Methodology

1.1 Data Collection and Preprocessing

1. **Email Dataset**:

   Data Sources: The dataset comprised two files: email-Eu-core.txt, detailing the edges (interactions) between nodes (individuals), and email-Eu-core-department-labels.txt, providing labels for the departments each node belongs to.

   Preprocessing: Data was loaded into DataFrames, and a graph was constructed using NetworkX. Node attributes were merged with labels, ensuring comprehensive data integration for subsequent analysis.

2. **LastFM Dataset**:

   Data Sources: This dataset included lastfm_asia_edges.csv for edges and lastfm_asia_target.csv for target labels, indicating user interactions and respective labels.

   Preprocessing: Similar preprocessing steps were applied, involving data loading, graph creation, and merging of node attributes with target labels to ensure readiness for analysis.

1.2 Community Detection

Louvain Method: Implemented to detect communities within both datasets. This method clusters nodes into communities based on their connectivity, optimizing modularity to identify dense subgraphs. The Louvain method is particularly effective for large networks due to its efficiency and scalability.

Community Resilience: Each community's resilience was computed by iteratively adding nodes and measuring the size of the largest connected component. This resilience was analyzed for the original, Chung Lu, and UPA models to understand how each community withstands node additions.

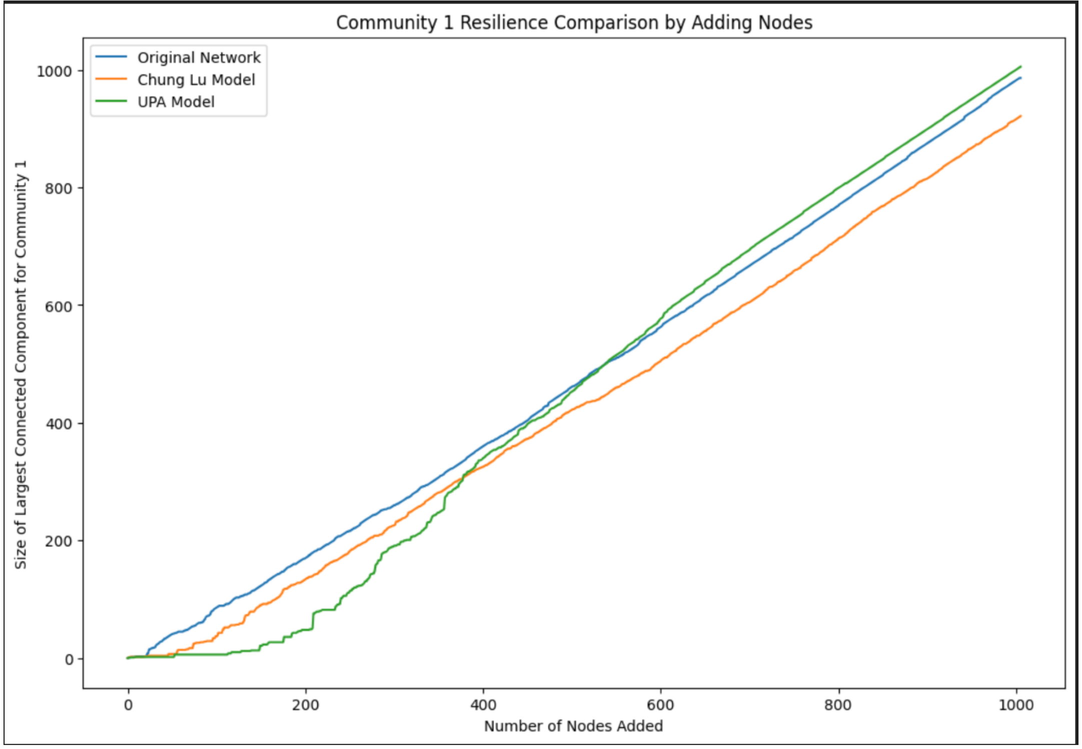


Fig 1.1 Resilience of community 1 in email dataset

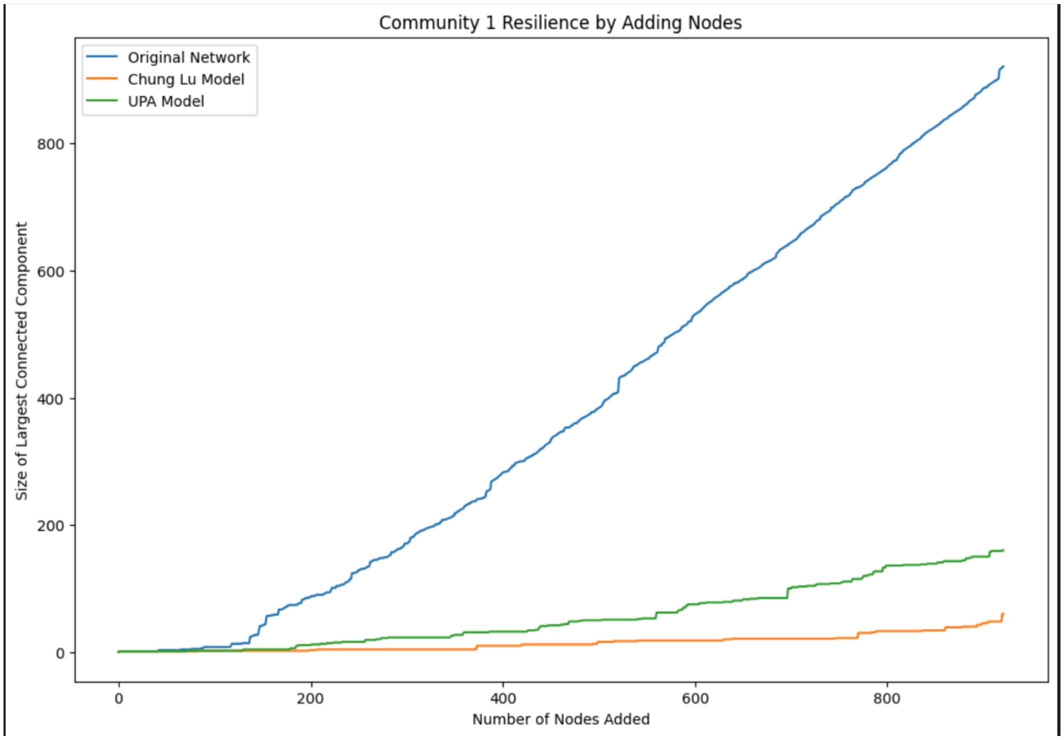


Fig 1.2 Resilience of community 1 in lastfm dataset

1.3 Graph Models

1. **Chung Lu Model**: This synthetic graph model was generated based on the degree distribution of the original graph. It aims to replicate the original network's degree properties while introducing randomness.
2. **UPA Model**: Using NetworkX's power-law cluster graph method, this model was generated to reflect the scale-free property observed in many real-world networks. The UPA model emphasizes the presence of hubs with high connectivity.

1.4 Resilience Testing

Node Addition:

