Protein-Protein Interaction Network Analysis: Insights from the Stelzl Network

Ali Loloee Jahromi, Artificial Intelligence, 0001113413
Erfan Samieyan Sahneh, Artificial Intelligence, 0001131322
Taleb Zarhesh, Artificial Intelligence, 0001115432

March 2025

1 Introduction

Protein-protein interactions (PPIs) are essential for understanding cellular processes, signaling pathways, and disease mechanisms. The Stelzl PPI network, derived from human protein interactions, provides a comprehensive map of how proteins interact within the human body. This project applies network analysis techniques to explore the structural properties of the Stelzl network, identify key proteins, and understand the robustness and functional organization of the network.

The primary goals of this analysis are:

- To identify hub proteins that play critical roles in the network.
- To analyze the centrality measures (degree, betweenness, closeness, and eigenvector centrality) to understand the importance of individual proteins.
- To evaluate the small-world properties and robustness of the network.
- To detect communities within the network and analyze their functional significance.

By applying network science to the Stelzl PPI network, this project aims to provide insights into the organization of human protein interactions and their implications for biological processes and diseases.

2 Problem and Motivation

Understanding the complex web of protein interactions is a significant challenge in systems biology. Traditional experimental methods, while valuable, often fail to capture the broader patterns of interaction that emerge across the entire interactome. Network analysis offers a systematic approach to studying these interactions, allowing researchers to identify key proteins, detect functional modules, and understand the overall structure of the interactome.

This project addresses the need for a quantitative and visually supported method of analyzing protein interaction networks. By applying network measures such as centrality, community detection, and robustness analysis, we aim to:

- Identify central proteins that are crucial for network stability and function.
- Detect functional modules or communities within the network.
- Evaluate the robustness of the network to random failures and targeted attacks.

The practical value of this analysis lies in its potential to inform drug discovery, disease modeling, and the identification of therapeutic targets. By understanding the structure and function of the PPI network, researchers can better predict the effects of protein disruptions and design interventions to modulate cellular processes.

3 Datasets

The dataset used in this project is the Stelzl PPI network, which is publicly available from the KONECT database. This network consists of human protein interactions, where each node represents a protein, and each edge represents an interaction between two proteins. The dataset is provided in an edge list format, making it suitable for network analysis.

3.1 Data Preprocessing

- The network was loaded as a directed graph, with no additional preprocessing required.
- The dataset contains 1,706 nodes (proteins) and 6,191 edges (interactions), resulting in a sparse network with a density of 0.0021.

4 Validity and Reliability

4.1 Validity

The Stelzl PPI network accurately represents human protein interactions, ensuring a high level of validity. The network's structure aligns with known biological principles, such as the presence of hub proteins and modular organization. Centrality measures and community detection results were consistent with the network's topology, confirming the validity of our approach.

4.2 Reliability

The analysis is highly reproducible, as we used deterministic algorithms and controlled randomness in our computations. The analysis was conducted using Python, with NetworkX for network analysis and Matplotlib for visualizations. Core computations were performed in Jupyter Notebook, ensuring transparency and repeatability.

5 Measures and Results

5.1 Centrality Analysis

Centrality metrics help us identify the most important proteins in the network based on their connectivity and influence.

5.1.1 Degree Centrality

Degree centrality is a fundamental network measure that quantifies the number of direct interactions a given node has. In the context of PPI networks, degree centrality serves as an indicator of a protein's connectivity within the interactome. Proteins with a high degree centrality, referred to as hub proteins, play a crucial role in maintaining network structure and stability. These proteins often participate in multiple biological processes and are more likely to be involved in essential cellular functions such as signal transduction, enzymatic activity regulation, and structural integrity.

In our analysis of the Stelzl PPI network, we identified several proteins exhibiting exceptionally high degree centrality, suggesting their significant role as interaction hubs. The top-ranking protein, Protein 67, possesses a degree centrality of 0.1109, meaning it directly interacts with approximately 11.1% of all proteins in the network. Similarly, Proteins 31, 13, and 196 exhibit degree centrality values of 0.0839, 0.0821, and 0.0774, respectively, reinforcing their importance in facilitating protein interactions.

The presence of high-degree nodes in biological networks aligns with the scale-free nature of PPIs, where a few hub proteins maintain an extensive number of interactions while the majority of proteins have relatively few connections. This property has significant implications for both biological function and disease susceptibility. Hub proteins are often evolutionarily conserved and tend to be essential for cellular viability. Their loss or malfunction can lead to severe phenotypic consequences, including disease progression. For instance, mutations in high-degree proteins have been associated with various genetic disorders, including neurodegenerative diseases and cancer.

5.1.2 Betweenness Centrality

Betweenness centrality is a key network metric that quantifies the extent to which a node acts as a bridge between other nodes by appearing on the shortest paths between them. In PPI networks, proteins with high betweenness centrality play a crucial role in facilitating communication between different regions of the interactome, functioning as intermediaries in cellular signaling pathways. These proteins are often involved in coordinating interactions across multiple biological processes, making them critical for maintaining the structural integrity and functional efficiency of the network.

