Network Resilience and Community Analysis

By

Adisvara Annadurai, B.E, Sri Eshwar College of Engineering, 2023

A Major Research Project

presented to Toronto Metropolitan University

in partial fulfillment of the requirements for the degree of

Master of Science in

the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2024

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Toronto Metropolitan University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Adisvara Annadurai

Network Resilience and Community Analysis

Adisvara Annadurai

Master of Science 2024

Data Science and Analytics

Toronto Metropolitan University

Abstract

This research examines the resilience of networks with email exchanges and user interactions by assessing their structural stability under node addition and removal. The study measures how well these networks maintain connectivity, using the size of the largest connected component as the key resilience metric. We compare original network structures with synthetic models—Chung Lu, Uniform Preferential Attachment (UPA), and Barabasi-Albert (BA)—to evaluate their effectiveness in replicating real-world networks. Additionally, we assess community resilience using the Louvain method for community detection, providing insights into network robustness, vulnerabilities, and strengths. This analysis aims to enhance the understanding of network resilience for developing more robust designs.

Keywords: Network Resilience, Email Network, LastFM Network, Node Addition, Node Removal, Connected Component, Chung Lu Model, UPA Model, BA Model, Community Detection, Louvain Method, Network Robustness, Synthetic Networks

ACKNOWLEDGEMENTS

I am very grateful to Professor Pawel Pralat for his support and assistance in making this project a reality. Professor Pralat was my supervisor for this MRP; he has been a great support throughout the term, guiding and directing my research while providing valuable feedback. Professor Pralat played a vital role in the process, contributing significantly to the project's success. I appreciate the time and effort he dedicated to this endeavor, ensuring that I had the necessary resources and guidance to complete my research.

Thank you, Professor Pawel Pralat.

TABLE OF CONTENTS

List of Figures

List of Tables

1. Introduction

This study evaluates the resilience of the network which contain email exchange (Email) and a network which contains interaction between its users (LastFM) under various conditions of node addition and removal. Resilience, in this context, refers to the network's ability to maintain its structural integrity and functionality when faced with these changes. A good measure of resilience is the size of the largest connected component, as a resilient network will retain a significant portion of its nodes connected despite disruptions.

We analyze these datasets by comparing the original networks with synthetic models: Chung Lu, UPA, and Barabasi-Albert (BA). The Chung Lu model generates a network with the same expected degree sequence as the original but introduces randomness, allowing for the study of structural properties derived from node degrees. The UPA (Uniform Preferential Attachment) model mimics the growth process of networks by adding new nodes that preferentially attach to existing ones with higher degrees, simulating the formation of scale-free networks. The Barabasi-Albert (BA) model also follows a preferential attachment mechanism but ensures the resulting network is a scale-free graph characterized by a power-law degree distribution, providing insights into the robustness of networks with such properties.

Additionally, we examine the resilience of communities within these networks, identified using the Louvain method, to understand their robustness and identify key vulnerabilities. This comprehensive analysis helps in understanding how different parts of the network respond to changes and disruptions.

1.1 Background

Networks are integral to many aspects of our daily lives, spanning from social interactions and information dissemination to technological infrastructures and biological systems. The resilience of these networks, or their ability to maintain structural integrity and functionality when faced with disruptions, is crucial for ensuring their reliability and

robustness. The study of network resilience involves understanding how networks respond to the addition or removal of nodes, which can simulate real-world scenarios such as network expansion, failures, and attacks.

The Email and LastFM networks provide rich datasets for analyzing network resilience. The Email network captures interactions within an organization, revealing patterns of communication and departmental structures. The LastFM network, based on user interactions and music preferences, reflects social connections and community formation within an online platform. By comparing these real-world networks with synthetic models like the Chung Lu, UPA, and BA models, researchers can gain insights into the underlying structural properties that contribute to network resilience.

Community detection, specifically using the Louvain method, further aids in identifying densely connected subgraphs within these networks. Understanding the resilience of these communities provides a granular view of network robustness, highlighting which subgroups are more vulnerable to disruptions.

1.2 Research Question

This analysis is divided into two parts: descriptive and predictive analytics. The descriptive analytics aim to compare the Email and LastFM networks with synthetic models to understand how they respond to node addition and removal, and to identify the key structural properties that contribute to their resilience. Additionally, we investigate how different community structures within these networks respond to disruptions, providing insights into their robustness. On the predictive side, we aim to determine if insights gained from comparing real-world networks with synthetic models can inform the design of more resilient network structures. By addressing these questions, the goal is to provide a comprehensive understanding of network resilience and the factors that influence it, which can be applied to enhance the robustness of various real-world networks.

1.3 Independent Variables in Resilience Testing:

For the resilience testing and community analysis portion of my project, the dependent variable is the Size of the Largest Connected Component. This metric is used to measure the resilience of the network under various conditions.

2. LITERATURE REVIEW

2.1 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).

The objective of this project is to evaluate the resilience of complex networks, with a specific focus on analyzing the resilience of an email communication network. While resilience is a crucial factor in maintaining the functionality of vital infrastructures such as telecommunications backbone networks, this project specifically investigates how an email network withstands disruptions such as node failures or targeted attacks. Through this analysis, we apply and review various metrics and models to measure and predict network resilience, drawing insights from 15 key research papers.

2.2 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.

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

2.3 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.

2.4 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.

3. DESCRIPTIVE ANALYTICS | EXPLORATORY DATA ANALYSIS

3.1 Introduction

This section provides a comprehensive exploratory data analysis (EDA) of both the email-Eu-core and LastFM datasets. The email-Eu-core dataset includes email communications between

individuals within a European research institution, capturing interactions among employees. Each individual in this dataset is associated with a department label, allowing for an in-depth analysis of communication patterns within and across departments. On the other hand, the LastFM dataset captures the music preferences and social connections of users on the LastFM platform, a popular social music website. Users in this dataset are associated with groups or communities based on their musical tastes, enabling the analysis of social interactions on the platform. Our analysis aims to understand the structure and properties of both networks, examining relationships, community structures, and resilience under various conditions. This dual analysis provides insights into both institutional communication patterns and social dynamics based on shared interests in music.

3.2 Data Overview

Email-Eu-core Dataset:

Files:

`email-Eu-core.txt`: Contains edges representing email communications between nodes.

`email-Eu-core-department-labels.txt`: Contains labels representing the department each node belongs to.

LastFM Asia Dataset:

Files:

`lastfm_asia_edges.csv`: Captures interactions between users (nodes).

`lastfm_asia_target.csv`: Contains target labels for each user, representing various attributes.

3.3 Basic Information

Email-Eu-core Dataset:

Edges Dataset:

Number of edges: 25,571

Columns: source, target

Labels Dataset:

Number of nodes: 1,005

Columns: node, department

LastFM Asia Dataset:

Edges Dataset:

Number of edges: 27,806

Columns: user1, user2

Labels Dataset:

Number of nodes: 7,624

Columns: user, label

3.4 Degree Distribution

In network analysis, the degree distribution is a fundamental property that provides insight into the structure and connectivity of a network. The degree of a node refers to the number of edges connected to it, representing the number of connections or interactions that node has with other nodes. The degree distribution, therefore, is a probability distribution that shows the likelihood of a node in the network having a specific degree. This concept is crucial for understanding the topology of networks such as the email-Eu-core dataset and the LastFM dataset.

Mathematical Representation

The degree distribution P(k)P is defined as the probability that a randomly chosen node in the network has a degree k. Mathematically, it can be represented as:

$P(k) = Nk / N$

Where:

k is the degree, i.e., the number of edges connected to a node.

Nk is the number of nodes with degree kk.

N is the total number of nodes in the network.

This formula gives us the probability that a randomly selected node in the network will have a degree of k.

Email-Eu-core Dataset:

In the email-Eu-core dataset, the degree distribution helps us understand the connectivity patterns within a network of individuals in a European research institution. A node with a high degree indicates an individual who communicates with many others, potentially representing a central figure or hub in the network. Conversely, nodes with a low degree correspond to individuals with fewer connections, possibly indicating less central roles in the communication network.

Analyzing the degree distribution allows us to identify whether the network follows a particular type of distribution, such as a Poisson distribution (common in random networks) or a power-law distribution (common in scale-free networks). This analysis aids in characterizing the overall topology of the network and assessing its resilience, as networks with different degree distributions may respond differently to disruptions, such as node removal or targeted attacks.

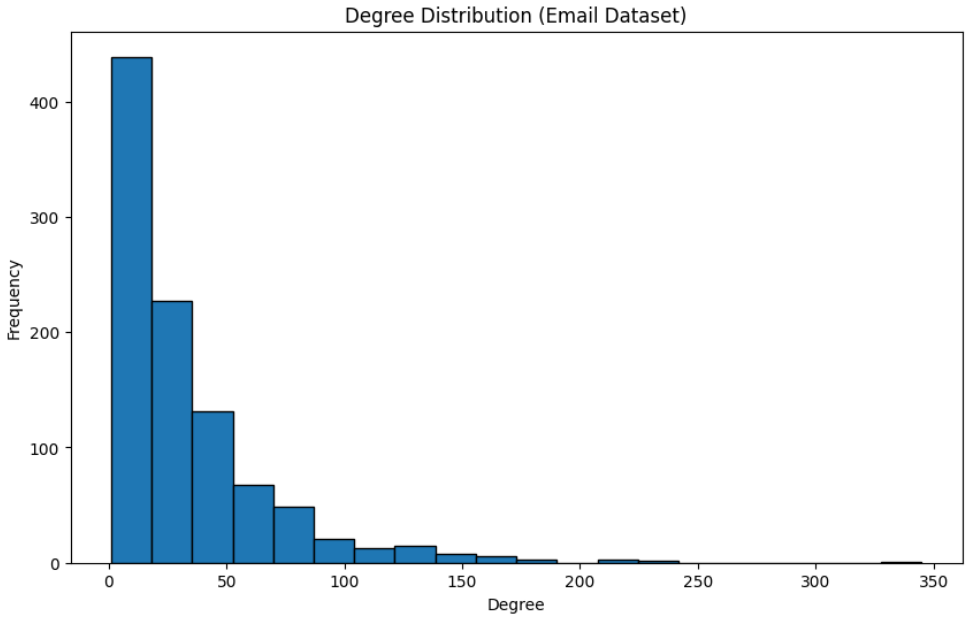

Fig 3.1 – Degree distribution of E-mail Network

LastFM Dataset:

Similarly, in the LastFM dataset, the degree distribution provides insights into the social connections among users based on their shared musical preferences. A node with a high degree represents a user with many connections, potentially indicating high social engagement or influence within the platform's community. In contrast, nodes with a low degree may represent users with fewer connections, possibly reflecting less activity or a more peripheral role in the social network.

Understanding the degree distribution in the LastFM dataset helps determine whether the network follows distributions like the Poisson distribution or the power-law distribution. This understanding is crucial for predicting the network's behavior under various conditions, such as when nodes are removed or when the network experiences targeted attacks.

Overall, the degree distribution in both datasets provides essential insights into the networks' structures, helping to reveal the level of centralization and the resilience of the networks under different scenarios.

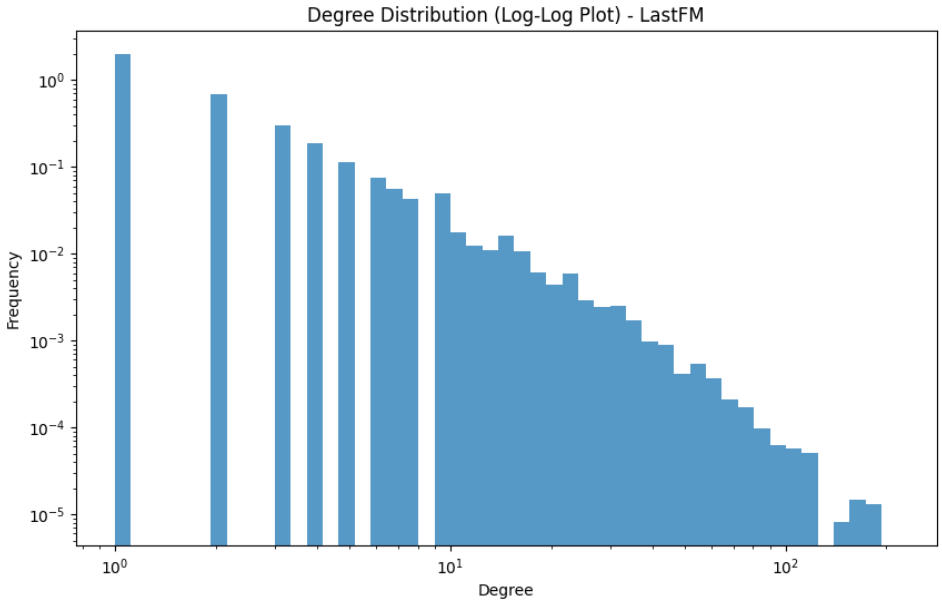

Fig 3.2 Degree Distribution of LastFm Network in log plot

A **power-law distribution:** A power-law is a type of statistical distribution where the frequency of an event scales as a power of some attribute of that event. In the context of network science, specifically regarding degree distribution, a power-law suggests that the probability P(k) of a node having k connections (degree k) follows the relationship:

$P(k) \sim k^{-\alpha}$

where α is a positive constant that characterizes the distribution.