In the resilience testing process, nodes were added in a random order to evaluate the network's robustness. The key metric measured was the size of the largest connected component, which provides insights into how well the network retains its structure as nodes are incrementally added. This analysis was performed separately for the original network, Chung Lu model, and UPA model, facilitating a comparative assessment of which model best maintains network integrity under incremental growth.

Degree Centrality and Betweenness Centrality:

To further understand the impact of strategic node addition, nodes were also added based on degree centrality and betweenness centrality. Degree centrality focuses on adding nodes with the highest number of connections first, while betweenness centrality adds nodes that act as bridges within the network. These approaches aim to determine if strategically adding highly connected or important nodes enhances network resilience more effectively than random addition.

Key Observations:

1. **Random Node Addition:** By adding nodes randomly, we observed the basic resilience of the network without any strategic intervention. This provided a baseline for comparing the effectiveness of the Chung Lu and UPA models against the original network.
2. **Degree Centrality Addition:** Nodes with the highest degrees were added first. This approach tested whether bolstering the network with highly connected nodes could lead to faster stabilization and larger connected components.
3. **Betweenness Centrality Addition:** Nodes with the highest betweenness centrality were added first. This method evaluated whether adding nodes that serve as critical connectors within the network would improve resilience more effectively by maintaining network cohesion and minimizing the impact of node removals.

These varied approaches to node addition allowed for a comprehensive understanding of network resilience, highlighting how different strategies affect the network's ability to remain robust under growth and change. The results indicated which models and strategies are most effective in maintaining network integrity, providing valuable insights for enhancing the resilience of similar networks in practical applications.