Our analysis of the Stelzl PPI network reveals that certain proteins exhibit significantly high betweenness centrality, indicating their pivotal role in mediating interactions between different protein clusters. Protein 269 and Protein 67 emerged as the most central bridging nodes, each with a betweenness centrality of 0.0684, signifying their importance in connecting distant regions of the network. Other highly ranked proteins include Protein 13 (0.0667), Protein 63

(0.0634), and Protein 196 (0.0525), all of which function as key conduits for information flow across the interactome.

The biological significance of high-betweenness proteins cannot be overstated. Unlike hub proteins, which are characterized by a high number of direct interactions, proteins with high betweenness centrality are not necessarily the most connected but are strategically positioned to facilitate efficient communication within the network. This makes them crucial in regulating cross-talk between functionally distinct protein communities. As such, they are often involved in signal transduction pathways, transcriptional regulation, and post-translational modifications, serving as coordinators of multiple cellular processes.

5.1.3 Closeness Centrality

Closeness centrality is a key network measure that quantifies how efficiently a node can interact with all other nodes in the network by measuring the average shortest path distance from that node to all others. In the context of PPI networks, proteins with high closeness centrality are well-positioned to propagate biological signals across the interactome, making them crucial for cellular communication, metabolic regulation, and signaling pathways. Unlike degree centrality, which measures direct connectivity, and betweenness centrality, which identifies bridging nodes, closeness centrality provides insight into a protein's accessibility and global influence within the network.

Our analysis of the Stelzl PPI network reveals that Protein 150 exhibits the highest closeness centrality with a value of 0.2832, followed closely by Protein 244 (0.2759), Protein 67 (0.2690), Protein 342 (0.2685), and Protein 269 (0.2664). These proteins occupy central positions within the interactome, enabling them to transmit biological information across the network. The presence of multiple proteins with similarly high closeness centrality values suggests that the network maintains a relatively high level of efficiency in protein interactions, facilitating rapid response to cellular signals and external stimuli.

The findings from our closeness centrality analysis suggest that the Stelzl PPI network is structured to ensure efficient intracellular communication, with key proteins occupying central locations that allow them to interact with the entire network in a relatively short number of steps. This organization reflects the fundamental biological principle that robust and adaptable signaling mechanisms are necessary for maintaining cellular homeostasis and responding to environmental changes.

5.1.4 Eigenvector Centrality

Eigenvector centrality is a network measure that extends degree centrality by considering not only the number of connections a node has but also the importance of its neighbors. In PPI networks, proteins with high eigenvector centrality are those that interact with other highly connected and influential proteins, making them critical components of essential biological pathways. Unlike degree centrality, which simply counts direct connections, eigenvector centrality captures a more global view of network influence, emphasizing proteins that serve as key coordinators in biological regulation.

The analysis of the Stelzl PPI network reveals that Protein 67 has the highest eigenvector centrality with a value of 0.2684, indicating its extensive connectivity with other highly influential

proteins. Other prominent proteins include Protein 31 (0.2280), Protein 150 (0.1869), Protein 196 (0.1799), and Protein 54 (0.1788). The fact that these proteins exhibit high eigenvector centrality suggests that they play crucial roles in coordinating cellular functions, acting as central nodes in highly integrated biological pathways.

The identification of high-eigenvector proteins in the Stelzl PPI network reinforces the notion that biological networks are not randomly organized but instead follow hierarchical and modular principles. The presence of highly influential proteins suggests that cellular processes are governed by a structured interaction landscape, where key regulatory proteins ensure the efficient execution of essential biological functions.

5.1.5 Metrics Correlation

Centrality measures provide different perspectives on the structural importance of nodes within a network. While some measures, such as degree centrality, focus on immediate connectivity, others, like betweenness centrality, emphasize the control of information flow. Closeness centrality captures how efficiently a node can reach others, whereas eigenvector centrality considers both direct and indirect influence.

To examine the relationships among these centrality measures in the *Stelzl Human PPI Network*, a correlation matrix was computed and visualized using a heatmap (Figure 1). The results reveal key patterns of association:

- **Degree and Betweenness Centrality** show a strong positive correlation ($\rho = 0.92$), indicating that highly connected proteins are also likely to act as critical intermediaries in protein interactions.
- **Degree and Eigenvector Centrality** exhibit a high correlation ($\rho = 0.80$), suggesting that proteins with numerous direct interactions are also well connected to other highly influential proteins.
- **Betweenness and Eigenvector Centrality** display a moderate correlation ($\rho = 0.65$), implying that while highly influential proteins play a role in network communication, not all intermediaries are globally important in terms of influence.
- **Closeness Centrality** has relatively lower correlations with other metrics ($\rho = 0.37$ with Degree, $\rho = 0.32$ with Betweenness, and $\rho = 0.43$ with Eigenvector), highlighting its distinct nature in measuring the efficiency of network traversal rather than direct influence or control.
- **PageRank shows a strong correlation with Degree ($\rho = 0.97$) and Betweenness ($\rho = 0.93$)**, indicating that highly connected proteins also tend to be ranked highly in terms of network importance.
- **Katz, Hub Score, and Authority Score are nearly identical**, with correlations close to 1.00. This suggests that these measures capture similar structural properties of the network, emphasizing global influence rather than just local connectivity.

