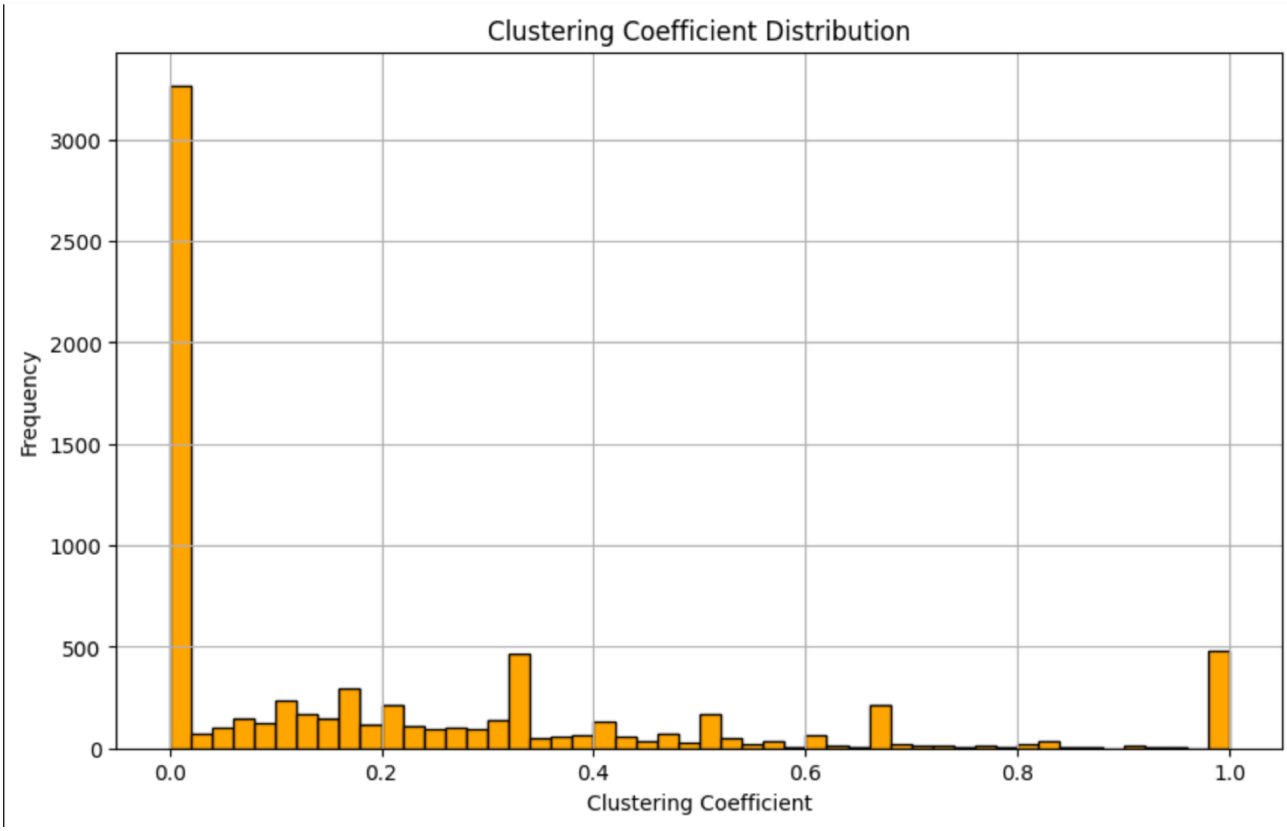
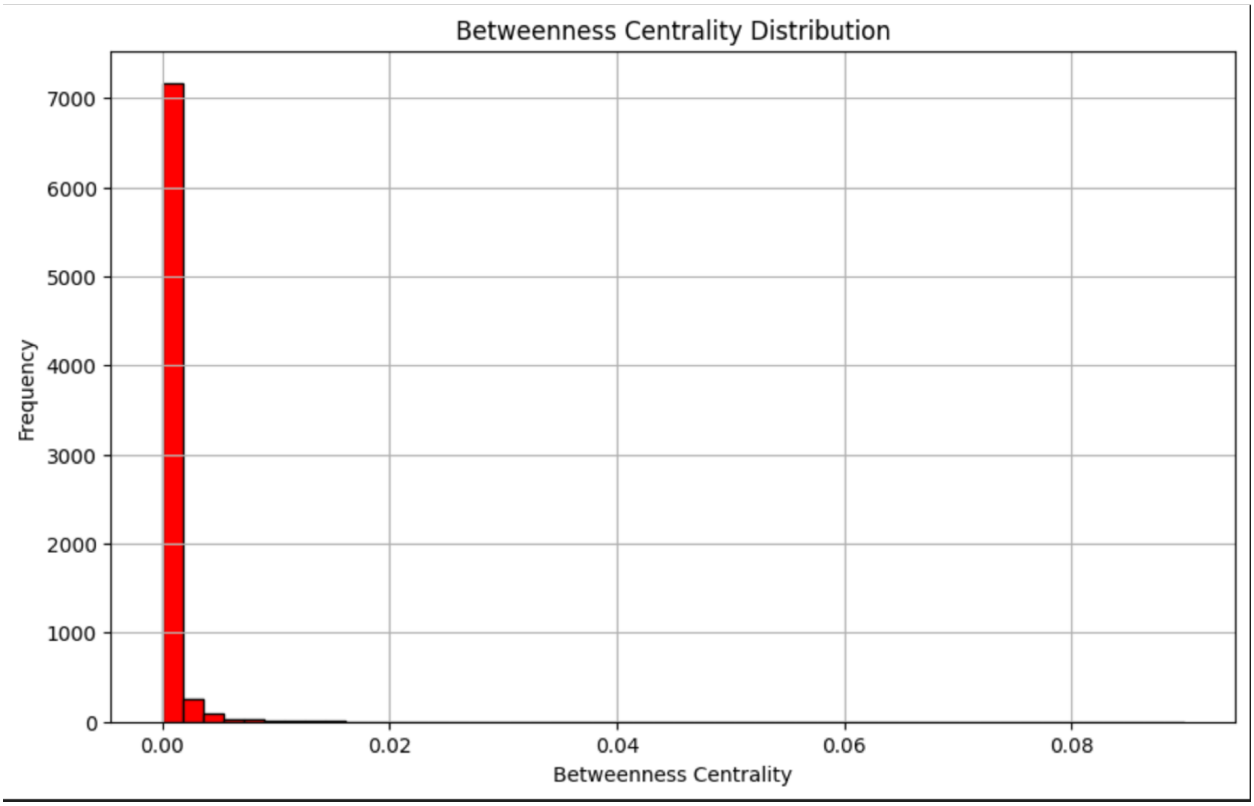