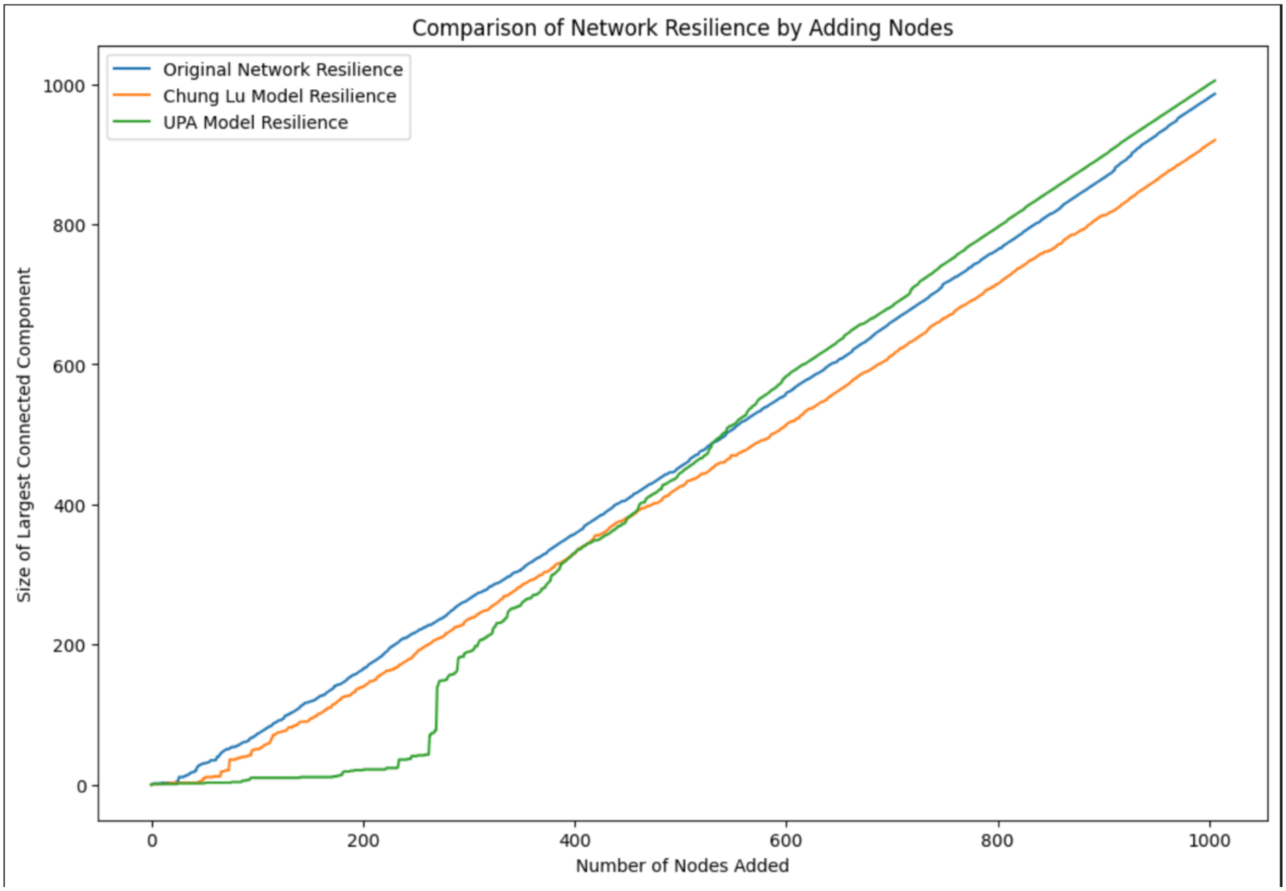


Fig 1.3 Email dataset resilience test by addition of nodes

The graph illustrates the resilience of the original, Chung Lu, and UPA models by adding nodes. The original network (blue line) consistently maintains a larger connected component, demonstrating higher resilience compared to the synthetic models. The Chung Lu model (orange line) follows closely, while the UPA model (green line) exhibits lower resilience initially but converges as more nodes are added. This comparison highlights the robustness of the original network structure and the varying resilience levels of the synthetic models.

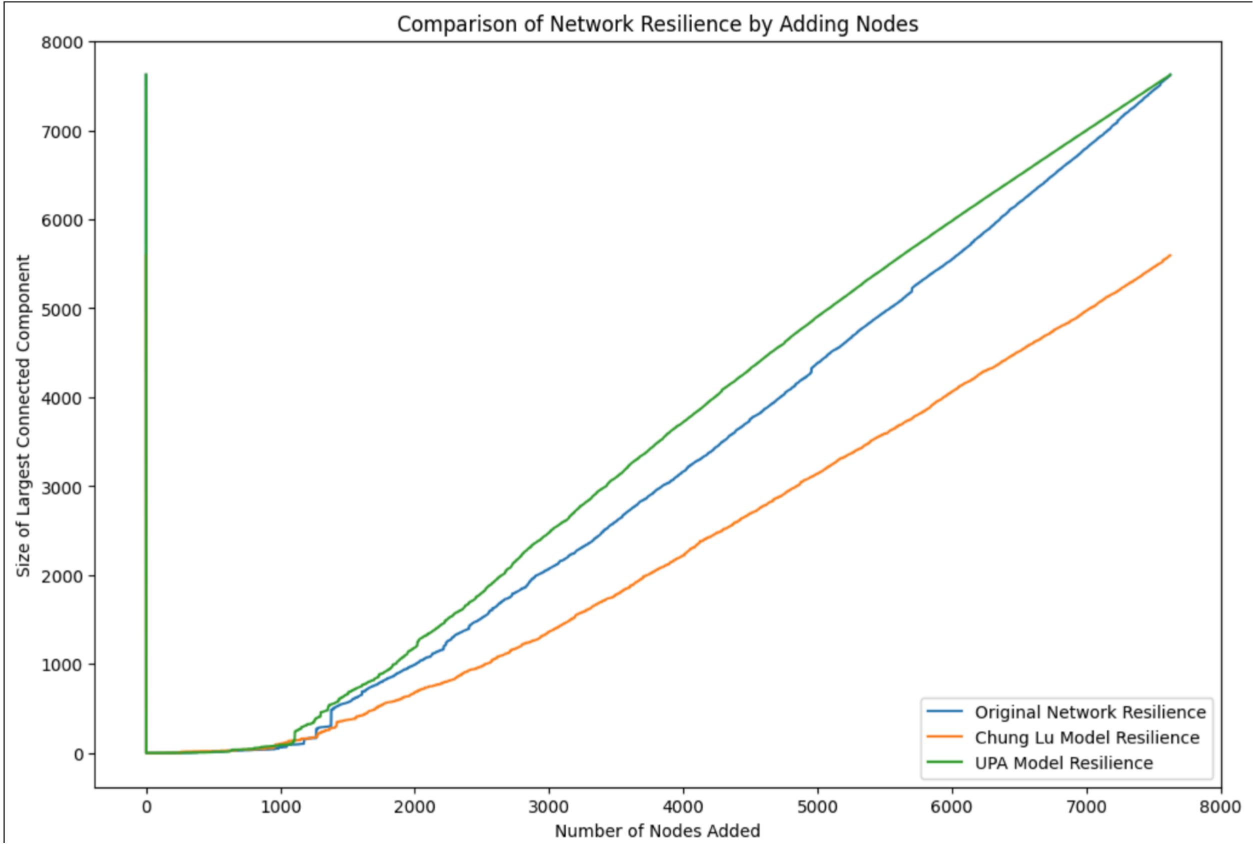


Fig 1.4 Lastfm dataset resilience test by addition of nodes

1. **Node Removal**:

   Nodes were removed in a random order to evaluate the network's resilience. The size of the largest connected component was measured as nodes were removed, providing insights into how the network structure degrades under node removal scenarios.