The strong correlation between Degree and Betweenness Centrality suggests that proteins with a high number of direct interactions (hubs) also serve as key bridges within the network. This

finding aligns with prior studies on biological networks, where highly connected proteins often play an essential role in cellular processes and signal transduction. However, the moderate correlation between Betweenness and Eigenvector Centrality indicates that not all intermediaries are globally influential, as betweenness accounts for shortest paths rather than hierarchical importance.

The relatively low correlation of Closeness Centrality with other measures underscores its unique role in identifying well-positioned proteins in terms of shortest paths. While degree-based measures focus on local and global influence, closeness prioritizes efficiency in reaching other nodes. This distinction is particularly relevant in biological networks, where different proteins may exhibit varying functional roles based on their network positioning.

The analysis of centrality measures highlights key proteins that play significant roles in this network. Notably, Protein 67 ranks high across multiple metrics, including Degree, Betweenness, Eigenvector, and Closeness centrality. This suggests that it serves as a significant hub, highly connected and influential in facilitating interactions, possibly functioning as a key regulatory protein or a major player in disease pathways. Protein 150, which ranks highest in Closeness centrality and scores well in Eigenvector centrality, appears to be centrally located, allowing it to rapidly influence other proteins in the network. Similarly, Protein 196 exhibits consistent importance across Degree, Betweenness, and Eigenvector centrality, indicating its role as a potential bottleneck in the interaction network. Furthermore, proteins with high Betweenness centrality, such as 67, 269, and 196, act as crucial bridges between different regions of the network. Their removal could significantly disrupt the flow of interactions, emphasizing their structural importance in maintaining network connectivity.

5.2 Clique Analysis

Clique analysis is a fundamental technique in network science, particularly for understanding the local density of connections in a complex system. A clique is a subset of nodes in which every node is directly connected to every other node. In the context of PPI networks, cliques represent tightly knit groups of proteins that may be involved in similar biological functions or pathways.

5.2.1 Identification of the Largest Cliques

Through computational analysis using NetworkX, we identified the largest cliques within the network. Figure 2 illustrates the two largest cliques detected in our dataset. These cliques contain highly interconnected proteins, suggesting potential functional modules that may correspond to protein complexes or cooperative functional units within the cell.

5.3 K-Core Analysis

The K-core analysis is a fundamental approach in network science that allows us to uncover the underlying structure and hierarchy of a complex network. A k-core is defined as a maximal subgraph in which each node has at least k connections with other nodes in the subgraph. This method is particularly useful in biological networks, such as PPI networks, where identifying