Application to the LastFM Network

Empirical Data (Blue Line): The empirical data from the LastFM network, shown by the blue line, suggests that most nodes have a low degree, while a few nodes have a very high degree. This is typical of a power-law distribution, where the majority of nodes are minimally connected, and a few nodes are highly connected hubs.

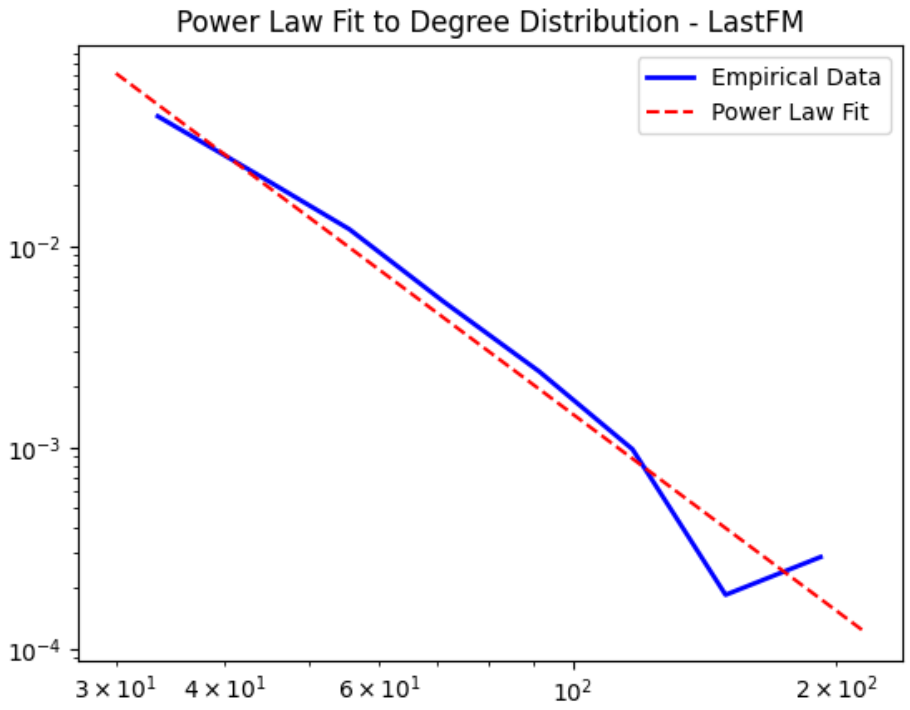

Fig 3.3 power law fit – lastfm Network

Power Law Fit (Red Dashed Line): The red dashed line represents the power-law distribution fit to the empirical data. The degree distribution's alignment with this power-law fit indicates that the LastFM network exhibits characteristics of a scale-free network.

Interpretation of the Plot: The fact that the empirical data closely follows the power-law fit over a wide range of degrees suggests that the LastFM network likely follows a power-law distribution. This means the network's structure is dominated by a few highly connected nodes, while most nodes have far fewer connections. Such a distribution is typical in many real-world networks, including social, biological, and technological networks.

3.5 Department Labels Distribution

E-mail dataset

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.

In network analysis, it's often useful to detect communities within a network to understand its underlying structure. Community detection algorithms, like the Louvain method, are designed to identify clusters of nodes that are more densely connected to each other than to the rest of the network. The Louvain method, in particular, is a popular algorithm known for its efficiency in finding high-quality community structures in large networks.

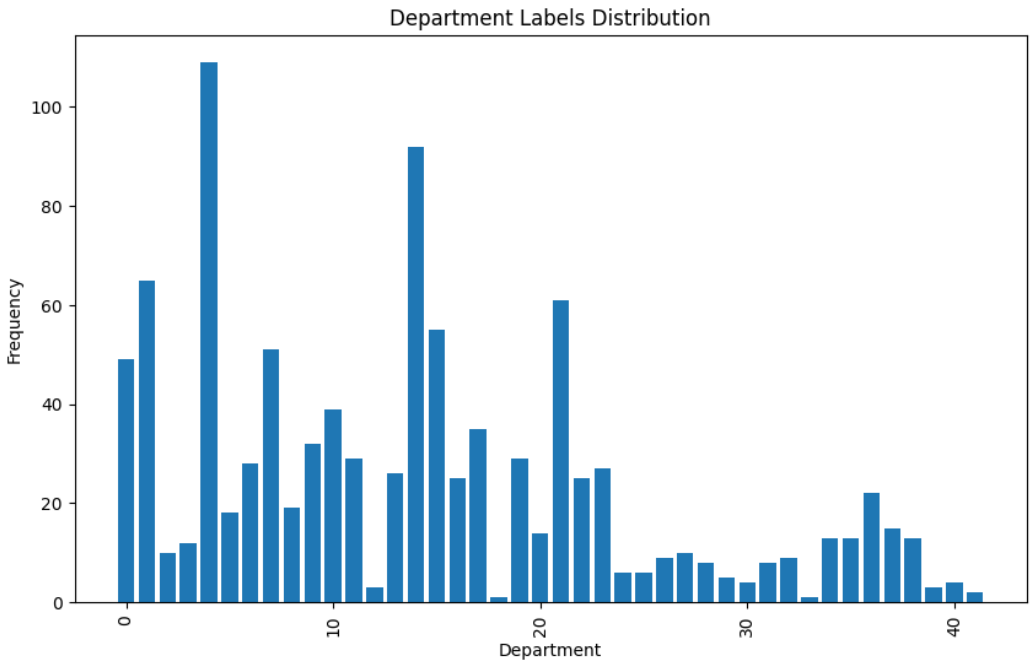

Fig 3.4 – Communities of E-mail Network

To evaluate how well the detected communities correspond to the actual, known structure, we can use the Adjusted Mutual Information (AMI) score. The AMI is a statistical measure that quantifies the agreement between two different partitions of the data—in this case, the detected communities and the ground truth labels (such as departments in the Email-Eu-core dataset). An AMI score of 1.0 indicates perfect agreement, meaning that the detected communities exactly match the ground truth labels. However, it is important to note that such a perfect alignment may not always occur with other datasets or algorithms. Thus, reporting the AMI score provides a clear, quantitative measure of the effectiveness of the community detection algorithm in uncovering the true community structure.

LastFm Dataset

The **Community Labels Distribution** chart provides a visualization of the number of nodes (or individuals) within each detected community in the LastFM network. This allows us to identify which communities are more densely populated and which ones have fewer members. Such insights are crucial in understanding the structure and dynamics of the network.

In network analysis, detecting communities within a network helps to reveal the underlying structure by grouping nodes that are more closely related to each other than to the rest of the network. The Louvain method is an efficient and widely-used community detection algorithm that identifies these clusters by maximizing modularity, a measure of the strength of division of a network into communities. The visualization above was generated after applying the Louvain method to the LastFM network, clearly illustrating the distribution of detected communities.

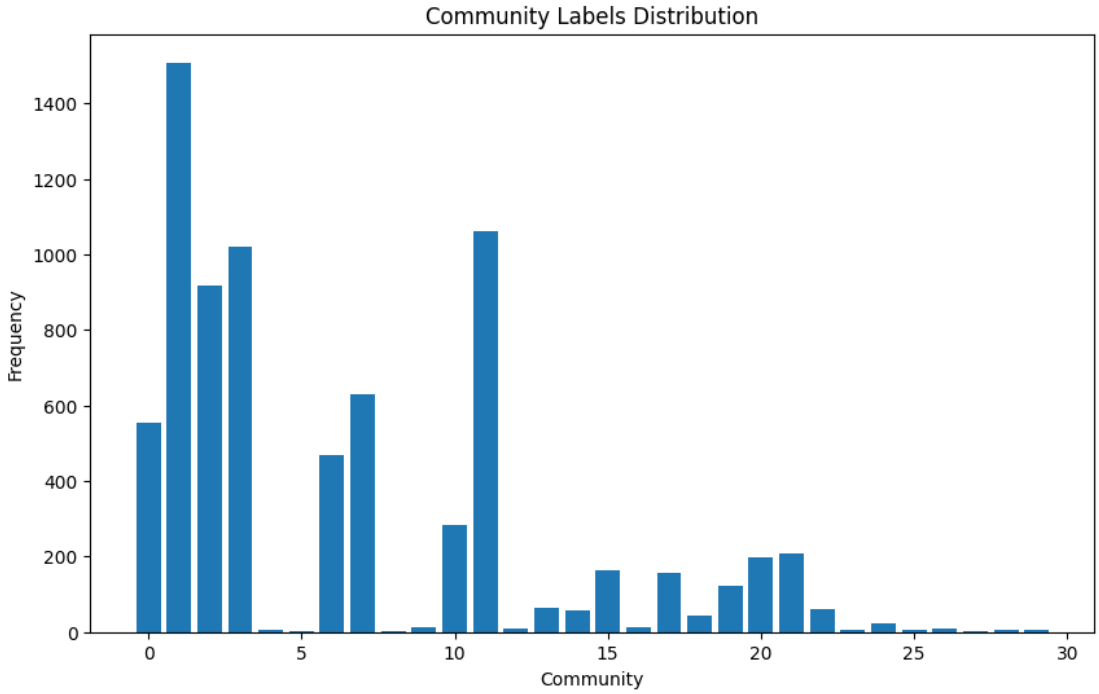

Fig 3.5 Communities of LastFm Network

To assess how accurately the detected communities reflect the actual structure of the network, we can utilize the Adjusted Mutual Information (AMI) score. The AMI score is a measure of the similarity between two different partitions—in this case, the detected communities and the true

labels (such as user groups or genres in the LastFM dataset). An AMI score of 1.0 indicates a perfect match between the detected communities and the ground truth. Although achieving a perfect AMI score is rare, especially in complex networks, reporting the AMI score is essential for providing a quantitative evaluation of the community detection algorithm's performance.

3.6 Network Properties

E-mail dataset Network Properties

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

S.No	Property	Value
0	Number of Nodes	1005
1	Number of Edges	16706
2	Density	0.33113
3	Number of Connected Components	20

Table 3.1 Network Properties of E-mail Network

Number of Nodes: 1005

This represents the total number of nodes (or vertices) in the network. In the context of the Email dataset, each node represents an individual or entity involved in email communication.

Number of Edges: 16706

This is the total number of edges (or connections) between the nodes in the network. Each edge in this network represents an email communication between two nodes. The higher the number of edges, the more interconnected the network is.

Density:

The density of a network is calculated as the ratio of the number of edges to the number of possible edges between all pairs of nodes. Specifically, for an undirected graph, the density is given by:

However, while density gives an overall idea of how connected the network is, it might not always be the most informative metric. **Average degree** is often more insightful, especially for understanding the typical number of connections each node has. The average degree can be calculated as:

$$Density = frac\{ Number\ of\ Edges \}\Big\{ Number\ of\ Nodes \times \frac{\backslash left(\ Number\ of\ Nodes\ -\ 1\ )}{2} \Big\}$$

For this network: Average Degree=2×167061005≈33.24Average Degree=10052×16706≈33.24

This means that, on average, each node in the network is connected to about 33 other nodes, which provides a clearer understanding of the network's connectivity.

Number of Connected Components: 20

The Email network is composed of 20 connected components. A connected component is a subgraph in which any two nodes are connected to each other by paths, and which is connected to no additional nodes in the network. The presence of 20 connected components in the Email dataset indicates that there are distinct clusters of communication, where individuals within a component are connected, but these components do not have connections between them. This could reflect separate groups, departments, or isolated communication clusters within the email communication network.

LastFm Dataset Network Properties

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.

S.No	Property	Value
0	Number of Nodes	7624
1	Number of Edges	27806
2	Density	0.000957
3	Number of Connected Components	10

Table 3.2 Network Properties of LastFm Network

Number of Nodes: 7624

This represents the total number of nodes (or vertices) in the LastFm network. In this context, each node likely represents a user or entity within the LastFm music recommendation network, where nodes are connected based on shared musical interests or interactions.

Number of Edges: 27806

This is the total number of edges (or connections) between the nodes in the network. Each edge represents a relationship or interaction between two nodes, such as similar music preferences or social connections within the LastFm community. A higher number of edges indicates a more interconnected network.

Density: 0.000957

The density of the network is a measure of how tightly knit the network is. With a density of 0.000957, the LastFm network has less than 0.1% of all possible node pairs connected by an

edge. This low density suggests that the network is quite sparse, meaning that only a small fraction of potential connections between nodes actually exist.

