Network Analysis and Simulation

Domenico Sosta, Tindaro Catalfamo

Year 2024 - 2025

Chapter 1

Phase 1: Data Analysis

1.1 Dataset Description

The dataset `bio-yeast.mtx` represents a network of biological interactions in yeast (*Saccharomyces cerevisiae*).

1.1.1 Network Dimensions:

- **Nodes:** 1458
- **Edges:** 1948

1.1.2 Network Characteristics:

- **Undirected:** The interactions are reciprocal.
- **Unweighted:** Each edge represents the presence of an interaction without an associated numerical value (weight).
- **Density:** 0.00183, suggesting a very sparse network.
- **Average Degree:** 2.67 (each node is, on average, connected to about 2-3 other nodes).

1.2 Network Analysis Results

1.2.1 Degree Distribution

The degree distribution of the network shows a typical scale-free structure. The distribution follows a power law, indicating the presence of hubs with a high number of connections and many nodes with few connections. The log-log scale distribution graph confirms this trend, a common characteristic in biological networks.
Power-law fitting results show:

- Exponent (alpha): 3.51
- Minimum degree (xmin): 5.0
- Goodness-of-fit test: R = 36.93, p-value = 5.2×10^{-5}.

The low p-value suggests the network significantly fits a power-law distribution, supporting the hypothesis of a scale-free network.

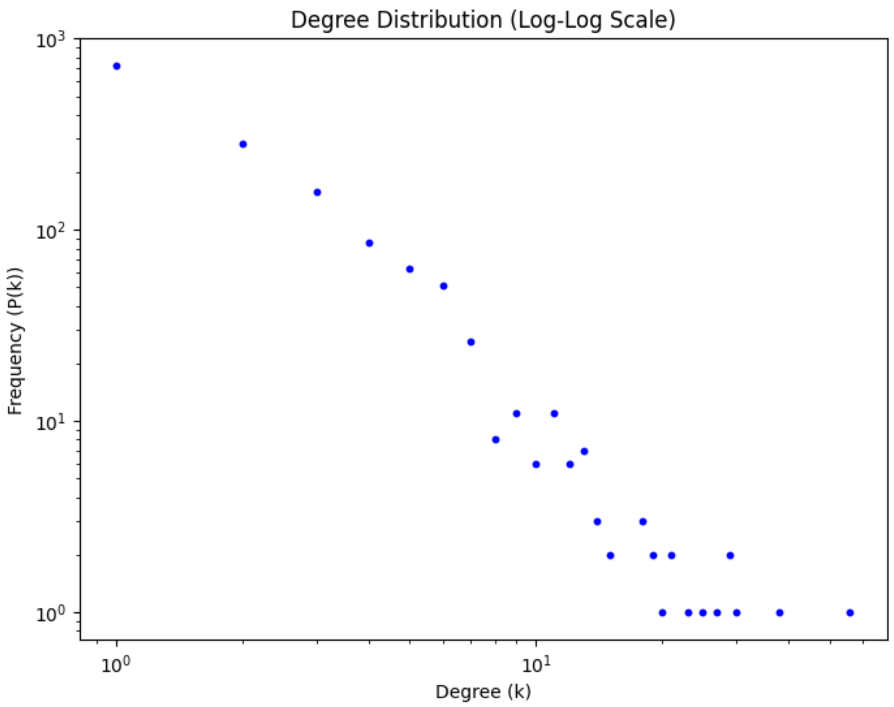


Figure 1.1: Degree distribution of the network in log-log scale.

1.2.2 Clustering Coefficient

The average clustering coefficient calculated for the network is $C = 0.0708$. This value indicates a moderate tendency of nodes to form triangles, suggesting some local cohesion among the network's nodes. However, the relatively low value reflects a globally less dense and more sparse structure.

1.2.3 Average Path Length

The average path length between any two nodes in the network is 6.81. This value confirms that the network exhibits a 'small-world' structure, where the average distance between nodes is relatively short despite the graph's size.

1.2.4 Community

Edge Betweenness Centrality
The analysis of edge betweenness centrality highlighted that the edges with the highest values play a crucial role in connecting different communities within the network. For instance, the edge with the highest value has a centrality of 0.0564, followed by others with slightly lower values.