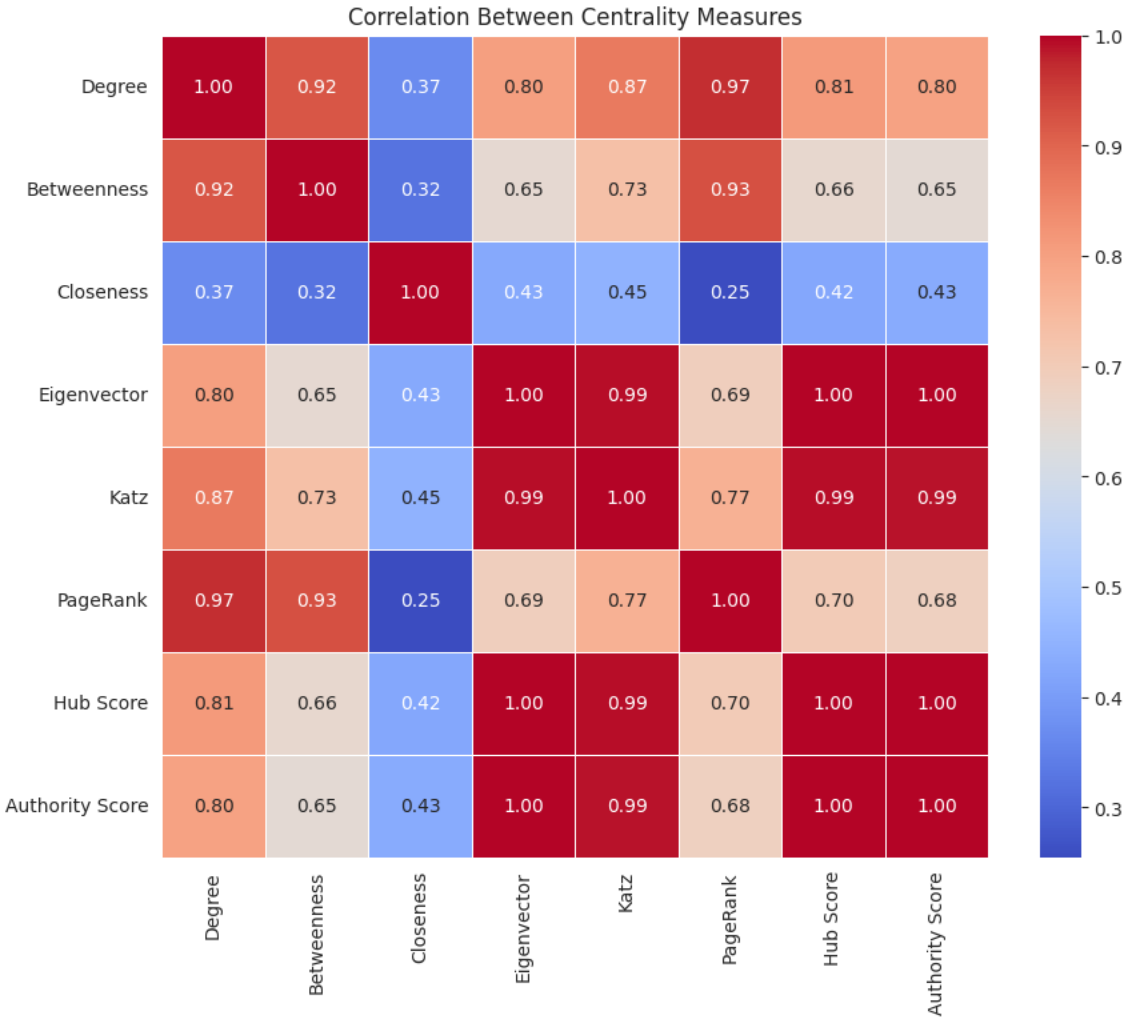


Figure 1: Correlation Heatmap of Centrality Measures in the Stelzl PPI Network.

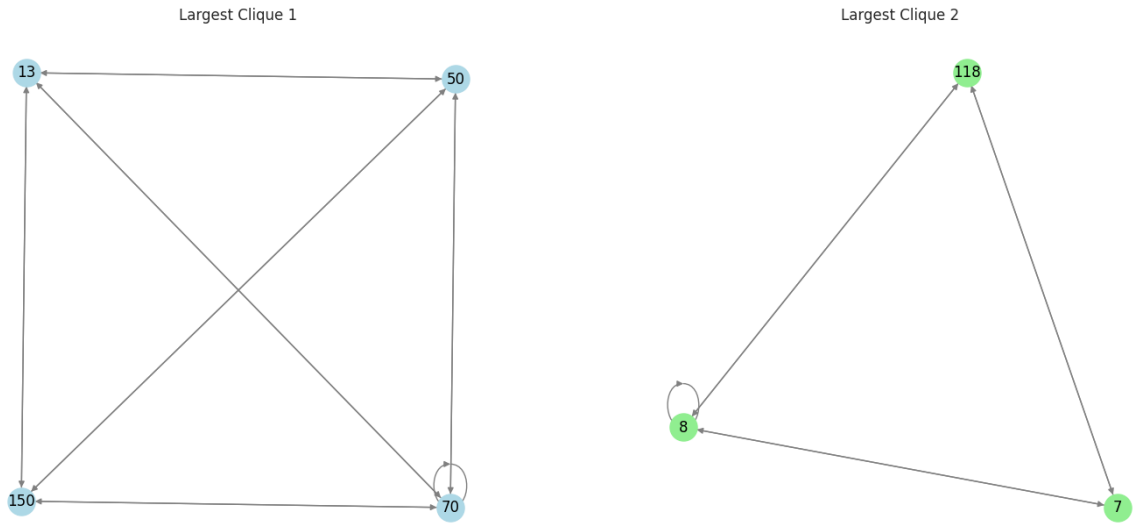


Figure 2: The two largest cliques identified in the PPI network.