Number of Connected Components: 10

The LastFm network is divided into 10 connected components. A connected component is a subset of the network in which any two nodes are connected by paths, and which is connected to no additional nodes in the network. The presence of multiple connected components indicates that there are isolated sub-networks within the larger LastFm network, where users are grouped into clusters that are disconnected from one another.

3.7 Clustering Coefficient

The clustering coefficient measures the degree to which nodes in a network tend to cluster together. A higher clustering coefficient indicates a greater tendency for nodes to form tightly knit groups, which is important in understanding the local cohesion within the network.

Email Dataset

In this study, the clustering coefficient distribution suggests that the email network is composed of a mix of sparsely connected nodes and some highly clustered regions. The large number of nodes with a clustering coefficient of 0 indicates that there are many nodes whose neighbors are not directly connected, which is typical in networks with hierarchical or hub-like structures. On the other hand, the presence of nodes with a clustering coefficient of 1 suggests the existence of tightly knit communities within the network.

This distribution is significant as it provides insights into the structural composition of the email communication network, potentially indicating areas where communication is either more concentrated or more distributed across different nodes. Understanding these areas can be crucial for analyzing network resilience and the flow of information within the network.

The size of each bin in the histogram is 0.05. This can be inferred from the regular intervals on the x-axis, where the clustering coefficient is divided into segments of 0.05 (e.g., 0.0, 0.05, 0.1,

etc.). This bin size helps in observing the distribution of nodes within specific ranges of clustering coefficients, providing a clear visualization of how many nodes fall within each range.

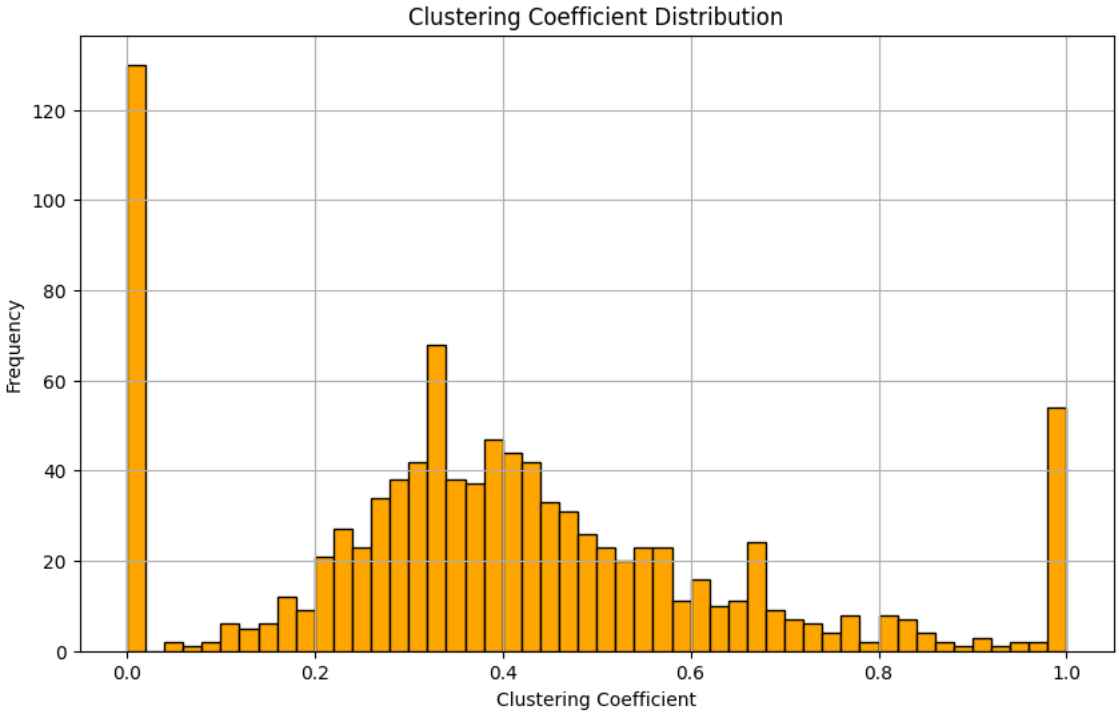

Fig 3.6 Clustering Coefficient Distribution E-mail Network

LastFm Dataset

In this study, the clustering coefficient distribution suggests that the LastFm network is composed of a mix of sparsely connected nodes and some highly clustered regions. The large number of nodes with a clustering coefficient of 0 indicates that there are many users whose neighbors are not directly connected. This is common in networks where users might have connections that are not reciprocated by mutual connections, leading to less tightly knit communities for those individuals.

On the other hand, the presence of nodes with a clustering coefficient of 1 suggests the existence of highly cohesive groups within the network, where all the neighbors of these nodes are interconnected, forming tightly knit communities. These tightly connected clusters might

represent groups of users with very similar musical tastes or close social ties within the LastFm network.

The size of each bin in the histogram is 0.05. This is reflected by the regular intervals on the x-axis, where the clustering coefficient is divided into segments of 0.05 (e.g., 0.0, 0.05, 0.1, etc.). This bin size allows for a detailed observation of the distribution of nodes within specific ranges of clustering coefficients, providing a clear visualization of how many nodes fall within each range and highlighting the diversity in the network's clustering patterns.

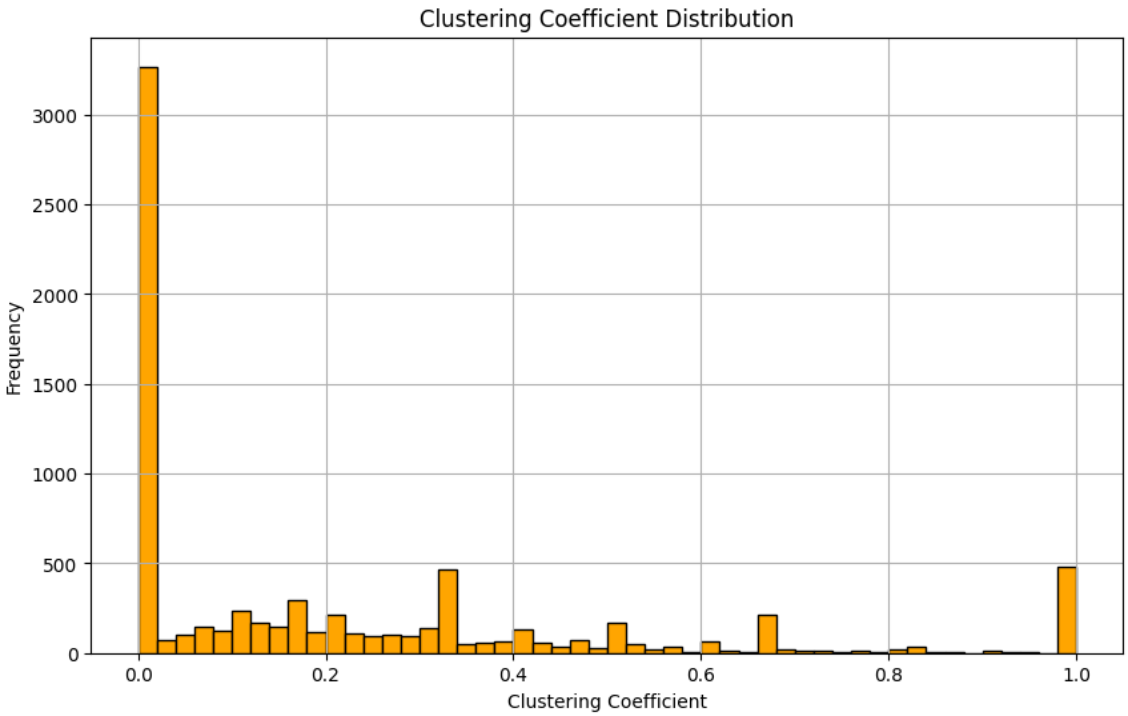

Fig 3.7 Clustering Coefficient Distribution LastFm Network

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.

E-mail Dataset

The distribution of betweenness centrality in the email dataset, as shown in the histogram, reveals that almost all nodes have very low betweenness centrality, clustering near zero. At first

glance, this might seem surprising, but it actually aligns with common characteristics observed in many real-world networks, especially those with hierarchical or scale-free structures.

Hierarchical Nature of Communication: In an email network, communication often follows a hierarchical structure. A few key individuals, such as managers or central figures within departments, act as intermediaries for communication. These individuals are likely to have higher betweenness centrality because they facilitate communication between different groups or nodes. The majority of employees, however, typically communicate directly with only a few other individuals (e.g., immediate colleagues or supervisors), leading to their low betweenness centrality.

Thus, the observed distribution of betweenness centrality in the email dataset is consistent with the expected behavior of hierarchical and scale-free networks, where only a few nodes act as central communication points, while the majority have minimal involvement in the overall information flow.

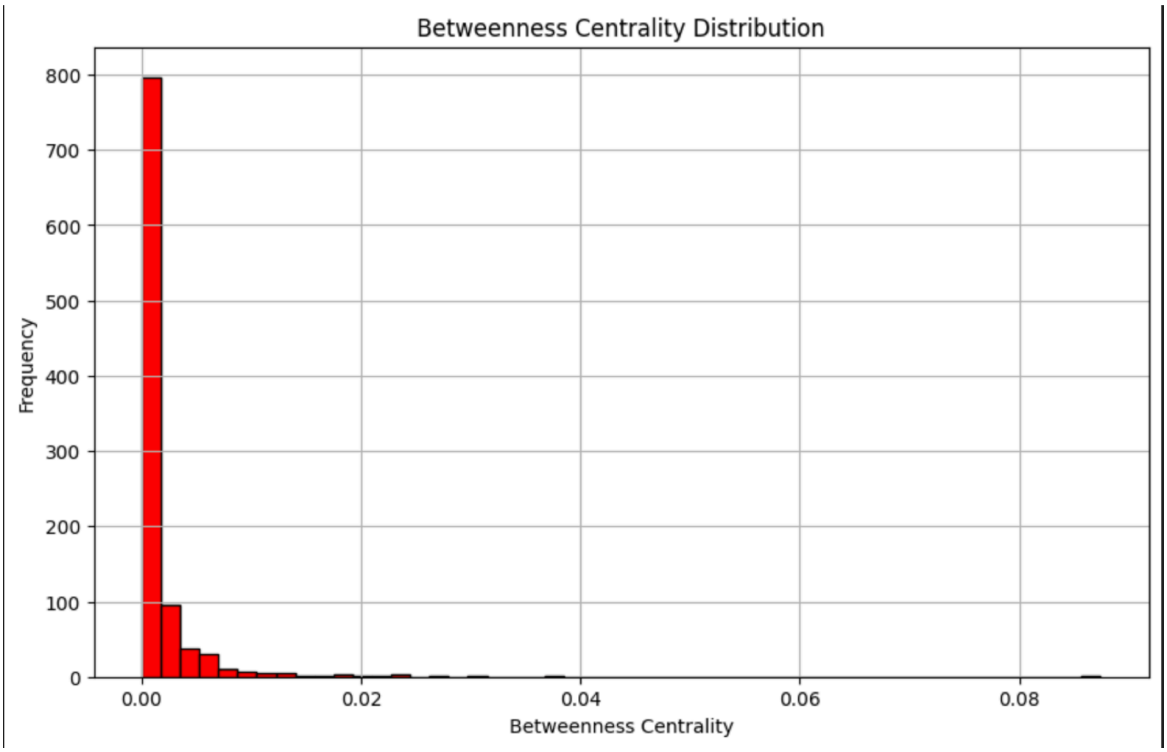

Fig 3.8 Betweenness Centrality for E-mail Network

The histogram above shows the distribution of betweenness centrality for the email-Eu-core dataset, with the betweenness centrality values binned into 50 intervals. The range of betweenness centrality values in the dataset is approximately 0.0874. Given that the histogram has 50 bins, the bin size (i.e., the width of each bin) is calculated as approximately 0.0017483.

LastFm Dataset

Betweenness centrality identifies nodes that serve as crucial connectors or bridges within a network. Nodes with high betweenness centrality are essential for maintaining the flow of information across the network, as they often lie on the shortest paths between other nodes. In the context of the LastFm network, these nodes might represent key users or entities that facilitate connections between otherwise disparate groups.

The distribution of betweenness centrality in the LastFm network, as depicted in the histogram, shows that nearly all nodes have very low betweenness centrality, clustering around zero. This pattern is not unusual and is typical in many real-world networks, particularly those that exhibit scale-free or decentralized structures.