Community Detection
The community detection algorithm based on edge betweenness (Girvan-Newman) identified several well-defined communities. The optimized modularity for these communities reached a significant value, indicating a structured and coherent subdivision of the network into internally interconnected groups.

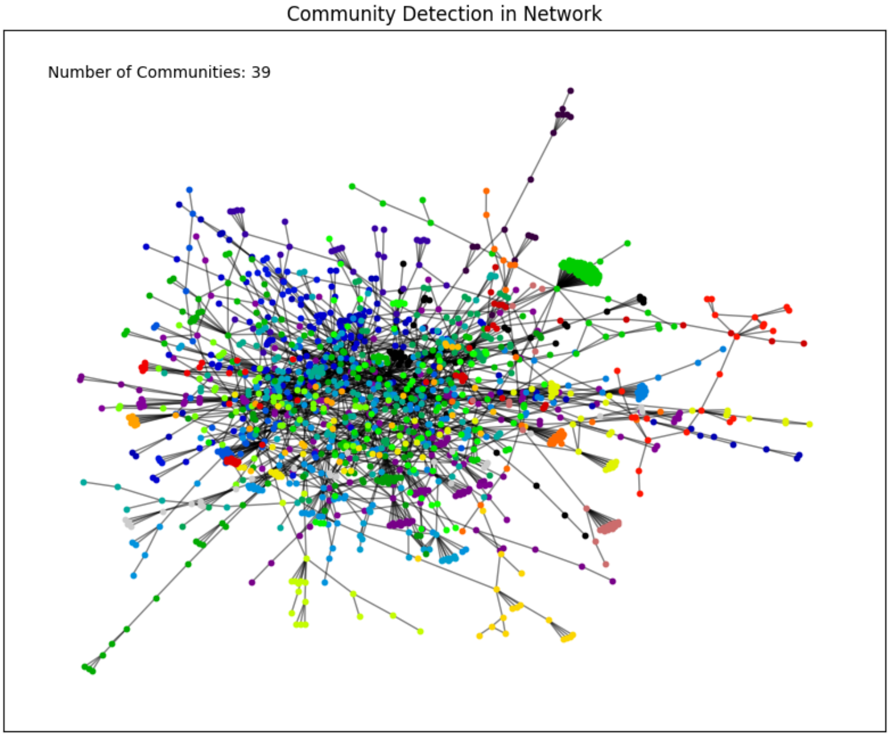
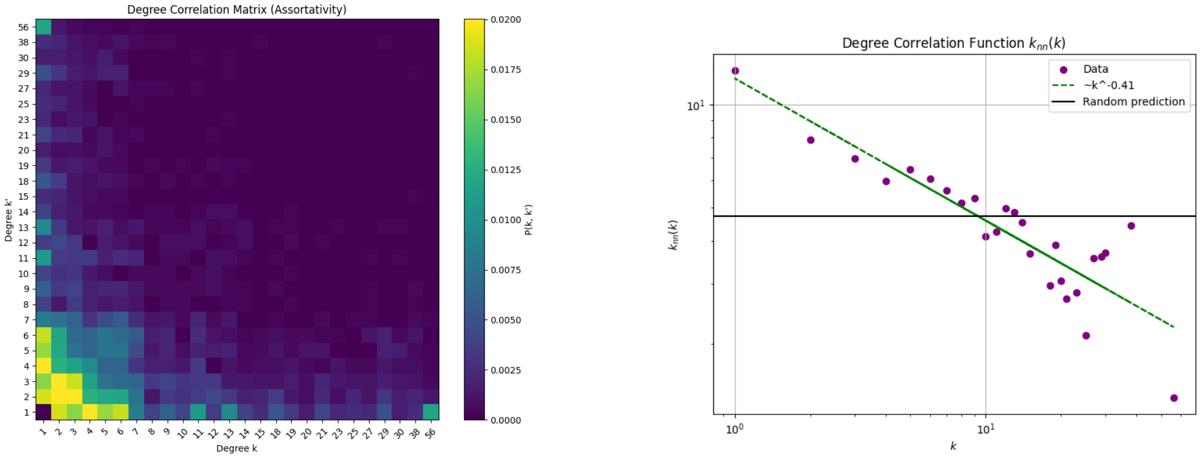


Figure 1.2: Visualization of detected communities in the network.