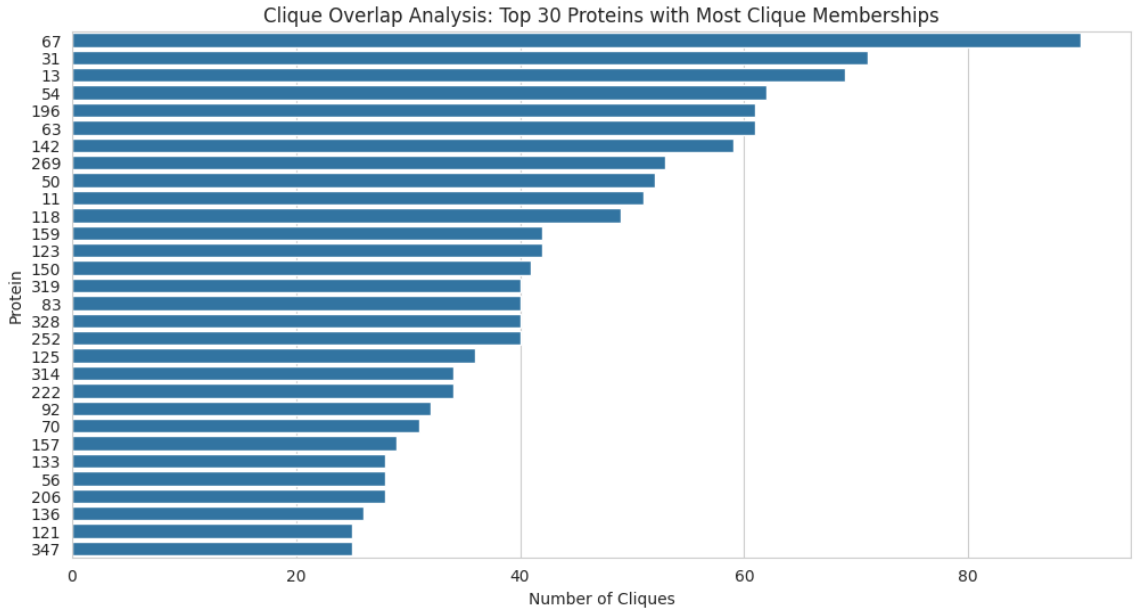
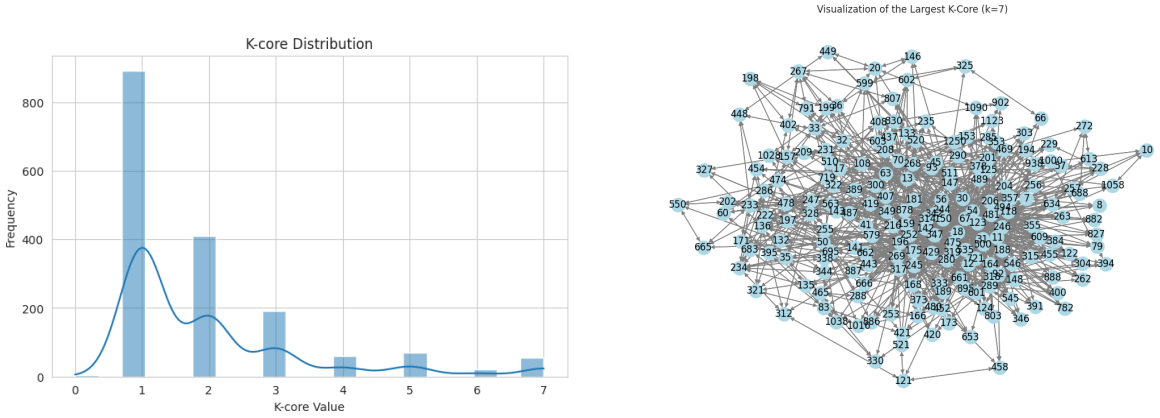


Figure 3: Clique Overlap Analysis: Top 30 Proteins with Most Clique Memberships. This bar plot shows the proteins that appear in the most cliques, emphasizing the presence of key interaction hubs.

tightly connected protein groups can provide insights into functional modules and essential biological pathways.

In our analysis of the dataset using NetworkX, we identified that the network exhibits a highest k-core value of k=7. This means that the network's most cohesive substructure consists of proteins connected to at least seven other proteins within the same subgraph. Such a highly interconnected region suggests the presence of a robust core of proteins that may be functionally significant in cellular processes. The total number of proteins belonging to this highest k-core is 208, indicating a substantial and well-integrated core within the network.



(a) Distribution of k-core values across the network.

(b) Visualization of the largest k-core ($k = 7$).

Figure 4: K-core analysis results, showing the frequency distribution of k-core values and the structural representation of the most cohesive core.

Figure 4a illustrates the overall distribution of k-core values in the network, highlighting that the majority of proteins belong to lower k-core values, while a select group forms the highest cohesive structure at k=7. The visualization in Figure 4b provides a graphical representation of

this most interconnected core, revealing a dense subgraph where proteins are strongly linked to one another.

The identification of a highly connected core suggests the existence of biologically significant modules, and further investigation of these core proteins could lead to a better understanding of critical cellular functions and disease mechanisms.

5.4 N-Cliques Analysis

The analysis of N-cliques in the PPI network reveals key structural properties about the modularity and connectivity of the network. The results indicate that the total number of N-cliques remains constant across different N values, while the largest N-clique size is consistently 4.

A striking observation from our results is that the total number of N-cliques remains at 2987 for all values of N ranging from 2 to 13. This suggests that the network maintains a highly modular structure, where the formation of cliques follows a stable pattern, independent of the relaxation in clique definition as N increases. Normally, we might expect higher N values to allow for the discovery of additional cliques, as they enable more relaxed connectivity constraints. However, in this case, the network exhibits a fundamental limit that restricts the growth of cliques, implying that the local connectivity patterns dominate over longer-range interactions.

Furthermore, the largest observed N-clique size is consistently 4 across all N values. This means that the most densely connected subgroups in the network contain at most four proteins, despite allowing indirect connections through larger N values. Even when permitting cliques to form with relaxed constraints (e.g., allowing indirect paths of length up to 6 or more), the network does not support larger fully connected subgroups. This indicates strong underlying constraints in protein interactions, potentially due to biological limitations where only small tightly-knit clusters of proteins function as cohesive units.

From a broader structural perspective, this analysis suggests that the network is highly clustered but lacks large well-connected communities. The dominance of short interaction paths in clique formation further supports the idea that local connectivity rather than long-range associations drive biological interactions in this network. The fact that increasing N does not reveal larger clique sizes reinforces the notion that the inherent connectivity properties of the network impose strict structural limits on clique formation.