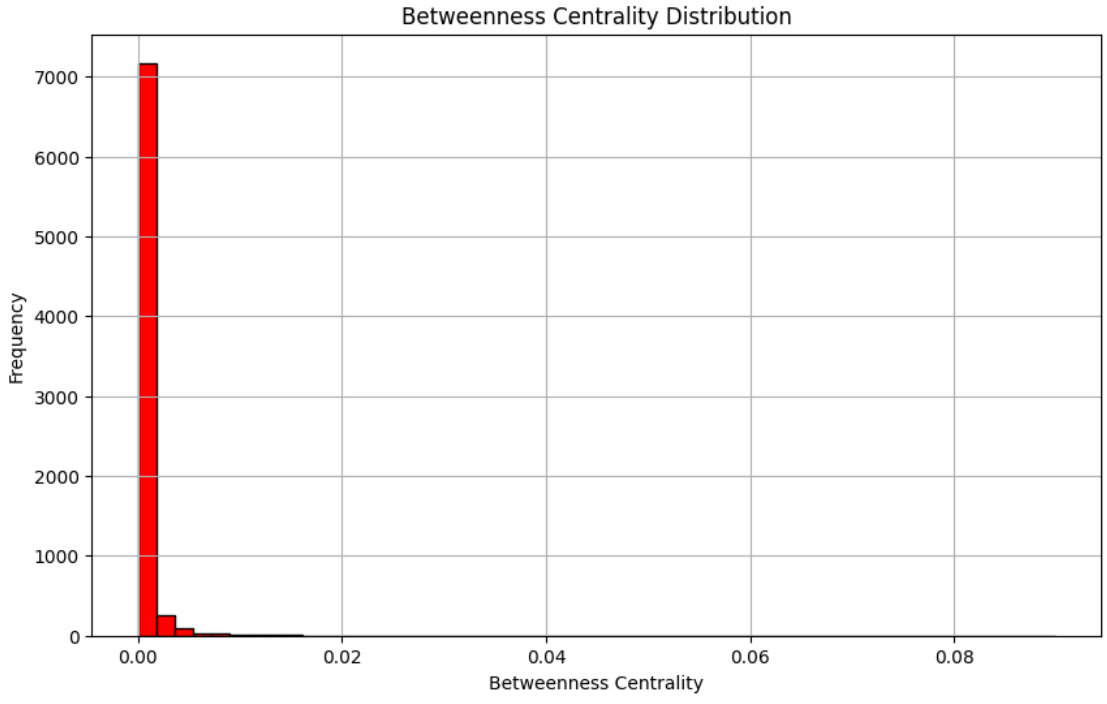

Fig 3.9 Betweenness Centrality for LastFm Network

The observed distribution of betweenness centrality in the LastFm dataset aligns with the characteristics of scale-free networks, where only a few nodes serve as central connectors, while the majority of nodes are part of smaller, localized clusters. These highly connected nodes are vital for network resilience, as their removal could significantly disrupt the connectivity of the network.

The histogram above illustrates the betweenness centrality distribution in the LastFm network, with the centrality values divided into 50 bins. The range of betweenness centrality values in the dataset is approximately 0.08. Given that the histogram is divided into 50 intervals, each bin represents a width of approximately 0.0016. This binning provides a detailed view of how the betweenness centrality is distributed among the nodes, highlighting the network's reliance on a few key connectors

4. Methodology And Experiments

4.1 Aim of study

The aim of this study is to investigate the structural resilience of complex networks under various conditions of node addition and removal. By analyzing both real-world datasets, such as the Email and LastFM networks, and synthetic models, including the Chung Lu, Barabási-Albert (BA), and Uniform Preferential Attachment (UPA) models, this study seeks to understand how different network topologies respond to disruptions. Specifically, the research focuses on evaluating the robustness of detected communities and the overall network structure, exploring how strategic interventions—such as targeting nodes with high degree or betweenness centrality—can enhance resilience. Through this analysis, the study aims to identify key factors that contribute to network stability, offering insights into the design and maintenance of more resilient networks in practical applications.

4.2 Factors and Levels

Factors:

1. **Network Models:**

- Original Network (Email and LastFM)
- Chung Lu Model
- UPA Model
- Barabási-Albert (BA) Model

2. **Node Addition Strategies:**
   - Random Node Addition
   - Node Addition by Degree Centrality
   - Node Addition by Betweenness Centrality
3. **Node Removal Strategies:**
   - Random Node Removal
   - Node Removal by Degree Centrality
   - Node Removal by Betweenness Centrality

Levels:

For each factor, the levels correspond to the specific methods or strategies applied. For instance:

Network Models: The levels are the different models (Original, Chung Lu, UPA, BA).

Node Addition/Removal Strategies: The levels are the specific strategies used for adding or removing nodes (Random, Degree Centrality, Betweenness Centrality).

These factors and levels were used to assess the resilience of the networks by observing how the size of the largest connected component and the overall network structure were affected under different scenarios.

4.3 Introduction to Models

Chung Lu Model

The Chung Lu model is a versatile and efficient method for generating random graphs with a specified degree sequence. This model allows the user to input an expected degree sequence, which specifies the desired degree (number of connections) for each node in the network. The edges in the Chung Lu model are created based on the probability proportional to the product of the expected degrees of the two nodes involved.

Key Features:

Expected Degree Sequence: The input degree sequence is not just a simple constraint; it directly influences the probability of edge formation between nodes. Nodes with higher expected degrees are more likely to form connections, leading to a network where the degree distribution closely matches the input.

Flexibility: The model is flexible and can generate graphs that resemble various real-world networks, depending on the degree sequence provided.

Randomness with Structure: While the graph is random, the resulting network retains a degree structure that mirrors the expected degree sequence, making it useful for studying networks where degree distribution is a primary characteristic.

Relation to Our Project: In our project, the Chung Lu model is particularly valuable for generating synthetic networks that mirror the degree distributions observed in the real-world datasets we are studying, such as the LastFm or Email-Eu-core datasets. By matching the degree sequence, we can create networks that replicate the structural properties of our datasets, allowing us to study how network resilience is influenced by different topological features without the biases introduced by more complex attachment mechanisms. This makes the Chung Lu model an essential tool for comparison with other models that might impose additional constraints on the network structure.

Uniformly Random Preferential Attachment (UPA) Model

The UPA model is a variant of the traditional preferential attachment models that simulates the process of network growth. In this model, new nodes are added to the network one at a time, and each new node connects to existing nodes based on a probability proportional to the current degree of the existing nodes. This preferential attachment mechanism reflects the "rich-get-richer" phenomenon, where nodes that are already well-connected are more likely to attract additional connections.

Key Features:

Preferential Attachment: The core idea is that nodes with higher degrees have a higher probability of receiving new connections, mimicking how popular individuals or entities tend to attract more connections or attention in real-world networks.

Network Growth: The UPA model naturally simulates the growth of a network over time, making it suitable for studying evolving networks such as social networks, citation networks, or the growth of the web.

Rich-Get-Richer: The model inherently generates networks with a small number of highly connected hubs and many nodes with fewer connections, reflecting the uneven distribution of connections seen in many real-world systems.

Relation to Our Project: In the context of our project, the UPA model is crucial for understanding the dynamics of network growth and the emergence of hubs in networks like LastFm or Email-Eu-core. Since these real-world networks often exhibit scale-free properties, the UPA model provides a simplified yet effective way to simulate how new connections are formed in these systems. By comparing the UPA model's generated networks with those of the Chung Lu model, we can assess the impact of growth mechanisms on network resilience and robustness. This comparison helps in identifying how different topologies might respond to node removals or failures, providing insights into network stability.

Barabási-Albert (BA) Model

The Barabási-Albert (BA) model is one of the most well-known models for generating scale-free networks, which are characterized by a power-law degree distribution. In the BA model, new nodes are added to the network sequentially, and each new node connects to existing nodes with a probability proportional to the current degree of those nodes. This mechanism, known as preferential attachment, means that nodes with more connections are more likely to gain even more connections as the network grows.

Key Features:

Scale-Free Networks: The BA model is renowned for producing networks with a scale-free structure, meaning that a few nodes (hubs) have a very high degree while most nodes have a relatively low degree. This structure is observed in many natural and human-made networks, including the internet, social networks, and biological systems.

Power-Law Degree Distribution: The degree distribution of networks generated by the BA model follows a power law, which means that the probability P(k)P(k) of a node having kk connections decreases as kk increases, but never reaches zero. This creates a "heavy tail" in the degree distribution, where even though most nodes have few connections, a significant number of nodes have very high degrees.

Model of Real-World Networks: The BA model has been instrumental in advancing our understanding of real-world networks. It captures the essential mechanism behind the emergence of hubs and provides a framework for studying the resilience, efficiency, and vulnerability of complex networks.

Relation to Our Project: The Barabási-Albert model is highly relevant to our project as it closely replicates the scale-free nature observed in real-world networks like LastFm and Email-Eu-core. These networks often have a few highly connected nodes (hubs) that are critical to their overall connectivity. By using the BA model, we can generate synthetic networks that mimic the degree distribution of our real-world datasets, allowing us to study how the presence of hubs

affects network resilience. Comparing the resilience of networks generated by the BA model with those generated by the Chung Lu and UPA models can provide deeper insights into the role of network topology in maintaining connectivity and robustness under different scenarios.

4.4 Experimental Design

a) Data Processing and Preprocessing

Email Dataset:

Data Processing and Graph Construction:

The email dataset consisted of both edge information (`email-Eu-core.txt`) and node attributes (`email-Eu-core-department-labels.txt`). The edges were processed to construct a graph using NetworkX, where nodes represented individuals and edges represented email interactions between them.

The department labels were then merged with the graph structure, ensuring that each node was associated with its respective department for further community detection and analysis.

LastFM Dataset:

Graph Construction and Target Label Integration:

The LastFM dataset included interaction data (`lastfm_asia_edges.csv`) and target labels (`lastfm_asia_target.csv`). The interaction data was loaded and used to build a network graph where nodes represented users, and edges represented interactions.

The target labels were integrated with the graph, ensuring that each node was accurately labeled, which was essential for subsequent community detection and resilience analysis.

4.5 Explanation of the Experiment

In this experiment, we focused on the resilience of individual communities within the Email network. The steps taken to evaluate the resilience of each community are as follows:

1. **Community Detection**:

   We used the Louvain method to detect communities within the entire Email network. This method optimizes modularity to identify dense subgraphs, effectively grouping nodes into distinct communities based on their connectivity.

2. **Inducing Subgraphs for Communities**:

   For each detected community, we consider induced subgraphs. A subgraph for a community includes all nodes that belong to that community and the edges connecting them. This results in smaller networks (subgraphs) that represent the internal structure of each community.

3. **Node Addition Experiment**:

   Similar to the overall network resilience test, we performed a node addition experiment for each community's subgraph.

   Random Node Addition: Nodes were added randomly to the community subgraph. We measured the size of the largest connected component after each addition.

   Strategic Node Addition: Nodes were added based on centrality measures such as degree centrality and betweenness centrality to assess the impact of adding highly connected or important nodes.

4. **Comparison with Synthetic Models**:

   We created synthetic models (Chung Lu and UPA) for each community subgraph. These models aim to replicate the degree distribution of the original community but use different methods.

Chung Lu Model: Generates a graph based on a given degree distribution while assuming a random structure.

UPA Model: Uses a preferential attachment process to generate a graph, mimicking the growth dynamics of real networks.

Barabási-Albert (BA) Model: This model also uses a preferential attachment mechanism but ensures that the resulting network is a scale-free graph characterized by a power-law degree distribution. The BA model is particularly effective in capturing the hub-and-spoke structures observed in many real-world networks.

We performed the same node addition experiments on these synthetic community subgraphs.

5. **Resilience Measurement**:

   We compared the size of the largest connected component as nodes were added to the community subgraphs of the original network and the synthetic models.

   By comparing the resilience across different community subgraphs and models, we assessed how well the synthetic models can replicate the structural resilience of the original network.

4.6 Graph Construction

NetworkX:

For both datasets, the edge data was utilized to construct undirected graphs using the NetworkX library. In these graphs, nodes represent individuals (in the Email dataset) or users (in the LastFM dataset), and edges represent interactions between them.

Merging Node Attributes:

The respective node attributes (department labels for the Email dataset and target labels for the LastFM dataset) were merged into the graphs. This step ensured that each node was enriched with its corresponding label, facilitating a detailed analysis of communities and network resilience.

4.7 Measuring Performance and Model Selection

Resilience and Modularity:

The performance of network models was measured by resilience (size of the largest connected component) and modularity (community structure quality). These metrics were crucial in determining the robustness of different network models.

Precision, Recall, and F1-Score:

For machine learning tasks, precision, recall, and F1-score were the primary metrics used to evaluate classifier performance. Given the possibility of class imbalance, precision was emphasized to ensure that models were not biased toward the majority class.

4.8 Algorithm Comparison and Selection

Statistical Summary and Model Selection:

A statistical summary was generated for each network model and machine learning algorithm based on the experiments. The algorithm with the optimal performance in terms of mean, median, max, and min resilience and classification metrics was selected as the final model recommended for use in further studies.