1.2.5 Degree Correlation

Assortativity

The calculated assortativity coefficient is $r = -0.2095$, indicating a disassortative network. This implies that high-degree nodes tend to connect with low-degree nodes, a typical behavior in biological networks where central nodes act as hubs connected to peripheral nodes with lower degrees.



(a) Degree correlation heatmap.

(b) Average neighbor degree $k_{nn}(k)$ vs k.

Figure 1.3: Analysis of degree correlation in the network.

The second image shows that low-degree nodes ($k = 1, 2, 3, \ldots$) connect to hubs with much higher average degrees ($k_{\text{nn}}(1) = 12.55$), while high-degree nodes ($k = 20, 25, 38, \ldots$) link to lower-degree nodes ($k_{\text{nn}}(20) = 3.05$, $k_{\text{nn}}(56) = 1.39$). The negative trend, following $k^{-0.41}$, highlights disassortativity typical of scale-free networks, where hubs prefer connections with peripheral nodes.

Chapter 2

Phase 2: Model Comparison

This chapter focuses on comparing the real-world network with synthetic models. Metrics such as degree distribution and clustering coefficient are analyzed.

2.1 Erdős–R'enyi model

- **Average Degree** The real average degree of the network is 2.67, while the model predicts an average degree of 2.73. The relative difference between the real and modeled degrees is 2.05%, indicating a good alignment with the theoretical Erdős–R'enyi model.

- **Degree Distribution** As discussed previously, the degree distribution of our real network follows a power law. In contrast, the degree distribution of the Erdős–Rényi model follows a binomial distribution, which can be approximated by a Poisson distribution when n is sufficiently large and p is very small—ensuring that np remains constant. In this context, the average degree can be written as $\langle k \rangle = np \approx np(1-p)$. In our case, with $n = 1458$ and $p = 0.00183$, the average degree is: $\langle k \rangle = np = 1458 \cdot 0.00183 \approx 2.67$, while $np(1-p) = 1458 \cdot 0.00183 \cdot (1 - 0.00183) \approx 2.66$.
  Given that p is small and n is moderately large, the Poisson approximation is justified, making this a reasonable simplification for analyzing the network's behavior.

- **Average Path Length** The graph is not connected; hence, the average path length cannot be calculated directly. This lack of connectivity is consistent with the expected behavior of a sparse Erdős–R'enyi graph at low connection probabilities.

- **Clustering Coefficient** The real average clustering coefficient is 0.0708, while the model predicts a much lower value of 0.0026. The relative difference of 96.39% highlights that the real graph exhibits a significantly higher clustering tendency compared to the theoretical expectation, suggesting deviations from randomness in local connectivity patterns.

- **Assortativity** The real network has an assortativity coefficient of -0.2095, indicating a disassortative mixing pattern where high-degree nodes tend to connect with low-degree nodes. The model predicts an assortativity of -0.0015, which is close to zero, as expected for a random model. The absolute difference of 0.2080 suggests that the real network exhibits stronger disassortative behavior than expected from the model.

- **Community Detection** Community detection analysis identifies 128 distinct communities in the network. Since the model is random, the number of communities identified is close to the number of connected components, highlighting how the random model is far from the real model representation.

2.2 Watts-Strogatz

Although we knew the real network exhibited scale-free properties, we tested the Watts-Strogatz model to systematically explore the role of clustering and small-world effects in approximating the network's structure. The Watts-Strogatz model, with its tunable rewiring probability (p), serves as an intermediary between completely random (Erdős-Rényi) and entirely regular networks, offering insights into how local cohesion and global efficiency interplay under specific parameter settings. Using $k = 2$ and $p = 0.1$, we sought to investigate whether its clustering and path-length dynamics, despite the model's known limitations for scale-free networks, could still provide partial insights or inform alternative perspectives on the real network's topology. This exploration allowed us to evaluate theoretical models beyond scale-free assumptions, ensuring a more comprehensive understanding of network behavior.