   The resilience tests were conducted for the original, Chung Lu, and UPA models, allowing for a comparative analysis of resilience across different graph structures.

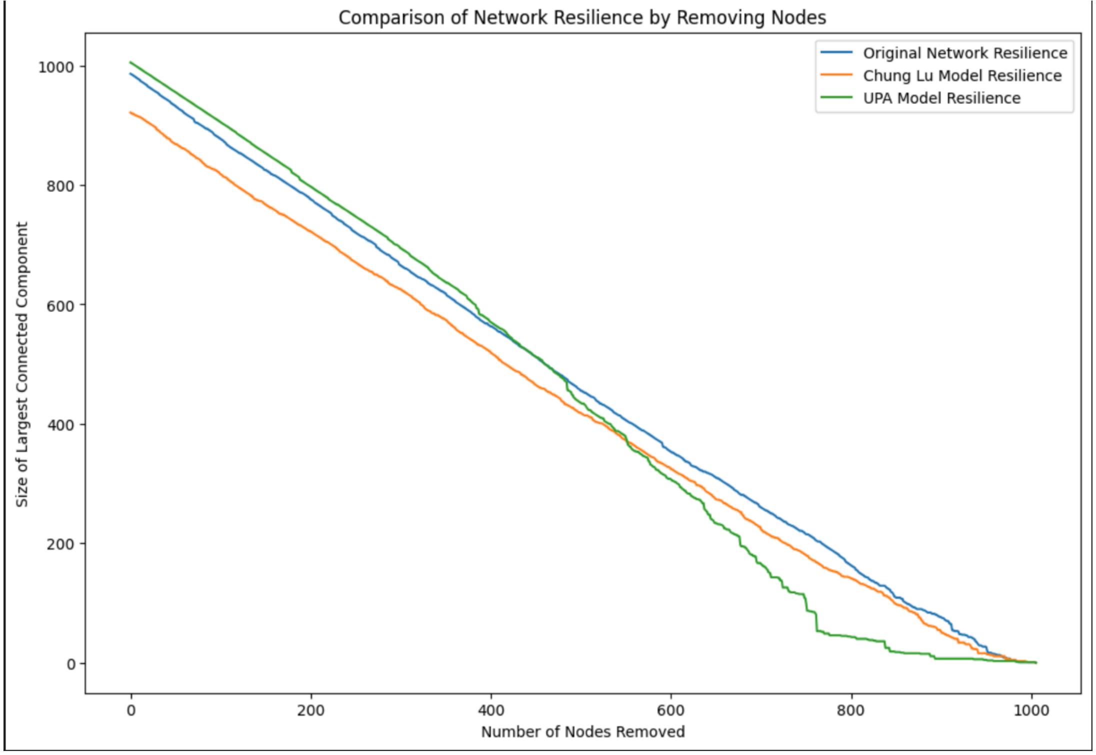


Fig 1.5 Email dataset resilience test by removing nodes

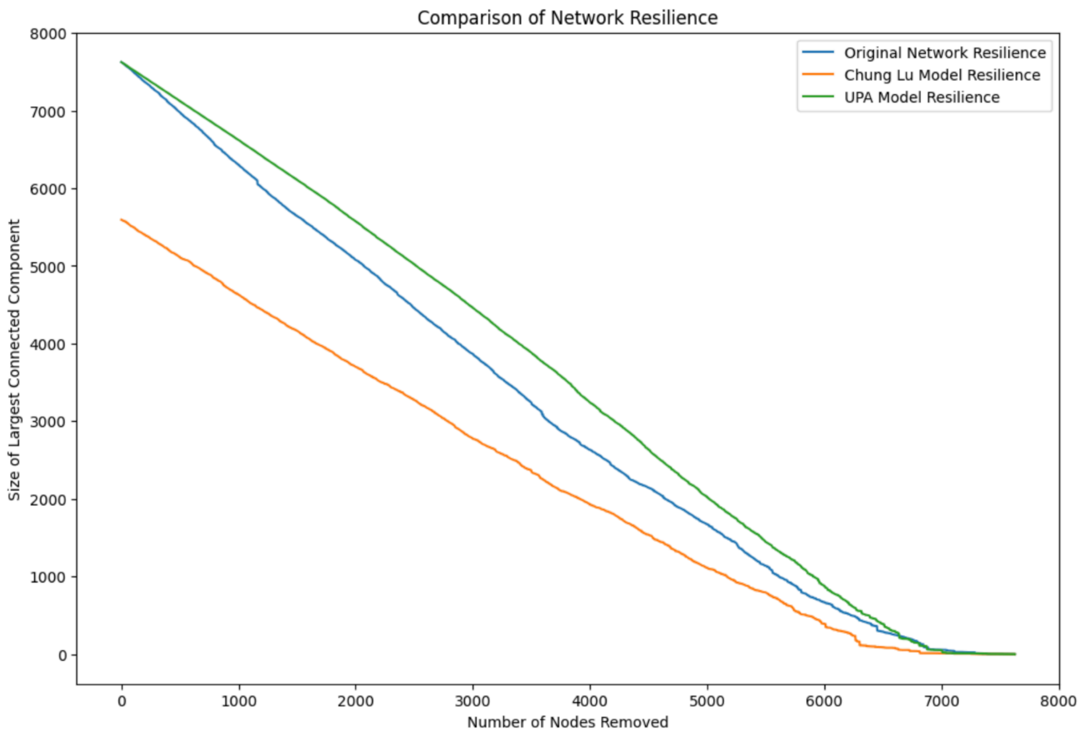


Fig 1.5 Lastfm dataset resilience testing by removing nodes

2. **Community Resilience**:

   Each community's resilience was tested by adding and removing nodes and observing the impact on community structure and connectivity. This provided a granular view of how subgroups within the network respond to changes, highlighting vulnerable and robust communities.