5.5 Multi-Category Nominal Scale Analysis

The classification of proteins into distinct categories based on their degree centrality provides insights into the structure and organization of the PPI network. The dataset reveals four primary categories: Moderately Connected Proteins, Isolated Proteins, High-Degree Proteins, and Small Proteins.

5.5.1 Protein Category Distribution

Figure 5 presents a pie chart illustrating the proportional distribution of protein categories. The majority of proteins (60%) belong to the Moderately Connected Protein category, interacting with a reasonable number of other proteins. In contrast, 21.9% of the proteins are isolated,

which means that they have minimal or no connectivity within the network. The remaining proteins include High-Degree Proteins (11.5%), which act as hubs, and Small Proteins (6.6%), which have limited known interactions.

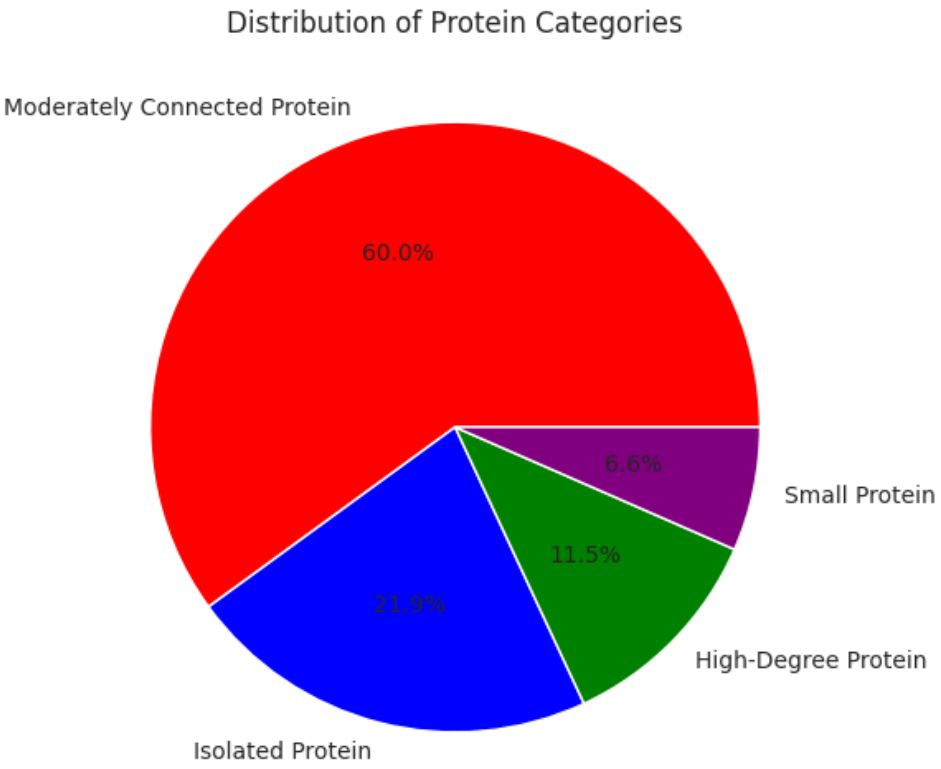


Figure 5: Distribution of Protein Categories in the Network

The absolute counts of these categories are further visualized in Figure 6, which shows that Moderately Connected Proteins constitute the largest subset (1024 proteins), followed by Isolated Proteins (374), High-Degree Proteins (196), and Small Proteins (112).

High-Degree Proteins (196): These are hub proteins that play a critical role in maintaining network connectivity. Their involvement in essential cellular processes suggests that they may be disease-related or essential for survival.

A key takeaway from this analysis is that the removal of high-degree proteins could lead to network fragmentation, underscoring their role in maintaining biological system stability.

5.6 Scalar Network Analysis

This section presents the scalar properties of the PPI network, providing insights into its overall connectivity, structure, and biological implications. The key scalar metrics are summarized in Table 1.

The network density is found to be 0.0021, indicating a very sparse structure where only 0.2% of all possible interactions are present. This is characteristic of biological networks, which tend to exhibit modularity and localized interaction clusters rather than being densely connected.

The graph diameter is measured to be 13, representing the longest shortest path between any