- **Average Degree**
  The real graph exhibits an average degree of 2.67, while the model produces an average degree of 2.00. This results in a relative difference of 25.15%, reflecting the inability of the Watts-Strogatz model under $k = 2$ to fully capture the connectivity patterns of the real network.

- **Degree Distribution**
  The degree distribution does not follow a binomial distribution or a power law as in other theoretical models, but it depends on the rewiring parameter. With a rewiring probability of 0.1 in our network, it retains many local connections, but some edges are randomly rewired. This results in a degree distribution that is not completely concentrated around the average degree, but shows a slight dispersion around it. So, no hubs are observed, but nodes with low degrees, with a slight variability in the degrees.

- **Average Path Length**
  The graph is not connected; hence, the average path length cannot be calculated directly. This lack of connectivity is consistent with the expected behavior of a Watts-Strogatz model with $k = 2$ and $p = 0.1$, where the network may form disconnected clusters due to the low average degree and limited rewiring probability, preventing the formation of a fully connected structure.

- **Clustering Coefficient**
  The clustering coefficient in the real network is 0.0708, whereas the model produces a clustering coefficient of 0.0000. This result is expected because, with $k = 2$, the network's sparsity significantly reduces the likelihood of forming triangles, making the Watts-Strogatz model unsuitable for approximating the real network's clustering properties under these parameters.

- **Assortativity**
  The real network's assortativity coefficient is -0.2095, indicating a strong disassortative mixing pattern, where high-degree nodes are more likely to connect with low-degree nodes. The Watts-Strogatz model, however, produces an assortativity coefficient of -0.0316, showing a much weaker tendency for disassortativity. This is expected because the model's structure and limited rewiring at $p = 0.1$ do not favor the formation of hubs or the significant degree mixing observed in the real network.

- **Community Detection**
  The real network exhibits a modular structure with 39 well-defined communities, reflecting its higher modularity and strong local cohesion. In contrast, the Watts-Strogatz model with $k = 2$ and $p = 0.1$ tends to produce chain-like structures within its communities. These elongated, sparsely connected clusters are a direct consequence of the low degree $k = 2$,

which limits the formation of densely interconnected groups. This structural limitation prevents the model from replicating the well-defined and compact communities observed in the real network.

2.3 Barab'asi–Albert model

- **Average Degree**
  The real graph exhibits an average degree of 2.67, while the model produces an average degree of 2.00. This results in a relative difference of 25.21%, indicating that the Barab'asi–Albert model underestimates the average degree compared to the real network. However, since the Barab'asi–Albert model can only produce integer degrees, due to the preferential attachment, the choice of degree is inherently limited, explaining the observed difference.

- **Degree Distribution**
  As in real networks, the degree distribution of the Barabási-Albert model follows a power law, suggesting a scale-free property.

- **Average Path Length**
  The real network has an average path length of 6.81, while the model network yields 6.72. The relative difference is only 1.31%, showing that the Barab'asi–Albert model closely approximates the real network's path length. This suggests that the model reproduces the small-world structure observed in many real networks.

- **Clustering Coefficient**
  The clustering coefficient in the real network is 0.0708, whereas the model produces a clustering coefficient of 0.0000. This leads to a relative difference of 100.00%, implying that the Barab'asi–Albert model fails to replicate clustering behavior and lacks triangle formation. This result arises because, with $m = 1$, the degree divided by 2 does not allow the formation of triangles.

- **Assortativity**
  The real network has an assortativity value of -0.2095, while the model reports -0.0781. The absolute difference of 0.1314 highlights that the model captures disassortative behavior but deviates from the real network's degree correlation structure. The low disassortativity observed in the Barab'asi–Albert model is expected, as it is a typical characteristic of scale-free networks.

- **Community Detection**
  The Barab'asi–Albert model may not accurately reproduce community structures, as it focuses on preferential attachment rather than modular organization but we can say that the distribution of communities is consistent with the presence of hubs; however, the model exhibits more communities, which may result from deviations driven by non-random processes.