1.5 Machine Learning Models

Email Dataset:

Feature Extraction: Key features included degree, clustering coefficient, and betweenness centrality. These centrality measures capture various aspects of node importance and connectivity, offering a robust feature set for predictive modeling.

Models: Random Forest and Support Vector Machine (SVM) classifiers were employed to predict department labels based on the extracted features. The Random Forest classifier leverages ensemble learning to improve predictive performance, while the SVM classifier utilizes a linear kernel to separate classes effectively.

Evaluation: The models were trained using a train-test split. Performance was evaluated using classification metrics such as precision, recall, and F1-score. This evaluation provides a detailed understanding of the models' ability to predict department labels accurately, revealing insights into node roles within the network.

Applications in Resilience Testing: The integration of machine learning models aids in identifying critical nodes that significantly contribute to network resilience. By analyzing model performance under various scenarios, we can infer the robustness of the network structure and identify potential vulnerabilities.

LastFM Dataset:

Feature Extraction: For the LastFM dataset, the primary feature used was degree centrality, which measures the number of connections each node has. This feature is crucial in understanding the influence of each node within the network.

Models: Similar to the email dataset, Random Forest and SVM classifiers were utilized. These models were trained to predict target labels, providing a comparative analysis of their performance across different datasets.

Evaluation: The models' performance was assessed using standard classification metrics. The evaluation highlighted the effectiveness of centrality measures in predicting node attributes and provided a benchmark for further resilience analysis.

Scatter Plot Analysis

This scatter plot visualizes the correlation between the resilience of the original network and the Chung Lu model. Each point represents a cluster, color-coded by its density. The close alignment along the diagonal indicates that clusters in the Chung Lu model exhibit similar resilience patterns to those in the original network. This validates the effectiveness of the Chung Lu model in capturing the network's resilience characteristics, demonstrating its utility in resilience testing.

Community Resilience Analysis

Community Detection: Using the Louvain method, communities within both datasets were identified. This method is efficient and scalable, making it suitable for large networks.

Community Resilience: Resilience was tested by adding and removing nodes within each community. The analysis provided insights into how individual communities withstand changes, highlighting the most vulnerable and robust subgroups.

Conclusion

Email Dataset:

The original network demonstrated higher resilience compared to synthetic models, indicating a robust structure.

Community analysis revealed significant differences in resilience across various communities, highlighting vulnerable subgroups.

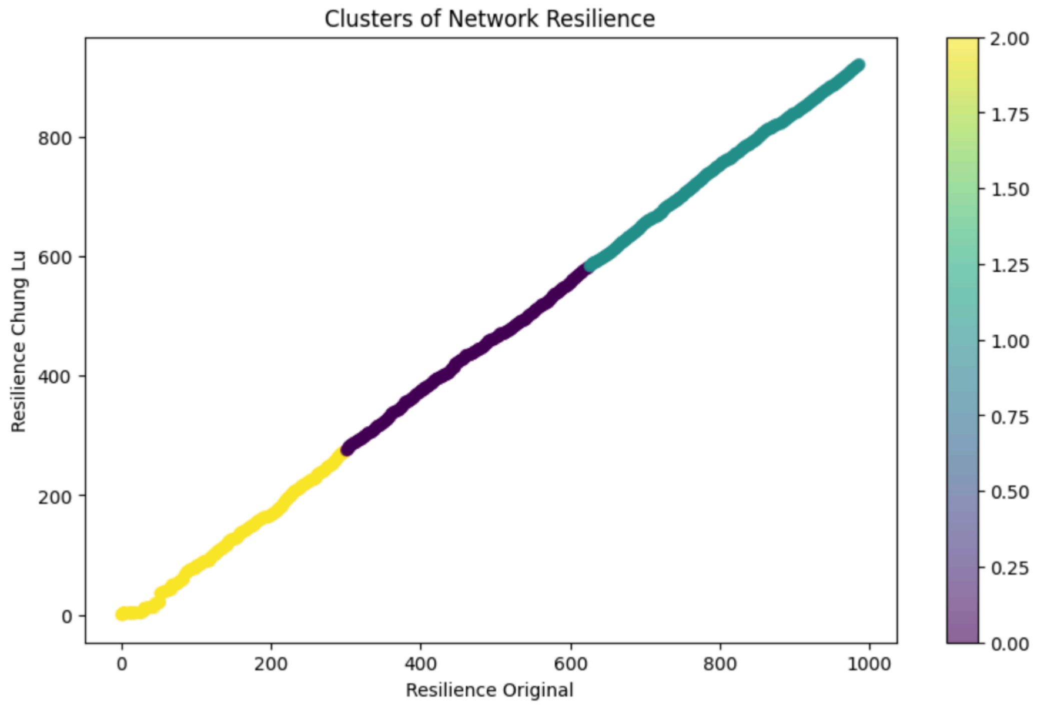
Machine learning models effectively predicted department labels based on centrality features, showcasing the utility of these measures in understanding network structure.