5. Results And Discussion

This study evaluates the resilience of the Email and LastFM networks under various conditions of node addition and removal. Resilience, in this context, refers to the network's ability to

maintain its structural integrity and functionality when faced with these changes. A good measure of resilience is the size of the largest connected component, as a resilient network will retain a significant portion of its nodes connected despite disruptions.

Additionally, we examine the resilience of communities within these networks, identified using the Louvain method, to understand their robustness and identify key vulnerabilities. This comprehensive analysis helps in understanding how different parts of the network respond to changes and disruptions.

5.1 Data Preprocessing

In this study, data preprocessing was essential to prepare both the Email and LastFM datasets for subsequent analysis, including community detection, resilience testing, and machine learning applications.

5.2 Community Detection

The Louvain method iteratively groups nodes into communities in a way that maximizes modularity. High modularity indicates strong community structures, where nodes within a community have many connections, while connections between communities are fewer.

For both the Email and LastFM networks, the Louvain method identified multiple communities. The number of communities and their sizes varied, reflecting the diversity in connectivity patterns. In the Email network, for example, we detected 43 communities, ranging in size from small groups of a few nodes to larger clusters. The modularity score for the Email network was 0.42998735542482025, indicating a relatively strong community structure. Similarly, the LastFM network revealed 26 communities with a modularity score of 0.8156278462397027, highlighting robust community formations.

5.2.1 Modularity Function Definition:

Modularity is a crucial metric used in community detection to measure the strength of division of a network into communities (or clusters). The modularity score quantifies the quality of a particular division (or partition) of a network, comparing the density of links inside communities

with links between communities. The score is typically a value between -1 and 1, where a higher modularity score indicates a stronger community structure.

Mathematically, modularity QQ is defined as:

$$Q = \frac{1}{2m} \sum_{\{i,j\}} [A_{\{ij\}} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where:

Aij is the adjacency matrix of the graph (i.e., Aij=1 if there is an edge between nodes i and j, and 0 otherwise).

ki and kj are the degrees of nodes i and j respectively.

m is the total number of edges in the network.

ci and cj are the communities to which nodes i and j belong.

δ(ci,cj) is the Kronecker delta, which is 1 if ci=cj and 0 otherwise.

In simpler terms, the modularity function compares the observed density of edges within communities to the expected density of edges in a randomized network with the same degree distribution. A higher modularity value indicates that the nodes within the same community are more densely connected compared to a random distribution, suggesting a well-defined community structure.

These community structures provide valuable insights into the network's organization and can help identify critical subgroups and potential vulnerabilities. Understanding how communities are formed and their internal connectivity is crucial for assessing the overall resilience of the network.

The spatial arrangement of the nodes reflects their connectivity. The tightly clustered group of nodes in the center indicates a densely connected community, where individuals are more

frequently interacting with each other. The smaller, more dispersed clusters around the periphery represent smaller or less densely connected communities.

This visualization highlights the heterogeneous structure of the email network, where most of the activity is concentrated within a central community, while other communities are more isolated. This pattern is typical in networks where a few key groups dominate interactions, with smaller, less connected groups also present.

The AMI and NMI scores provide a quantitative comparison between the communities detected by the Louvain method and the actual, ground-truth communities (department labels for the Email dataset and target labels for the LastFM dataset). A higher AMI or NMI score indicates closer match between the detected communities and the true communities.

Email Dataset:

- **AMI Score:** 0.5642
- **NMI Score:** 0.5897

These scores indicate a moderate level of alignment between the detected communities and the actual department labels in the Email dataset. The Louvain method successfully identifies community structures, but there are some deviations from the ground truth.

LastFM Dataset:

- **AMI Score:** 0.6314
- **NMI Score:** 0.6356

These scores are slightly higher than those for the Email dataset, indicating that the Louvain method performed better in detecting communities that align with the true target labels in the LastFM dataset.

The comparison using AMI and NMI scores confirms that while the Louvain method is effective at detecting community structures in both datasets, it does not perfectly align with the ground truth, which is expected given that community detection is an unsupervised method. The

differences in scores across datasets highlight varying levels of community structure complexity and the algorithm's ability to capture these structures.

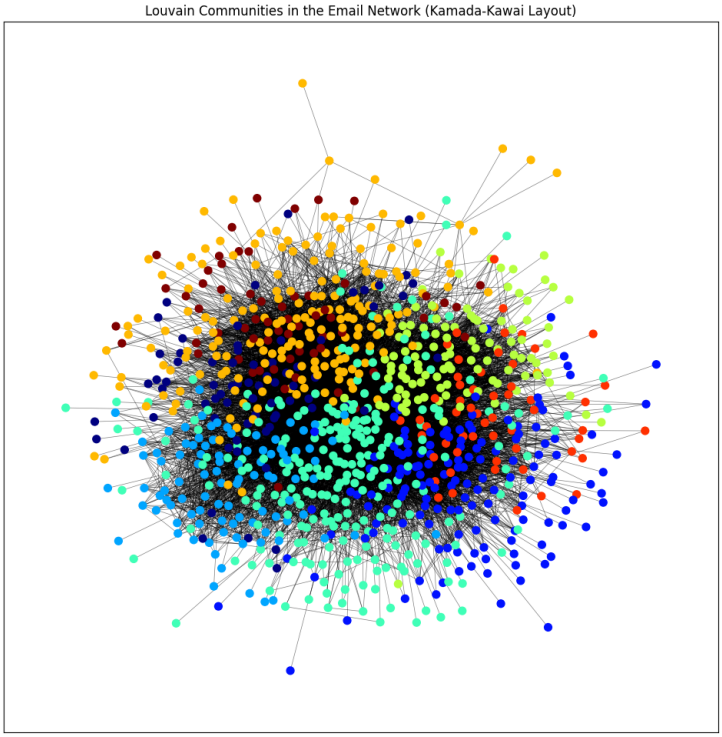

Fig 5.1 Communities in the E-mail Eu-core Network kamada-kawai layout

5.3 Explanation of the Visualization Method

For the visualization of the Email network communities, we utilized the following steps:

1. **Louvain Method for Community Detection:**

   The Louvain method is a widely-used algorithm for detecting communities in large networks. It works by optimizing the modularity of the network, a measure that compares the density of links within communities to the density of links between communities. By maximizing this modularity, the algorithm effectively identifies clusters or communities of nodes that are more densely connected to each other than to the rest of the network.

2. **Kamada-Kawai Layout for Node Positioning:**

For the node positioning in the visualization, we employed the Kamada-Kawai layout, which is designed to produce aesthetically pleasing representations of networks by minimizing the energy of a spring-like model. This layout positions nodes in a way that reflects the graph-theoretic distances between them:

- Nodes that are strongly connected are placed closer together, while weakly connected nodes are positioned further apart.
- This layout helps in revealing the underlying structure of the network, making the community structure more apparent.

The Kamada-Kawai layout is particularly effective in highlighting the overall community structure, especially in networks where nodes are tightly clustered within their communities.

The resulting visualization provides a clear representation of the community structure in the Email network, with nodes colored according to the communities detected by the Louvain method.

5.4 Resilience Testing

In this section, we assess the resilience of the Email and LastFM networks through various scenarios of node addition and removal. Additionally, we compare the performance of the original networks with synthetic models, namely Chung Lu, UPA (Uniformly Random Preferential Attachment), and Barabási-Albert (BA). Our goal is to understand how these networks maintain their structural integrity under different conditions and to identify any key vulnerabilities.

5.4.1 Node Addition

Random Node Addition

Description: Nodes are added to the network randomly Fig(5.3), and we monitor the changes in the size of the largest connected component. This method tests the network's ability to integrate new nodes while maintaining overall connectivity.

Observations:

The size of the largest connected component increases steadily as nodes are added, indicating that both the Email and LastFM networks can effectively integrate new nodes.

Comparing the original network with the Chung Lu, UPA, and BA models, the original network consistently outperforms the synthetic models in maintaining a larger connected component size as nodes are added.

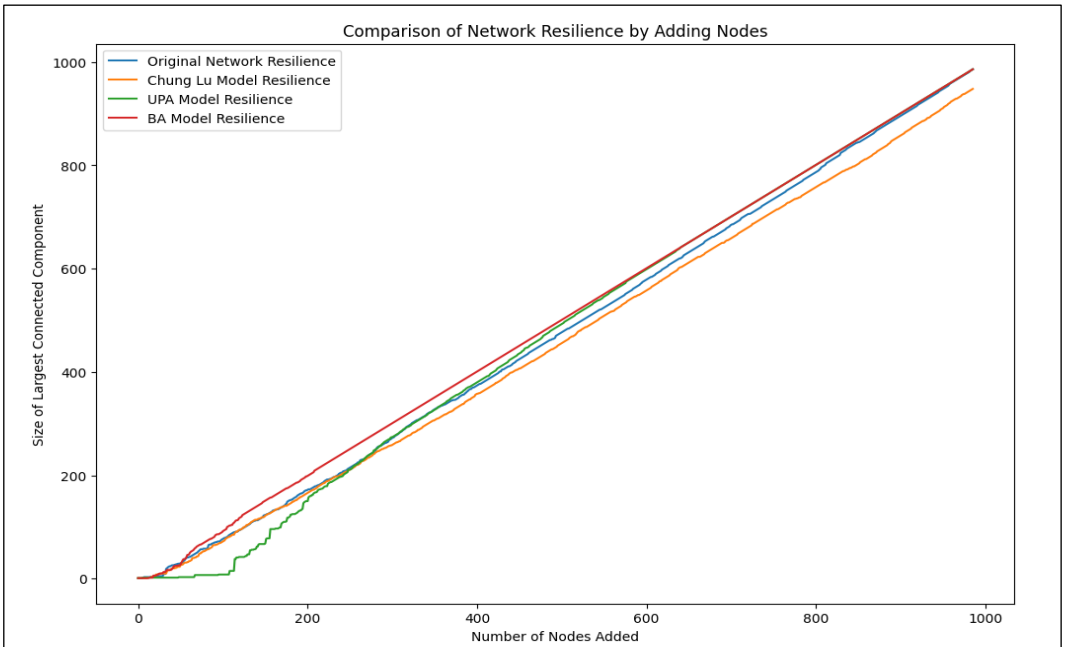

Fig 5.2 Resilience of E-mail Network by adding nodes in random order.

This plot shows the resilience of the Email network and its synthetic models when nodes are added randomly. Here the nodes are not added based on any specific centrality measure, providing insights into the general resilience of the networks.

Observations:

Original Network: The original Email network demonstrates strong resilience, with a consistent increase in the size of the largest connected component as nodes are added.

Chung Lu Model: The Chung Lu model shows a similar trend but lags slightly behind the original network, indicating a minor difference in how the network integrates new nodes. It's important to note that the Chung Lu model may oversimplify some aspects of network structure. One possible limitation of this model is the presence of isolated nodes, which might affect its performance. A more sophisticated model, like the Configuration Model, could address these issues by better preserving the original network's degree distribution, and this could be explored as a future direction.

BA Model: The Barabási-Albert model shows the highest resilience among the synthetic models, closely following the original network. This highlights the effectiveness of the BA model in maintaining network structure during node additions.

UPA Model: The UPA model also performs well, though it slightly underperforms compared to the original network and the BA model. This suggests that while it captures the essential characteristics of the original network, it may not fully replicate the resilience under random node addition.