2.4 Best model

The Barabási-Albert model is considered the best approximation of the real-world network due to its ability to capture key structural properties. It reproduces the small-world characteristic, with an average path length very close to that of the real network (6.72 vs. 6.81, a relative difference of only 1.31%). Additionally, its preferential attachment mechanism reflects the presence of hubs, which are common in real-world networks, and partially aligns with the observed disassortative mixing pattern. While the model does not replicate the clustering coefficient accurately, this limitation stems from its design and does not undermine its overall suitability for approximating the network's broader characteristics, outperforming the Erdős-Rényi model in this regard.

Chapter 3

Phase 3 & 4: Simulation and Strategies

The phase 3 evaluates the network's robustness, efficiency, and plasticity through simulations based on specific hypotheses. The analyses are grouped into three main categories:
Node Removal (Robustness) (H1, H2).
H1: Hub node removal drastically reduces net- work connectivity.
H2: Random node remo- val has a smaller impact than targeted hub remo- val.
Edge Addition (Efficiency) (H3, H4)
H3: Adding random ed- ges reduces average shor- test path length.
H4: Adding random ed- ges between communi- ties reduces the number of communities and modularity.
Rewiring (Plasticity) (H5).
H5: Random rewiring reduces modularity but improves path length.
The following section present detailed results for each hypothesis, supported by tables summarizing the observed metrics and trends.

3.1 Hypoteses, Test Summary and Observed Results

Table 3.1: Hypotheses and Tests Summary

H	**Rationale**	**Test**
H1	Hub proteins (high-degree nodes) are crucial for connectivity; removing them might cause fragmentation.	Sequentially remove high-degree nodes and measure the number of connected components, the size of the largest component, and changes in the APL.
H2	Random failures are less likely to affect critical hubs.	Compare the impact of random vs targeted node removal on network connectivity.
H3	Additional links might provide shortcuts, improving efficiency in information transfer.	Add random edges and observe changes in the average path length.
H4	Random edges can simulate mutations, disrupting modular organization and reducing functional specialization.	Evaluate changes in clustering coefficient and modularity after adding intra-cluster edges.
H5	Biological systems balance modularity and global efficiency; rewiring disrupts this balance.	Rewire edges randomly and track changes in modularity and average shortest path length.

Table 3.2: Observed Results for Hypotheses

H	**Observed Results**
H1	The removal of 1, 2, and 3 hubs resulted in 50, 64, and 73 **connected components**, respectively. **Largest component size** decreased to 1401, 1375, and 1360, with **average path lengths** (on the largest cc) of 6.85, 7.08, and 7.24.
H2	Random node removal caused no fragmentation, preserving a single **connected component** and a minimal change in **path length**.
H3	Adding 10, 20, and 30 edges progressively reduced the **average shortest path length** from 6.81 to 6.79, 6.74, and 6.72, respectively. Further additions up to 300 edges continued to decrease the path length, reaching 5.96 with 300 additional edges. The network diameter decreased from 19 to 14, while average clustering coefficient showed a slight reduction from 0.0708 to 0.0599, indicating improved global efficiency at the cost of local connectivity.
H4	**Modularity** dropped from 0.82 to 0.73 after adding random edges incrementally. The **average clustering coefficient** decreased slightly from 0.071 to 0.051, suggesting a loss of local cohesiveness.
H5	Rewiring progressively reduced **modularity** from 0.82 to 0.68 as 90% of edges were rewired. The **average path length** (largest connected component) initially decreased from 6.81 to 6.31 but then increased to 6.67, reflecting a temporary improvement in efficiency followed by a decline as structural coherence was disrupted. The network diameter oscillated, showing periods of reduction but ultimately stabilizing at a slightly higher value, indicating inconsistent global efficiency. The clustering coefficient fell sharply, and the number of connected components increased, highlighting a significant loss of local cohesion and fragmentation into smaller subgraphs.

3.2 Strategies and Simulation Results

3.2.1 Experiment Setup and Objectives

The central experiment involves the removal of the hub with the highest degree from the network and the evaluation of its impact. The primary goal is to mitigate the resulting fragmentation by employing edge-adding strategies. Additionally, we introduce a novel metric, network efficiency, to assess the effectiveness of these strategies in comparison to the Average Path Length (APL), which has certain limitations in fragmented networks.