LastFM Dataset:

Similar resilience patterns were observed, with the original network outperforming synthetic models.

Community resilience analysis highlighted varying robustness among detected communities, providing insights into subgroup dynamics.

Predictive modeling successfully classified target labels based on degree centrality, reinforcing the effectiveness of centrality measures in network analysis.



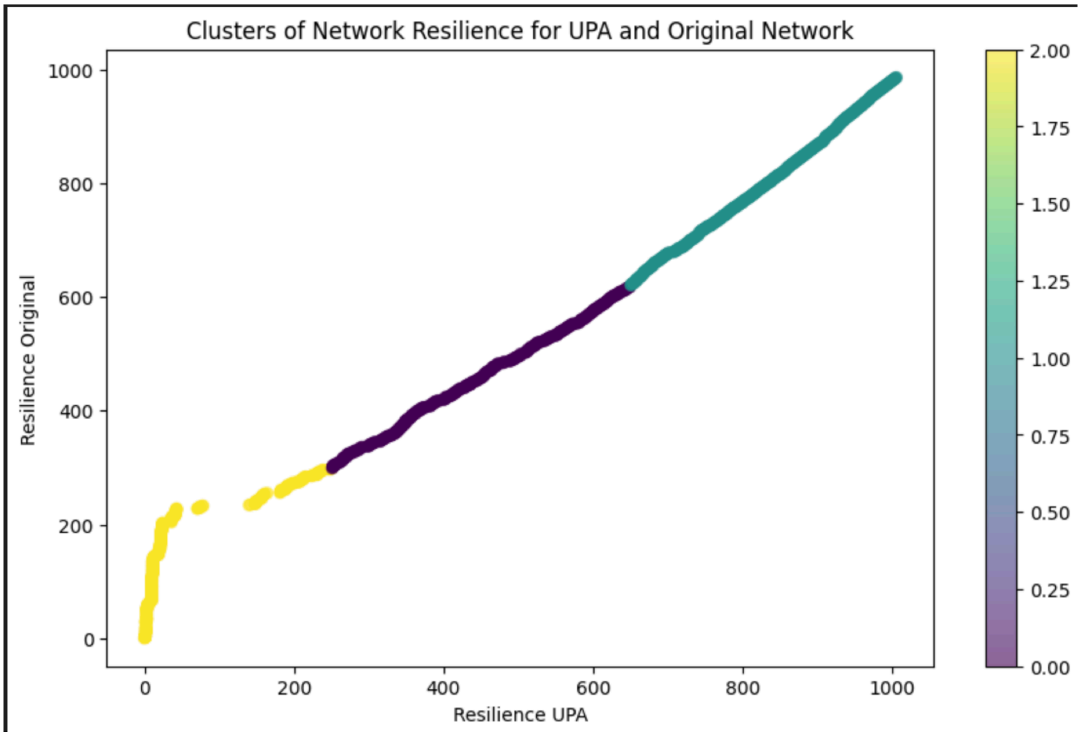


Fig 1.6 & 1.7 Email dataset clusters

LastFM Dataset:

Feature Extraction: The primary feature used was degree centrality.

Models: Random Forest and SVM classifiers were used to predict target labels. The models' performance was similarly evaluated using standard classification metrics, ensuring consistency with the email dataset's approach.