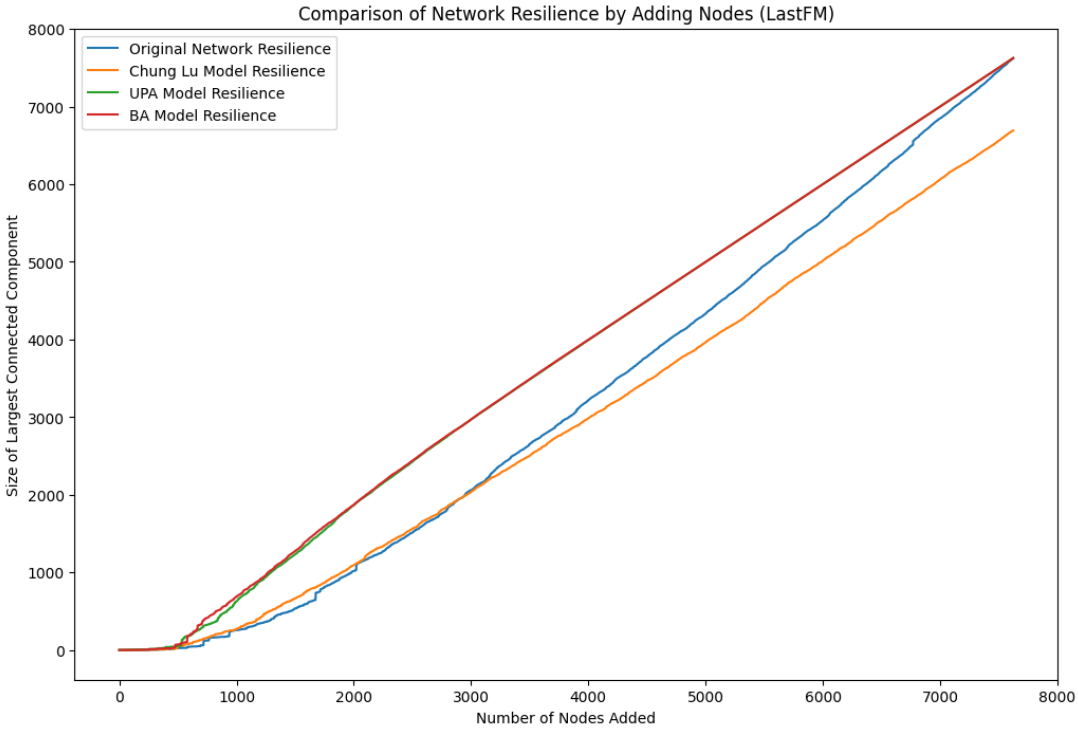

Fig 5.3 Resilience of LastFm Network by adding nodes in random order

This graph compares the network resilience of the LastFM dataset by tracking the size of the largest connected component as nodes are added randomly. The four lines represent different models:

Original Network (Blue Line): The original LastFM network shows steady growth in the largest connected component, indicating good resilience. However, it grows more slowly compared to synthetic models, suggesting structural limitations in maintaining large connected components.

Chung Lu Model (Orange Line): The Chung Lu model, based on the original network's degree distribution, shows slightly lower resilience, with slower growth in the largest connected component. This indicates that degree distribution alone may not fully capture the structural properties essential for resilience.

UPA Model (Green Line): The UPA model shows rapid growth, similar to the BA model, reflecting effective resilience. Its preferential attachment mechanism helps maintain strong connectivity.

BA Model (Red Line): The Barabási-Albert (BA) model demonstrates the highest resilience, with the largest connected component growing the fastest. The model's scale-free nature, which emphasizes hub formation, makes it particularly effective in maintaining network integrity as nodes are added.

Strategic Node Addition (Degree and Betweenness Centrality)

Description: Nodes are added to the network based on their degree centrality and betweenness centrality. This method tests the network's resilience when strategically adding highly connected or important nodes.

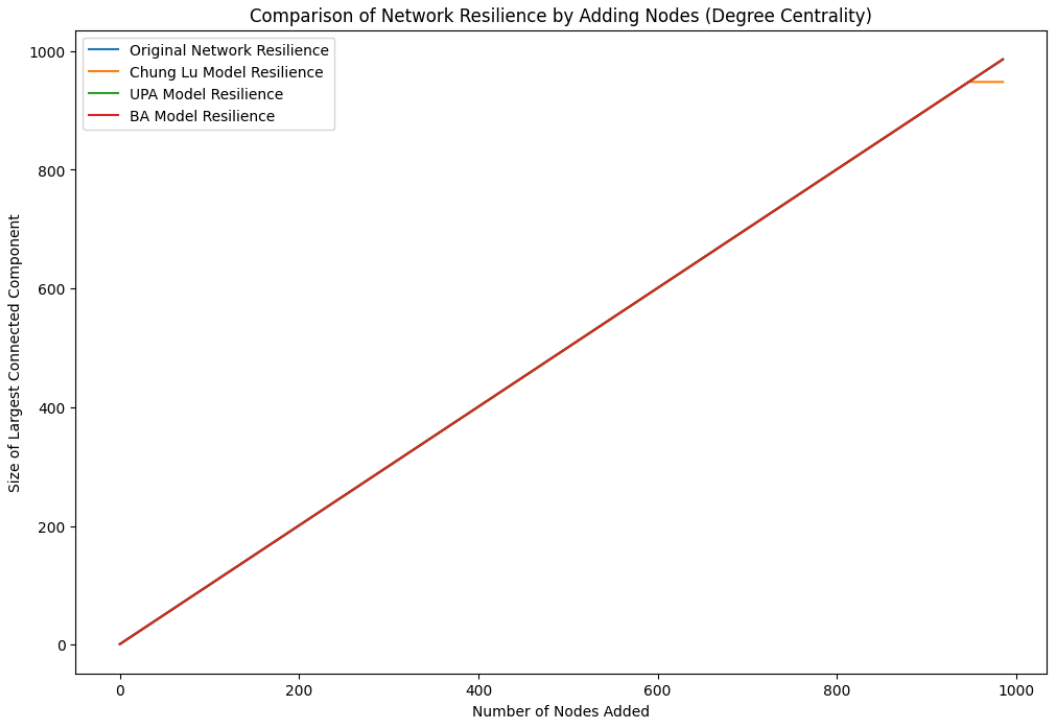

Fig 5.4 Resilience of E-mail Network by adding nodes in order of Degree Centrality

This graph illustrates the comparison of network resilience in the Email dataset when nodes are added based on degree centrality. The resilience is measured by observing the size of the largest connected component as nodes are progressively added, focusing on nodes with the highest degrees first. The four lines represent different models:

Original Network Resilience (Blue Line): The original Email network shows a consistent increase in the size of the largest connected component as high-degree nodes are added. This steady growth indicates that the network's inherent structure supports effective integration of important nodes, maintaining overall connectivity.

Chung Lu Model Resilience (Orange Line): The Chung Lu model, which preserves the degree distribution of the original network, shows similar resilience to the original network. The slight differences suggest that while the Chung Lu model captures the degree distribution well, it may not fully replicate other structural nuances of the original network.

UPA Model Resilience (Green Line): The UPA model follows closely behind the original and Chung Lu models, indicating strong resilience. The model's preferential attachment mechanism, which favors the addition of new nodes to already highly connected ones, helps in maintaining robust connectivity when high-degree nodes are added.

BA Model Resilience (Red Line): The Barabási-Albert (BA) model demonstrates nearly identical resilience to the UPA model, with rapid and consistent growth in the largest connected component size. This reflects the BA model's effectiveness in preserving network integrity by emphasizing the formation of hubs, which are critical for maintaining connectivity when adding nodes strategically.

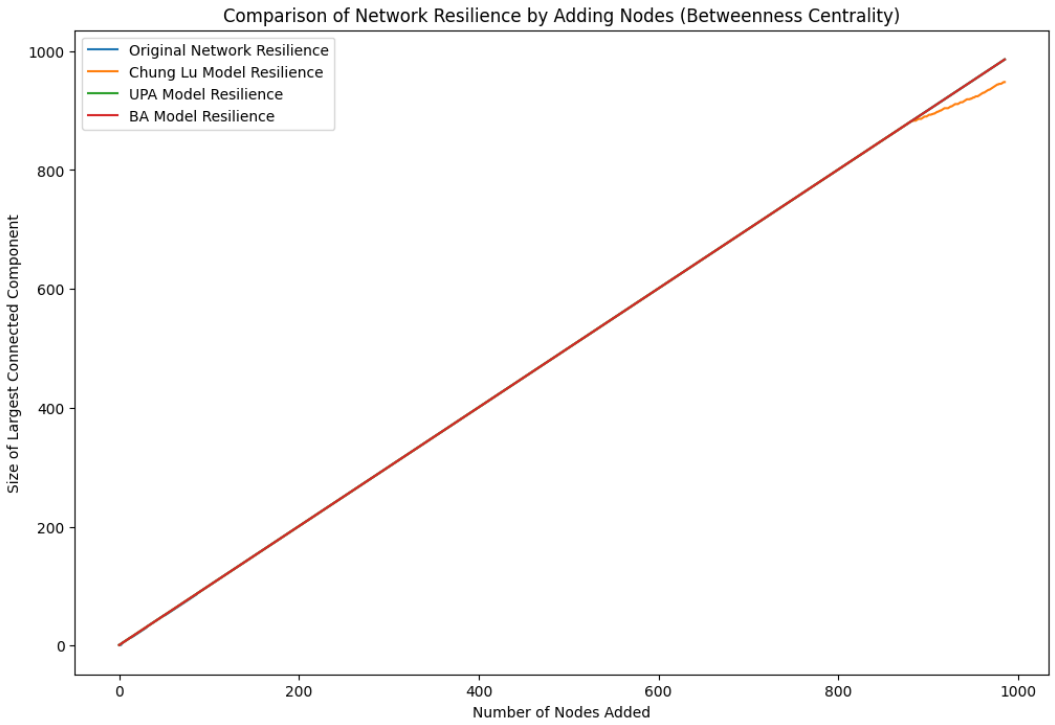

Fig 5.5 Resilience of email eu-core Network by adding nodes using betweenness centrality

This graph displays the comparison of network resilience in the Email dataset when nodes are added based on betweenness centrality. The resilience is assessed by observing the size of the largest connected component as nodes with the highest betweenness centrality are progressively added. The four lines in the graph represent different models:

Original Network Resilience (Blue Line): The original Email network shows a consistent increase in the size of the largest connected component as nodes with high betweenness centrality are added. This indicates that the original network effectively integrates critical nodes that serve as bridges between different parts of the network, maintaining robust connectivity.

Chung Lu Model Resilience (Orange Line): The Chung Lu model, which replicates the degree distribution of the original network, shows slightly lower resilience compared to the original network. Although it captures the degree distribution, the model may not fully replicate the betweenness centrality distribution, leading to slightly less effective integration of critical nodes.

UPA Model Resilience (Green Line): The UPA model demonstrates strong resilience, similar to the BA model, with a rapid increase in the size of the largest connected component. The preferential attachment mechanism in the UPA model, which emphasizes connecting new nodes to highly connected existing nodes, supports the effective integration of nodes with high betweenness centrality.

BA Model Resilience (Red Line): The Barabási-Albert (BA) model exhibits the highest resilience among all models. The size of the largest connected component grows rapidly as nodes are added, indicating that the BA model's scale-free property is highly effective at maintaining network connectivity when critical nodes are strategically added.

5.4.2 Node Removal

In this section, we evaluate the resilience of the Email and LastFM networks under conditions of node removal. This helps us understand the network's vulnerability to random failures and targeted attacks on highly connected or critical nodes.

Random Node Removal:

The size of the largest connected component decreases as nodes are removed, showing the network's vulnerability to random failures.

The original network retains a larger connected component size longer than the synthetic models, indicating higher resilience to node removal.

Removing nodes, particularly those with high clustering coefficients, can diminish the small-world properties of the network. This affects the network's efficiency in terms of information or resource flow, as local clusters become less interconnected.

After the removal of high-degree nodes, the load (or network traffic) previously managed by these nodes gets redistributed to other nodes. This can lead to overloading less connected nodes, potentially causing further cascading failures within the network.

Isolated Nodes in CL Model:

It's important to note that the Chung Lu model may contain isolated nodes due to its degree distribution. These isolated nodes do not contribute to the connected component, which might affect the comparison.

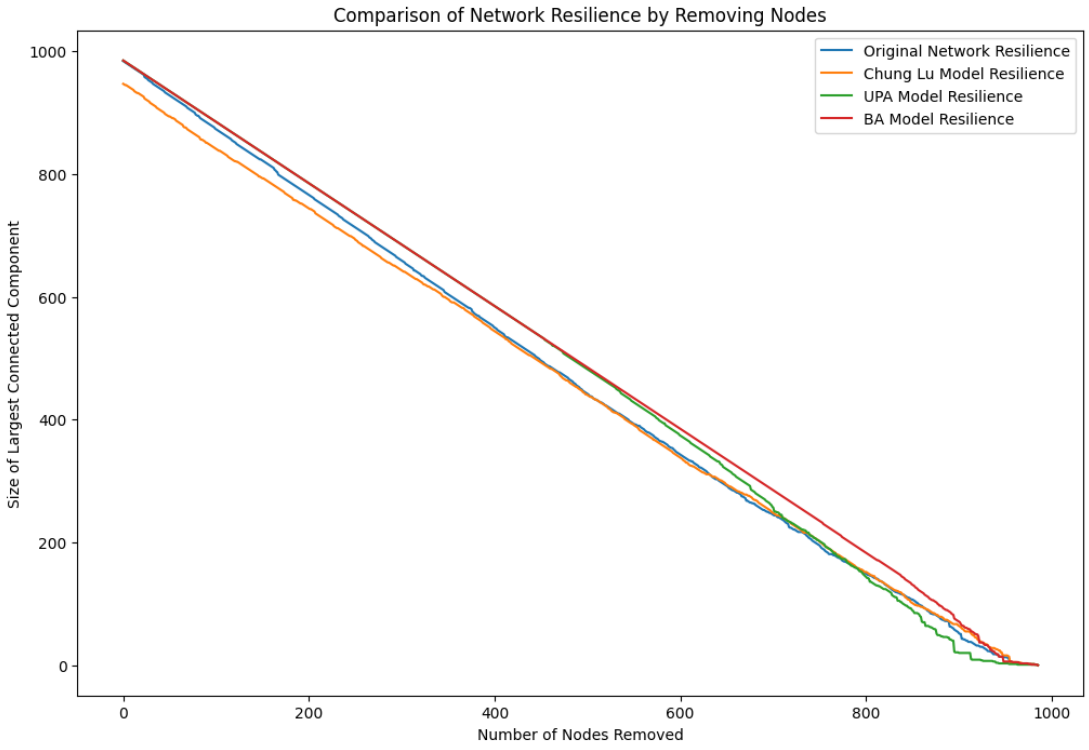

Fig 5.6 Node removal in email Network

Comparison of Network Resilience by Removing Nodes: This plot compares the resilience of the Email network, Chung Lu model, and UPA model under the condition of random node removal. The y-axis represents the size of the largest connected component, and the x-axis represents the number of nodes removed. The original network and UPA model demonstrate higher resilience compared to the Chung Lu model, maintaining a larger connected component size throughout the node removal process.