3.2.2 Metrics for Network Analysis

1. Average Path Length (APL) APL measures the average shortest path between all node pairs in a connected component. While useful, it fails to capture the effects of disconnected components, where the shortest path is undefined.
2. Global Efficiency Global efficiency, defined as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{3.1}$$

where d_{ij} is the shortest path between nodes i and j, is a more robust measure. It assigns a contribution of 0 to disconnected pairs, making it suitable for fragmented networks.
To quantify the improvement in global efficiency achieved by applying a strategy, we define the Global Efficiency Improvement (GEI) as:

$$\text{GEI} = \frac{E_{\text{strategy}} - E_{\text{fragmented}}}{E_{\text{original}}} \times 100, \tag{3.2}$$

The GEI provides a normalized measure of how effectively a strategy restores the network's efficiency relative to its original state.

3.2.3 Hub Removal and Repair Strategies

The analysis from Phase 3 revealed that the removal of hub nodes significantly fragmented the network, as shown in table 3.3. The number of connected components (CCs) increased sharply, while the size of the LCC remained relatively intact, demonstrating robustness, whereas in the Barabási-Albert model, it dropped significantly, as detailed in table 3.3.
Additionally, random node failures had minimal impact, highlighting the network's inherent resilience to random disruptions. The addition of random edges improved global efficiency by reducing average path lengths, though this slightly reduced local clustering. Similarly, adding edges between communities enhanced global connectivity at the expense of modularity.
These findings underline the necessity of targeted repair strategies to mitigate fragmentation caused by hub removal. In Phase 4, two main strategies were developed and tested to address these challenges.

Observation	**Barabási-Albert Network**	**Bio-Yeast Network**
Number of Connected Components (CCs)	Increased from 1 to 120	Increased from 1 to 50
Size of Largest Connected Component (LCC)	Dropped to 49% of the original network size	Retained 96% of the original network size
Global Efficiency	Decreased significantly from 0.1668 to 0.0492	Decreased slightly from 0.1638 to 0.1508
Average Path Length (APL) within the LCC	Decreased from 6.78 to 6.29	Increased slightly from 6.81 to 6.85

Table 3.3: Impact of Hub Removal on Barabási-Albert Model and Bio-Yeast Network

Proposed Strategies

To address this fragmentation, we propose two generalizable strategies for reconnecting the network efficiently. These strategies focus on representative nodes of the connected components (CCs) and operate effectively across scenarios where $k \leq n$, where k is the number of connected components and n is the number of orphan nodes disconnected from the hub.
1. **Fully Connected (FC):** Maximizes robustness by connecting all representative nodes of the connected components (CCs) with edges between every pair.
2. **Spanning Tree (ST):** Minimizes redundancy by connecting representative nodes of the CCs with the minimal number of edges needed to maintain connectivity.
The strategies were initially tested on the Barabási-Albert network to assess their generalizability and later applied to the real-world Bio-Yeast network to evaluate their effectiveness in a biological context.

Str	**Description**	**Advantages**	**Disadvantages**
FC	Fully connect all representative nodes to ensure complete component connectivity.	Maximizes resilience and restores network robustness by connecting all components.	Significant edge count $k(k-1)/2$, leading to redundancy and increased complexity.
ST	Construct a minimal spanning tree among all rappresentative nodes.	Minimizes the number of added edges $(k-1)$, providing an efficient solution.	Less resilient to targeted attacks compared to fully connected networks.

Table 3.4: Comparison Strategies (STR): Description, Advantages, and Disadvantages

3.2.4 Application to Barabási Albert Network

Str	GEI	#EA	Key Observations
FC	Significant: from 0.0492 to 0.1949 (87.35%)	$\frac{k(k-1)}{2}$	Introduces redundancy, making the network robust but less efficient in edge management.
ST	Notable: from 0.0492 to 0.1699 (72.36%)	$k-1$	Provides an efficient solution with minimal redundancy but less resilience to targeted attacks.