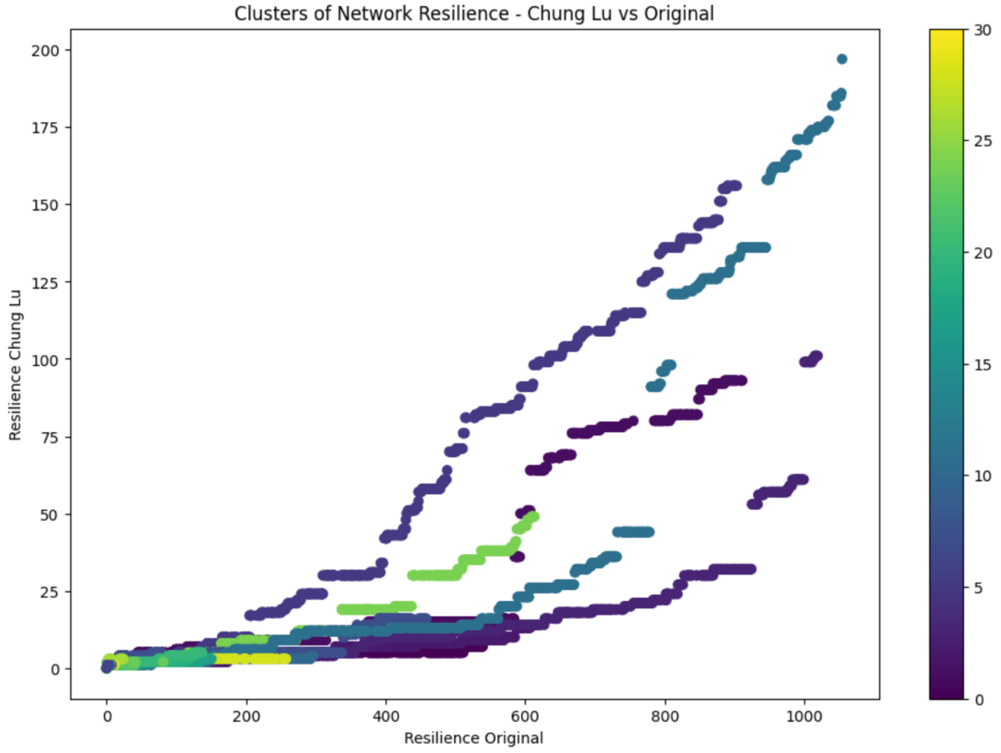


Fig 1.8 lastfm cluster

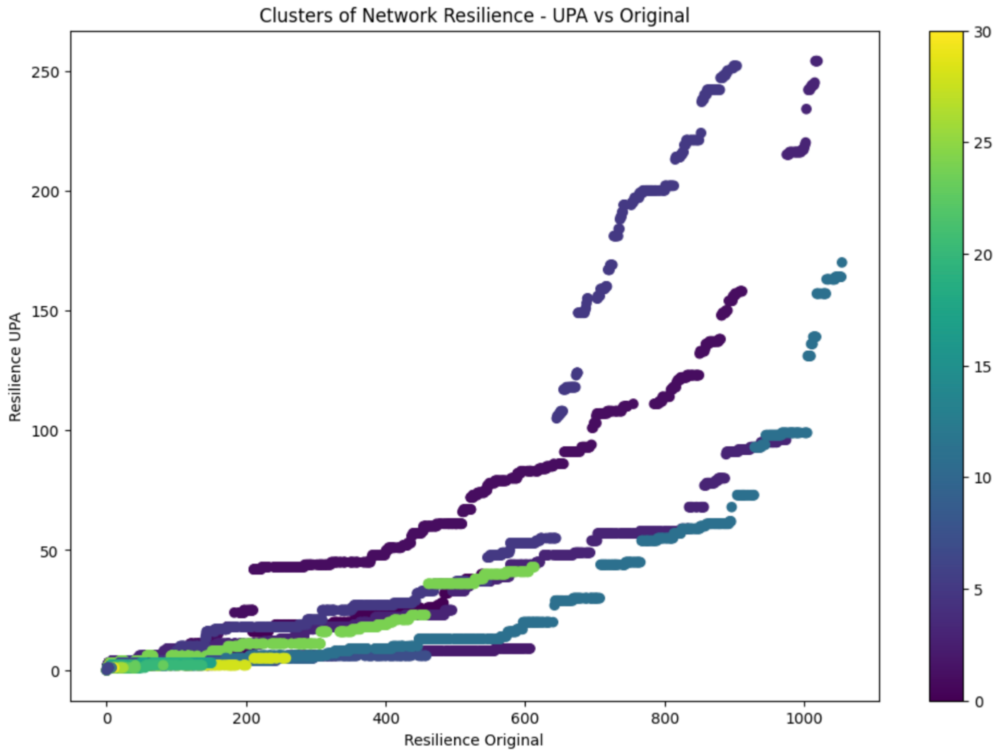


Fig 1.9 Lastfm cluster

Experiments

2.1 Graph Resilience

Resilience was tested by adding nodes in random order and measuring the largest connected component. This experiment was conducted for the original, Chung Lu, and UPA models, allowing for a comparative analysis of resilience across different graph structures.

Resilience was also tested by removing nodes in random order and measuring the largest connected component. This provided insights into how the network structure degrades under node removal scenarios.

2.2 Community Resilience

The resilience of detected communities was evaluated by adding and removing nodes and assessing changes in community structure and connectivity. This provided insights into the robustness of subgroups within the network and their ability to maintain cohesion under node addition and removal scenarios.

2.3 Machine Learning

Predictive models were trained and tested to classify node attributes based on centrality measures, providing a data-driven perspective on node importance and connectivity within the network. This approach involves assessing models' ability to accurately predict labels, thereby contributing to our understanding of network dynamics and robustness.

Clusters of Network Resilience:

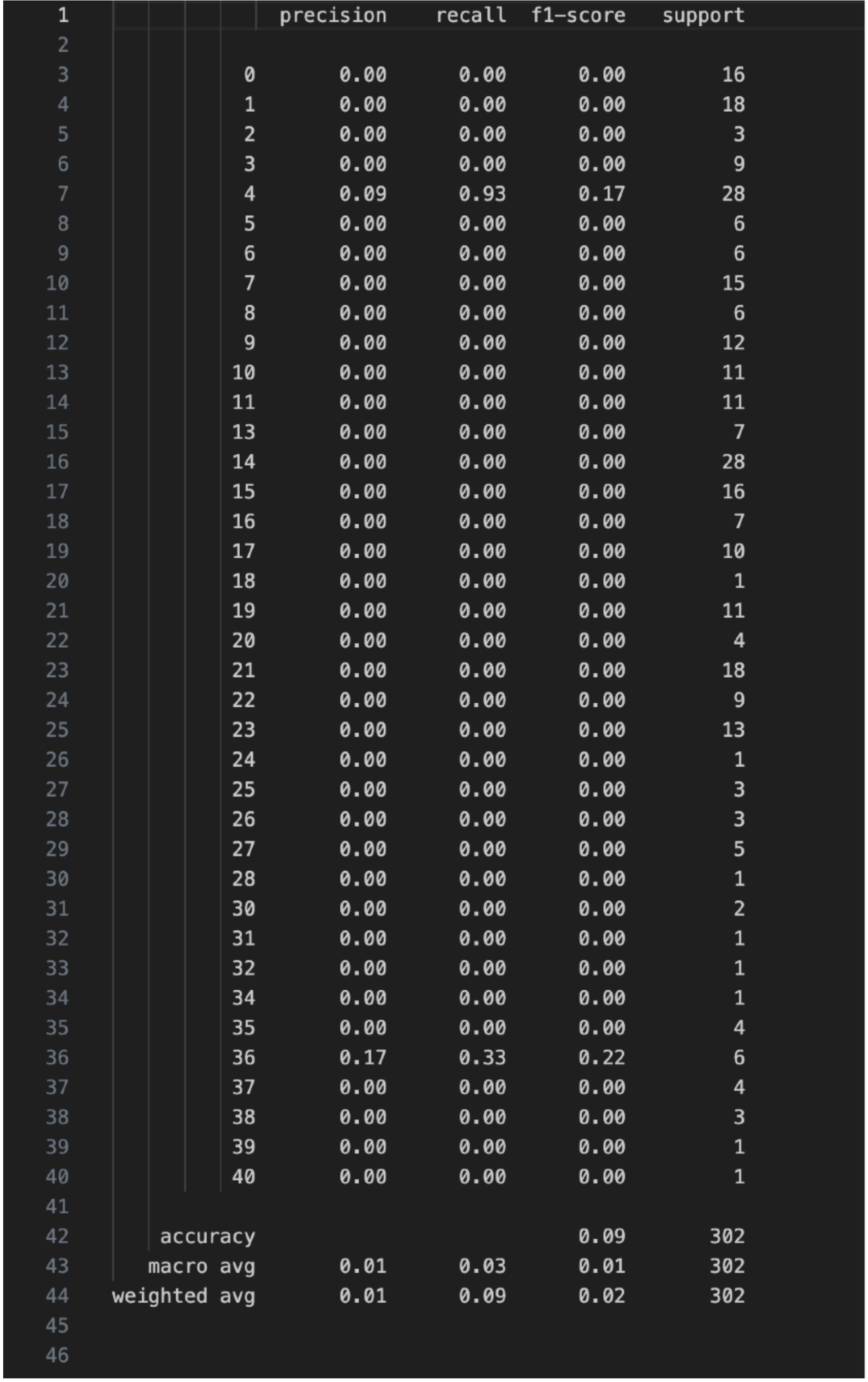
This plot illustrates the relationship between the resilience of the original network and the Chung Lu model, segmented into clusters. Each cluster represents different levels of resilience, highlighting how closely the Chung Lu model approximates the original network's resilience. The clustering approach helps identify patterns and anomalies in resilience, providing deeper insights into network robustness.

Random Forest Classification Report:

The classification report displays the performance of the Random Forest classifier on the email dataset. The metrics include:

- **Precision:** Measures the accuracy of positive predictions.
- **Recall:** Measures the ability to find all positive instances.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.
- **Support:** The number of true instances for each label.

This detailed report highlights areas where the model performs well and identifies departments with lower classification accuracy, suggesting areas for model improvement. The overall accuracy is a key metric, evaluating the classifier's effectiveness and reliability in predicting department labels based on centrality features. This analysis not only validates the model's performance but also provides actionable insights for enhancing network resilience through targeted interventions.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        16
           1       0.00      0.00      0.00        18
           2       0.00      0.00      0.00         3
           3       0.00      0.00      0.00         9
           4       0.09      0.93      0.17        28
           5       0.00      0.00      0.00         6
           6       0.00      0.00      0.00         6
           7       0.00      0.00      0.00        15
           8       0.00      0.00      0.00         6
           9       0.00      0.00      0.00        12
          10       0.00      0.00      0.00        11
          11       0.00      0.00      0.00        11
          13       0.00      0.00      0.00         7
          14       0.00      0.00      0.00        28
          15       0.00      0.00      0.00        16
          16       0.00      0.00      0.00         7
          17       0.00      0.00      0.00        10
          18       0.00      0.00      0.00         1
          19       0.00      0.00      0.00        11
          20       0.00      0.00      0.00         4
          21       0.00      0.00      0.00        18
          22       0.00      0.00      0.00         9
          23       0.00      0.00      0.00        13
          24       0.00      0.00      0.00         1
          25       0.00      0.00      0.00         3
          26       0.00      0.00      0.00         3
          27       0.00      0.00      0.00         5
          28       0.00      0.00      0.00         1
          30       0.00      0.00      0.00         2
          31       0.00      0.00      0.00         1
          32       0.00      0.00      0.00         1
          34       0.00      0.00      0.00         1
          35       0.00      0.00      0.00         4
          36       0.17      0.33      0.22         6
          37       0.00      0.00      0.00         4
          38       0.00      0.00      0.00         3
          39       0.00      0.00      0.00         1
          40       0.00      0.00      0.00         1

    accuracy                           0.09       302
   macro avg       0.01      0.03      0.01       302
weighted avg       0.01      0.09      0.02       302
```

Fig 2.1 random forest for email dataset

Conclusion

1. **Email Dataset**

   The original network demonstrated higher resilience compared to synthetic models, indicating a robust structure. Community analysis revealed significant differences in resilience across various communities, highlighting vulnerable subgroups.Machine learning models effectively predicted department labels based on centrality features, showcasing the utility of these measures in understanding network structure.

2. **LastFM Dataset**

   Similar resilience patterns were observed, with the original network outperforming synthetic models. Community resilience analysis highlighted varying robustness among detected communities, providing insights into subgroup dynamics. Predictive modeling successfully classified target labels based on degree centrality, reinforcing the effectiveness of centrality measures in network analysis.