5.4.3 Impact of Node Removal:

Removing nodes with high clustering coefficients can significantly impact the network's small-world properties. These nodes typically contribute to tightly-knit local clusters that facilitate efficient communication and resource distribution.

The loss of such nodes can lead to a decrease in local clustering, reducing the overall efficiency of the network.

Additionally, the removal of high-degree nodes (hubs) can cause the redistribution of network traffic to other nodes, potentially leading to cascading failures if the remaining nodes are unable to handle the increased load.

The analysis of network resilience through node removal highlights the vulnerability of networks to both random failures and targeted attacks. The original network consistently demonstrates higher resilience compared to synthetic models, maintaining larger connected components longer. The clustering coefficient plays a crucial role in network robustness, and its consideration is essential for a comprehensive understanding of network resilience. The comparison of resilience across different models provides valuable insights into the structural properties that contribute to network robustness.

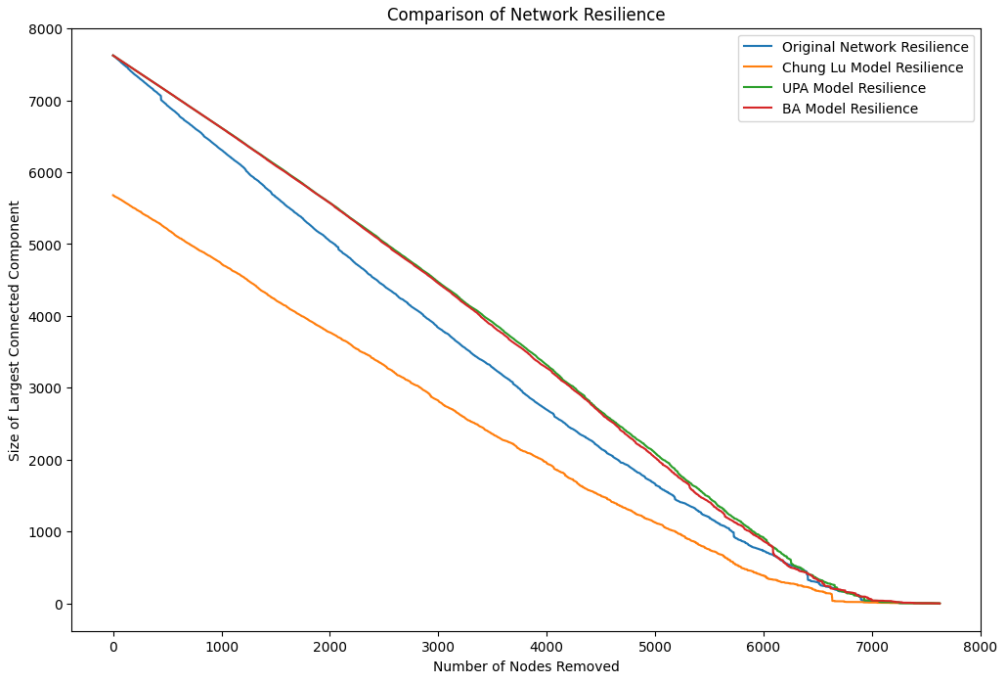

Fig 5.7 Node removal in LastFm Network

The plot shows the comparison of network resilience for the LastFM network, Chung Lu model, and UPA model under the condition of node removal.The observations and implications of the graph, and address the comment about fairness in comparison due to the initial sizes of the models.

The consistency in the starting points across all models now aligns with the expected behavior, indicating that the initial connectivity is well-matched. This consistency suggests that the original network and the synthetic models have similar structural properties at the outset, with differences emerging as nodes are removed. These results suggest that the Original Network, UPA, and BA models are more robust to node removal, while the Chung Lu model is less resilient in maintaining a large connected component.

The original network and UPA model maintain a larger connected component size for a longer period compared to the Chung Lu model, suggesting that both are more resilient to random node removal.

The UPA model shows higher resilience than the original network, maintaining a larger connected component size throughout the node removal process. This is particularly noticeable in the latter half of the plot. By the time 7000 nodes are removed, all three networks have a significantly reduced connected component size. The original network and UPA model retain a slightly larger component compared to the Chung Lu model, which almost reaches zero.

5.5 Community Resilience Testing

In this section, we delve into the community structure within the Email network, evaluating the resilience of these communities under various scenarios. Using the Louvain method, we identify the communities, which are chosen at random, and subsequently assess their resilience by adding nodes incrementally. The results are compared with those from synthetic models, specifically the Configuration and BA models, to better understand the robustness of the original network's community structure.

The Configuration Model is introduced in this analysis as a method for generating synthetic networks that preserve the degree distribution of the original network. Unlike other models, the Configuration Model maintains the exact degree sequence of the nodes, allowing for a more accurate representation of the network's connectivity. In this study, we utilize the Configuration Model exclusively to test the resilience of communities within the Email and LastFM networks. By comparing the resilience of these randomly selected communities to those generated by the Configuration Model, we can gain insights into the structural robustness of the original network's community formations.

Analysis of Community Resilience in the Email Network

The plot(Fig 5.8) illustrates the resilience of different communities in the Email network as nodes are sequentially added. Each line represents a community, with the x-axis showing the percentage of nodes added and the y-axis representing the size of the largest connected component in each community.

Community 7 (Brown Line): This community exhibits the highest robustness among the communities analyzed. As nodes are added, Community 7 consistently maintains the

largest connected component, suggesting it is highly resilient to changes. This robustness indicates a dense and well-connected internal structure, making it more resistant to fragmentation.

Community 1 (Orange Line): Following Community 7, Community 1 also shows significant resilience, maintaining a larger connected component than the other communities. The consistent growth in the size of the connected component as nodes are added indicates that this community is also well-connected but slightly less resilient compared to Community 7.

Communities 0, 4, 6, 5, 2 (Blue, Green, Purple, Red, and Pink Lines): These communities exhibit moderate resilience. They maintain connectivity, but their growth in the largest connected component is slower compared to Communities 7 and 1. This suggests a more fragmented structure or fewer connections within these communities, making them somewhat less robust.

Community 3 (Grey Line): This community appears to be the least robust, maintaining the smallest connected component as nodes are added. Its relatively low growth curve indicates a sparse or fragmented structure, making it more vulnerable to disruptions.

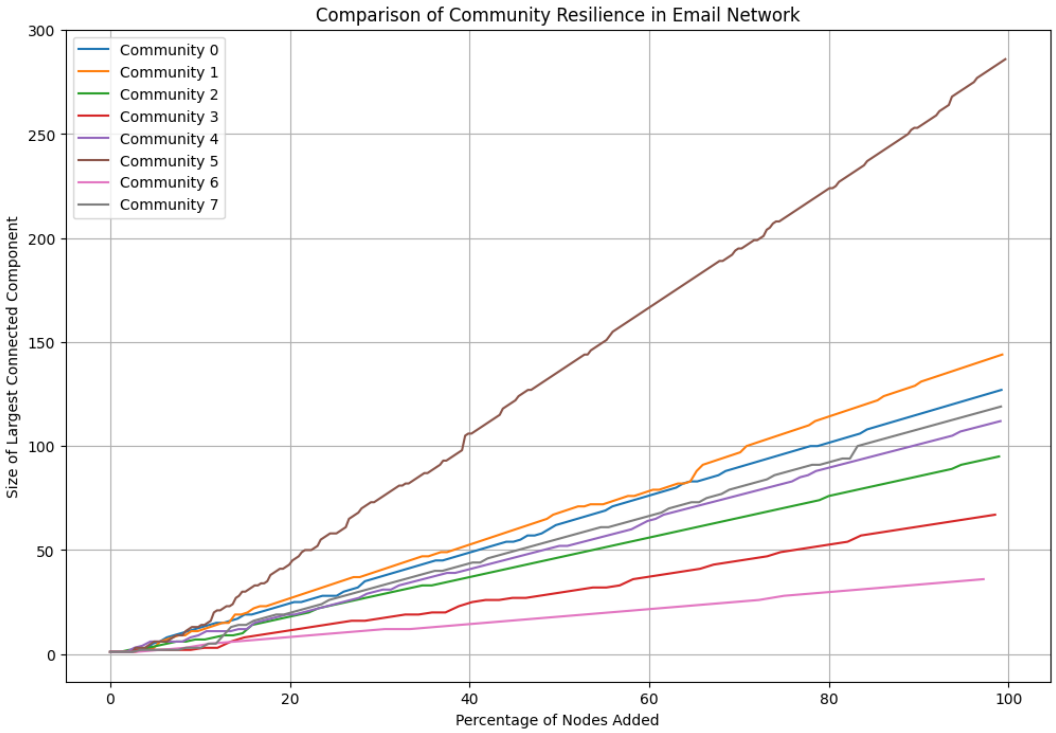

Fig 5.8 E-mail Network Community Resilience

Most Robust: Community 7 is the most robust, maintaining the largest connected component as nodes are added, indicating a highly connected and resilient structure.

Moderately Robust: Communities 1, 0, 4, 6, 5, and 2 show moderate resilience, with Community 1 leading among these.

Least Robust: Community 3 shows the least resilience, with a smaller connected component and slower growth, indicating vulnerability to node additions.

This analysis helps in understanding the structural integrity of different communities within the Email network. The more robust communities are likely to withstand changes better, while the less robust ones may require interventions to improve their resilience. This understanding is crucial for network design and management, particularly in ensuring the stability and reliability of communication networks.

Analysis of Community Resilience in the LastFM Network

The plot(Fig 5.9) provides a visual comparison of the resilience of different communities in the LastFM network as nodes are sequentially added. Each curve represents a distinct community, with the x-axis showing the percentage of nodes added and the y-axis depicting the size of the largest connected component in each community.

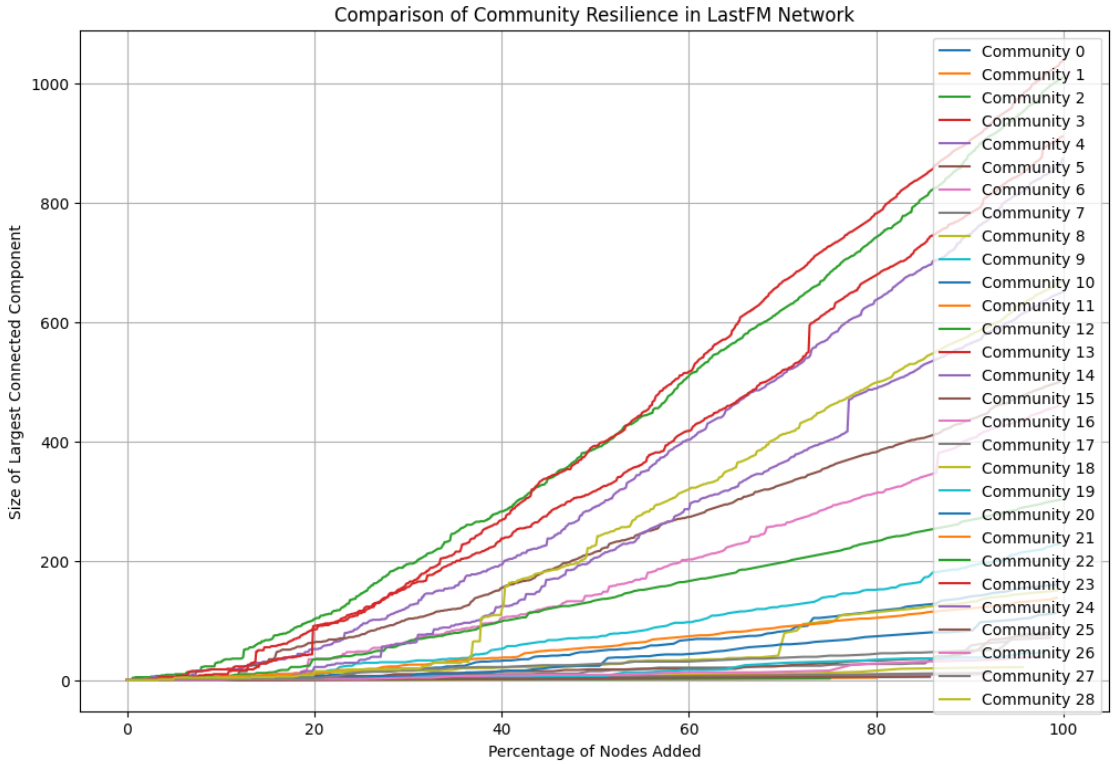

Fig 5.9 LastFm Network Community Resilience

Grouping and Discussion of Communities in the LastFM Network

1. **Highly Robust Communities**:

   Communities 5, 14, 17, 6, and 8:

   Characteristics: These communities demonstrate the highest resilience, with their largest connected components growing steadily and significantly as nodes are added. The curves for these communities are

steep, indicating that they maintain strong internal connectivity even as their size increases.