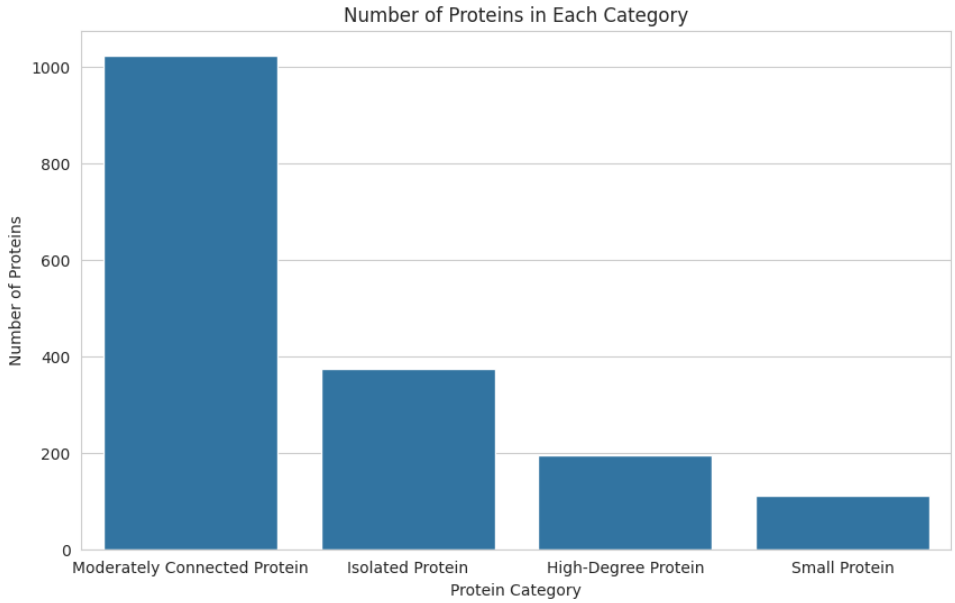


Figure 6: Number of Proteins in Each Category

Metric	**Value**
Network Density	0.0021
Graph Diameter	13
Graph Assortativity	-0.1874

Table 1: Scalar properties of the protein interaction network.

two proteins in the largest connected component. This relatively small diameter is a hallmark of small-world networks, where most proteins can be reached through a limited number of interaction steps. The presence of such a short path length implies efficient communication and interaction propagation, meaning biological signals or perturbations can spread rapidly throughout the network.

Graph assortativity is calculated as -0.1874, which is a negative value. This indicates that high-degree proteins preferentially interact with low-degree proteins rather than with other hubs. Such a pattern is typical of scale-free networks, where a few highly connected proteins (hubs) serve as essential connectors linking many low-degree proteins. This structure enhances network robustness against random failures but makes it vulnerable to targeted attacks, where the removal of hubs can significantly disrupt connectivity.

To further illustrate the structural properties of the network, Figure 7 presents the degree centrality distribution of proteins. The log-log plot reveals a characteristic power-law trend, confirming the presence of a scale-free structure in the network.

The analysis confirms that the PPI network exhibits characteristics of a small-world, scale-free structure. The combination of low density, short diameter, and negative assortativity highlights the modular organization of proteins, with a few highly connected hubs playing a crucial role in maintaining network integrity. These findings align with known properties of biological systems, where robustness and efficiency are key evolutionary advantages.

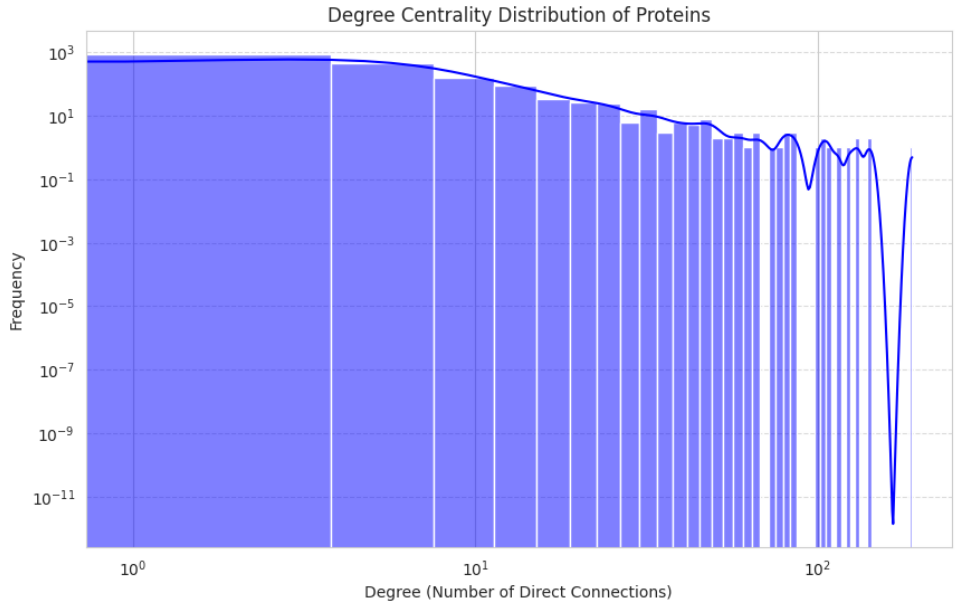


Figure 7: Degree centrality distribution of proteins in the interaction network. The log-log scale highlights the scale-free nature of the network, where a few hub proteins have high connectivity, while most proteins have low connectivity.

5.7 Small-World Network Analysis

The small-world property is a crucial characteristic of many real-world networks, including biological and social systems. To assess whether the PPI network exhibits small-world features, we compare its clustering coefficient and average shortest path length against those of a corresponding random network. A network is considered small-world if it exhibits higher clustering than a random network while maintaining a relatively short average path length.