Fig 2.5

5.2.9 Betweenness Centrality

Betweenness centrality measures the frequency at which a node appears on the shortest paths between other nodes. We calculate and plot the betweenness centrality for all nodes to identify nodes that act as critical bridges within the network.



6 Conclusion for Email-Eu-core and LastFM Datasets

6.1 Email-Eu-core Dataset

The Email-Eu-core dataset represents email communications within a large European research institution. This dataset is highly structured and exhibits typical characteristics of social networks, including a few highly connected nodes (hubs) and many nodes with fewer connections. The network's density and clustering coefficients indicate moderate local clustering, while the presence of a giant component suggests a strong core structure. These properties make it suitable for studying network resilience, particularly in understanding how network connectivity is maintained or disrupted under various failure and attack scenarios.

6.2 LastFM Dataset

The LastFM dataset represents the social network of users from the LastFM music streaming service. It consists of user connections and their associated country information, highlighting both the social interactions and geographical distribution of users. The degree distribution shows a typical power-law behavior, indicating a few users with many connections and many users with fewer connections. The network's clustering coefficient suggests significant local clustering, and centrality measures identify key influential users. This dataset is valuable for examining network resilience, particularly in how the network responds to the addition or removal of nodes based on their centrality measures.

6.3 Conclusion

Both datasets provide rich insights into the structure and dynamics of social networks. The Email-Eu-core dataset is ideal for studying the resilience of professional communication networks, while the LastFM dataset offers a broader perspective on social interactions in an entertainment context. By examining these datasets, we can develop robust models for network resilience, understanding how networks can withstand and recover from various disruptions, whether through node removal or strategic attacks, and ensuring continuous connectivity and functionality.