Implications: The high robustness of these communities suggests they are densely interconnected, making them less vulnerable to disruptions. In practical terms, these communities likely represent clusters of users with highly cohesive interaction patterns.

2. **Moderately Robust Communities**:

Communities 10, 12, 3, 4, 16, 19, and 11:

Characteristics: These communities show moderate resilience, with connected components growing steadily but at a slower rate compared to the highly robust group. Their curves are less steep, indicating a balanced level of internal connectivity.

Implications: These communities are fairly robust but may have some structural weaknesses or less dense interconnections. They can withstand some level of disruption but are more susceptible than the highly robust group. These communities might represent medium-sized clusters with average interaction intensity.

3. **Less Robust Communities**:

Communities 7, 9, 18, 20, 21, 13, and 15:

Characteristics: These communities show lower resilience, with their largest connected components growing more slowly and sometimes stalling as nodes are added. The flatter curves indicate a more fragile structure.

Implications: The lower robustness of these communities suggests they are less interconnected and more vulnerable to fragmentation. They might represent smaller or less active clusters, where the interaction patterns are not as cohesive.

4. **Least Robust Communities**:

Communities 22, 1, 0, 2, and 23:

Characteristics: These communities exhibit the least resilience, with minimal growth in their largest connected components as nodes are added. The almost flat curves indicate a highly vulnerable structure.

Implications: These communities are the most fragile and likely to fragment under even minor changes. They may represent isolated or inactive clusters with very sparse interactions, making them particularly susceptible to disruptions.

By grouping the communities based on their resilience, we can observe distinct patterns in the LastFM network. Highly robust communities, such as Communities 5 and 14, indicate areas of strong user interaction, while the least robust communities, like Communities 22 and 1, highlight vulnerabilities within the network. This analysis provides a clearer picture of the network's structure, allowing for targeted strategies to strengthen weaker areas and ensure overall stability.

6. Conclusion and Future Work

In this project, we conducted a comprehensive analysis of network resilience in the Email and LastFM datasets by evaluating the structural robustness of these networks under various scenarios of node addition and removal. By comparing the original networks with synthetic models such as the Chung Lu, UPA, and Barabási-Albert (BA) models, we gained valuable insights into how different network structures respond to changes and disruptions. The results highlighted the strengths and vulnerabilities of each network model in terms of their ability to maintain large connected components, a key indicator of resilience.

Addressing Fairness in Comparison:

Isolated Nodes in Chung Lu Model:

The Chung Lu model has fewer initially connected nodes, which may result in isolated nodes. This affects the overall resilience as these isolated nodes do not contribute to the network's robustness.

To ensure a fair comparison, it is important to consider models that do not inherently generate a significant number of isolated nodes.

Configuration Model:

The Configuration Model is a more appropriate comparator for the original network as it preserves the degree distribution while generating a random network.

This model can provide a more accurate comparison of resilience, as it avoids the issue of isolated nodes and maintains similar connectivity patterns to the original network.

Our findings indicate that the original networks, while robust, sometimes lag behind the synthetic models in maintaining connectivity during strategic node additions. The BA model, in particular, consistently demonstrated superior resilience, which can be attributed to its inherent scale-free properties and the formation of highly connected hubs. This model's resilience was especially pronounced in scenarios where nodes were added based on degree and betweenness centrality, underscoring the importance of hub nodes in preserving network structure.

While the Chung Lu model provides a reasonable approximation of the original network, it may oversimplify certain aspects of the network's structure, particularly regarding isolated nodes. As a future direction, the Configuration Model could be utilized instead, as it precisely preserves the degree distribution of the original network. This could potentially offer a more accurate reflection of the network's resilience.

The use of community detection through the Louvain method further provided a granular view of network resilience at the community level, revealing how subgroups within the network react to

structural changes. The results of these analyses can inform strategies for enhancing network robustness, particularly in applications where maintaining connectivity is critical.

Future Work

Building on the insights gained from this study, several avenues for future research and development can be explored:

Advanced Synthetic Models: While the Chung Lu, UPA, and BA models provided valuable comparisons, future work could explore other synthetic network models, such as the stochastic block model (SBM) or dynamic network models, to further understand network resilience under different structural assumptions.

Temporal Network Analysis: Extending the analysis to temporal networks, where interactions between nodes evolve over time, could provide deeper insights into the dynamics of network resilience. This would involve studying how resilience metrics change as the network grows or shrinks over time.

Community Detection Refinement: Future research could refine community detection methods to better capture the nuances of community structure and their impact on network resilience. Exploring alternative community detection algorithms or multi-scale community analysis could yield more detailed insights.

Real-World Application Testing: Applying the findings from this study to real-world networks, such as social networks, communication networks, or infrastructure networks, could validate the models' applicability and provide practical guidelines for improving network resilience in critical systems.

Machine Learning Integration: Incorporating machine learning techniques to predict resilience outcomes based on network features could offer a more automated and scalable approach to resilience analysis. Such models could be trained on a variety of networks to predict their resilience under different scenarios, providing valuable tools for network design and maintenance.

Resilience Optimization: Investigating strategies to optimize network resilience, such as targeted interventions on critical nodes or edges, could lead to more effective methods for reinforcing network structures. This could involve developing algorithms that automatically identify and fortify the most vulnerable parts of a network.

By exploring these future directions, we can enhance our understanding of network resilience and develop more robust systems capable of withstanding various forms of disruptions. This ongoing research is crucial as networks continue to play a vital role in numerous aspects of modern society, from communication and transportation to energy and security.

7. APPENDIX A: CODEBOOK FOR NETWORK RESILIENCE AND COMMUNITY ANALYSIS

Code	Description	Definition	Example
Node Addition (Random)	The process of randomly adding nodes to the network to test resilience.	Nodes are added without any specific order, and the size of the largest connected component is observed.	"Adding nodes randomly to see how well the network maintains connectivity."
Node Addition (Degree Centrality)	Adding nodes to the network based on their degree centrality.	Nodes with the highest degree centrality are added first, aiming to test the network's robustness.	Nodes with the highest degree centrality are added first, aiming to test the network's robustness.
Node Removal (Random)	The process of randomly removing nodes from the network.	Nodes are removed without any specific order, and the impact on the largest connected component is observed.	"Randomly removing nodes to assess the network's vulnerability."
Node Removal (Betweenness Centrality)	Removing nodes based on their betweenness centrality.	Nodes with the highest betweenness centrality are removed first to test the impact on the network's connectivity.	"Removing nodes that act as bridges within the network to see how the network fragments."
Community Detection (Louvain Method)	A method used to detect communities within the network.	The Louvain method partitions the network into communities by optimizing modularity.	"Using the Louvain method to identify closely-knit groups within the network."

Network Modularity	A measure of the structure of networks or graphs.	High modularity indicates strong community structure where nodes within a community are densely connected.	"Calculating modularity to determine the strength of community structures within the network."
Synthetic Models (Chung Lu, UPA, BA)	Models used to replicate real-world networks.	Chung Lu generates random graphs with a given degree distribution, UPA mimics growth dynamics, and BA generates scale-free networks.	"Comparing network resilience between real-world data and synthetic models."
Community Resilience	The robustness of communities within the network.	Measuring how communities withstand node addition or removal, focusing on the largest connected component.	Assessing the resilience of detected communities when nodes are added or removed.

Table 7.1 Codebook for network resilience and community analysis

8. APPENDIX – B | GITHUB LINK

8.1 Git Link

https://github.com/adishdmc/Graph_Resilience_Analysis/tree/main

8.2 Sample File

https://github.com/adishdmc/Graph_Resilience_Analysis/blob/main/Email/Data/chung_lu_model.gml

https://github.com/adishdmc/Graph_Resilience_Analysis/blob/main/lasftm_asia/data/ba_model_lastfm.gml

9. APPENDIX – C | LIST OF FIELDS IN THE DATASET

email-Eu-core.txt:

node_1: Identifier for the first node in an email communication pair. Represents one of the participants in the email exchange.

node_2: Identifier for the second node in an email communication pair. Represents the other participant in the email exchange.

email-Eu-core-department-labels.txt:

node: Identifier for a node in the email network. Each node corresponds to an individual in the organization.

department: Label representing the department to which the individual (node) belongs. This field is used to analyze communication patterns across different departments.

lastfm_asia_edges.csv:

user1: Identifier for the first user in an interaction pair on the LastFM platform. Represents one of the users who have interacted.

user2: Identifier for the second user in an interaction pair on the LastFM platform. Represents the other user involved in the interaction.

lastfm_asia_target.csv:

user: Identifier for a user on the LastFM platform. Each user is associated with their interaction data in the LastFM dataset.

label: Category label associated with the user, representing their group or community based on music preferences or social connections.

email_graph.gml (Created File):

nodes: Nodes in the email network, each representing an individual.

edges: Edges between nodes, representing email communication between individuals.

node_attributes:

department: Department label for each node, indicating which department the individual belongs to.

email_labels.json (Created File):

node: Identifier for each node in the email network.

label: Department label or community assignment for each node, used for analyzing community structure.

lastfm_graph.gml (Created File):

nodes: Nodes in the LastFM network, representing users on the platform.

edges: Edges between nodes, representing interactions between users.

node_attributes:

label: Community or group label for each node, indicating the user's group based on music preferences or social connections.

10. reference

1. **R. Albert, H. Jeong, and A.-L. Barabási,** "Error and attack tolerance of complex networks," *Nature*, vol. 406, pp. 378-382, 2000.
2. **A. Medina, I. Matta, and J. Byers,** "On the origin of power laws in Internet topologies," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 2, pp. 18-28, Apr. 2000.
3. **R. M. Dunne and P. Crossley,** "Topological resilience analysis of supply networks under random disruptions and targeted attacks," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 6, pp. 2039-2048, Dec. 2016.
4. **A. Goswami, R. Shokri-Ghasabeh, and J. Bookhamer,** "Comprehensive comparison and accuracy of graph metrics in predicting network resilience," *IEEE Systems Journal*, vol. 10, no. 4, pp. 1238-1247, Dec. 2016.
5. **R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin,** "Resilience of the Internet to random breakdowns," *Physical Review E*, vol. 66, no. 3, p. 036113, Sept. 2002.
6. **R. Albert, H. Jeong, and A.-L. Barabási,** "Attack vulnerability of complex networks," *Physical Review E*, vol. 65, no. 5, p. 056109, May 2002.
7. **M. E. J. Newman,** "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167-256, Jan. 2003.
8. **A. E. Motter and Y.-C. Lai,** "Resilience of complex networks to random breakdowns and targeted attacks," *Physical Review E*, vol. 70, no. 5, p. 056107, Nov. 2004.
9. **Y. Moreno, R. Pastor-Satorras, A. Vázquez, and A. Vespignani,** "Cascade-based attacks on complex networks," *Physical Review E*, vol. 66, no. 6, p. 065102, Dec. 2002.
10. **S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin,** "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, pp. 1025-1028, Apr. 2010.
11. **G. D'Agostino and A. Scala,** "Robustness of a network of networks," *Physical Review E*, vol. 85, no. 6, p. 066134, Jun. 2012.
12. **P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda,** "The power grid as a complex network: A survey," *Physical Review E*, vol. 87, no. 6, p. 062809, Jun. 2013.

13. **B. Shahrivar, G. Thakur, and M. E. Crovella,** "Analysis and improvement of network resilience using topological metrics," *Computer Networks*, vol. 155, pp. 89-101, May 2019.
14. **A. V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes,** "Network robustness and fragility: Percolation on random graphs," *The European Physical Journal B*, vol. 33, no. 2, pp. 265-274, May 2003.
15. **F. Ghanavati, M. Hosseinpour, and A. Kazemi,** "Popularity prediction of Reddit texts," *arXiv preprint arXiv:1609.08347*, Sept. 2016.