Table 3.5: Simulation Results for Edge-Adding Strategies in the Barabási–Albert Network. Strategy (STR), Global Efficiency improvement (GEI), Number of edges addes (#EA)

Strategy Selection Based on k and n

1. **When $k \approx n$:** Each node disconnected from the hub forms its own CC. In this case, the Spanning Tree Approach is preferred, as it adds the minimal number of edges $(k-1)$ to achieve a single connected component.

2. **When $k << n$:** Many orphan nodes are already well-connected to the Largest Connected Component (LCC) or to each other. For such cases, the Fully Connected Network approach can be considered, as it provides maximum resilience with a higher edge count $(k(k-1)/2)$.

Both strategies are designed to operate effectively across various scenarios, ensuring that the network is reconnected while balancing resilience and efficiency. Depending on the specific network structure and requirements, either strategy can be selected to optimize connectivity restoration.

3.2.5 Application to the Bio Yeast Network

The Bio Yeast network shows higher resilience to hub removal than the Barabási-Albert model, with global efficiency decreasing slightly from $E_{\text{original}} = 0.1638$ to $E_{\text{fragmented}} = 0.1508$, unlike the significant drop observed in the synthetic model. The proposed strategies restored global efficiency in the Bio Yeast network with marginal improvements, as shown in table 3.6.

Str	GEI	#EA	Key Observations
FC	Marginal: from 0.1508 to 0.1647 (8.49%)	$\frac{k(k-1)}{2}$	Introduces redundancy, making the network robust but less efficient biologically.
ST	Marginal: from 0.1508 to 0.1642 (8.12%)	$k-1$	Provides an efficient solution with minimal redundancy but less resilience to targeted attacks.

Table 3.6: Simulation Results for Edge-Adding Strategies in the Bio Yeast Network

Biological Relevance of Restored Connections: Although the strategies successfully restore global efficiency and structural integrity, the biological significance of the newly introduced connections remains uncertain. Specifically:

- The restoration of connectivity does not guarantee the preservation of the eliminated hub's biological function, such as its role in metabolic pathways or enzymatic interactions.

- The added edges might not correspond to biologically plausible interactions, as they are determined purely based on structural considerations.

This limitation underscores the need for additional analysis to incorporate functional metrics that assess whether key biological pathways and interactions are maintained after the intervention. Future work should focus on integrating biochemical constraints and simulating dynamic network behaviors, such as metabolic flux, to ensure that the restored connections align with real-world biological processes.

Chapter 4

Key Findings and Recommendations

4.1 Key Findings

This study analyzed the yeast **Saccharomyces cerevisiae** interaction network through quantitative and simulation-based methods, benchmarking against theoretical models and testing reconnection strategies. The findings are summarized below:

Biological Network Analysis

The network, comprising 1458 nodes and 1948 edges, exhibits low density (0.00183), a sparse structure, and a power-law degree distribution (exponent 3.51), typical of scale-free networks.

Comparison with Theoretical Models

- **Erdős-Rényi (ER):** Matches the average degree but fails to replicate clustering and scale-free features.
- **Barabási-Albert (BA):** Captures scale-free behavior and path length but lacks clustering, underrepresenting local cohesion.

Simulations and Reconnection Strategies

- **Hub Removal:** Leads to significant fragmentation, increasing components and reducing global efficiency (0.1668 → 0.0492).
- **Reconnection Strategies:**
  - **Fully Connected (FC):** Maximizes robustness but adds redundancy.
  - **Spanning Tree (ST):** Efficient edge use but less robust against attacks.
- **Application to the Yeast Network:** Exhibits higher resilience; efficiency decreases minimally after fragmentation and reconnection.

Impact of Node Distribution within Components: Node distribution minimally affects the performance of FC and ST strategies, which focus on representative nodes, ensuring adaptability across fragmented structures.

4.2 Recommendations

- Use **FC strategies** for critical systems demanding maximum robustness, implementing **ST strategies** for systems prioritizing efficiency and minimal redundancy. Explore **hybrid approaches** combining FC and ST strengths to balance resilience and efficiency.
- Ensure biological reconnections respect functional constraints, preserving key pathways and interactions.