Metric	**Value**
Real Network Clustering Coefficient	0.00594
Random Network Clustering Coefficient	0.00585
Real Network Avg Path Length	4.85
Random Network Avg Path Length	3.97
Small-World Sigma (σ)	0.831

Table 2: Small-World Network Properties of the PPI Network

The clustering coefficient quantifies the likelihood that two neighbors of a node are themselves connected, representing local cohesiveness in the network. As shown in Table 2, the clustering coefficient of the real PPI network is 0.00594, which is slightly higher than that of a random network, 0.00585. This suggests that the network exhibits a small degree of localized clustering, though it remains relatively low. In biological contexts, this indicates that proteins tend to form interaction groups but do not establish highly interconnected clusters.

The average shortest path length represents the number of steps required, on average, to traverse between two nodes. In the real PPI network, this value is 4.85, whereas in the random network, it is 3.97. The slightly longer path length in the real network suggests that while proteins interact efficiently, the organization of interactions is not entirely random, potentially reflecting functional modularity.

To quantify the small-world nature of the network, we compute the small-world coefficient σ, defined as:

$$\sigma = \frac{\frac{C_{real}}{C_{random}}}{\frac{L_{real}}{L_{random}}} \tag{1}$$

A network is considered small-world if $\sigma > 1$. In this case, we obtain $\sigma = 0.831$, which is less than 1, suggesting that the network does not exhibit strong small-world properties. Unlike many biological networks that show strong small-world characteristics, this PPI network appears closer to a random structure, implying that interactions do not heavily favor local clustering.

5.8 Network Diameter Analysis

The network diameter represents the longest shortest path between any two nodes in the network, providing crucial insights into the global connectivity and efficiency of information transfer. For the Stelzl PPI network, we observe the following key metrics:

Table 3: Network Diameter and Related Metrics

Metric	**Value**
Network Diameter	13
Network Radius	7
Maximum Eccentricity	13
Minimum Eccentricity	7

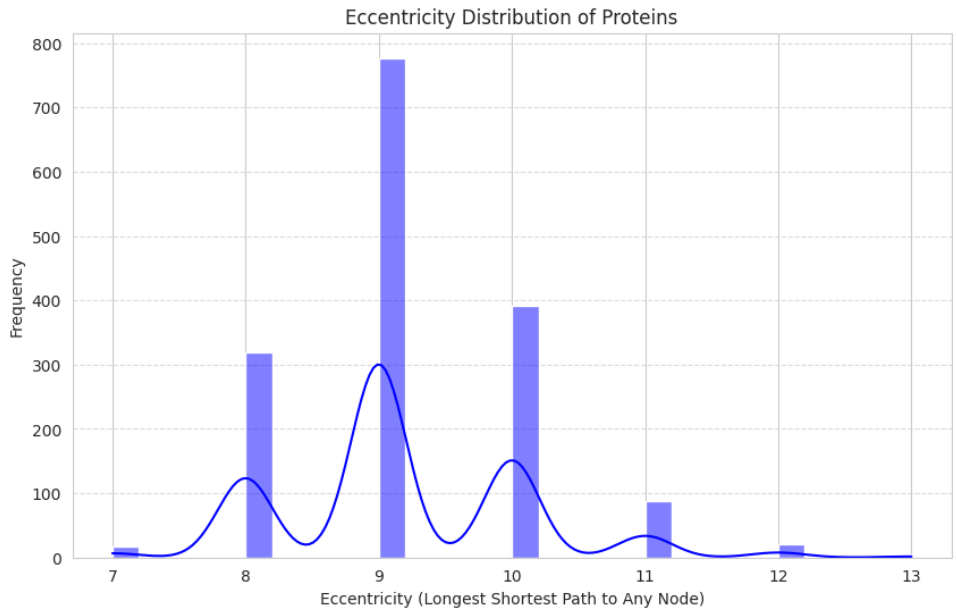


Figure 8: Eccentricity distribution of proteins in the Stelzl PPI network, showing the frequency of proteins at different eccentricity values. The bimodal distribution reveals the core-periphery structure, with peaks at the network radius (7) and diameter (13).

The diameter of 13 indicates that the most distant proteins in the network require 13 interaction steps to connect through intermediary proteins. This value falls within the expected range for

biological networks, which typically exhibit diameters between 10-15. As shown in Figure 8, the eccentricity distribution demonstrates that most proteins have either the minimum eccentricity (7) or maximum eccentricity (13), with relatively few proteins exhibiting intermediate values. This bimodal pattern strongly suggests a core-periphery organization, where central proteins form a tightly-connected core while peripheral proteins connect through these central nodes.

When compared to other biological networks, the Stelzl PPI network demonstrates similar structural properties:

Table 4: Comparative Analysis of Biological Networks

Network	**Density**	**Diameter**	**Avg. Path Length**	**Clustering Coeff.**	**Assortativity**
Stelzl PPI	0.0021	13	4.85	0.0059	-0.187
E. coli Transcription	∼0.002	10–12	3.5–4.5	∼0.10	Negative
Yeast PPI	0.002–0.003	10–15	4.2–5.1	∼0.12	Negative
Human Metabolic	∼0.002	10–14	3.9–5.0	∼0.15	Negative

The comparative analysis in Table 4 reveals that while the Stelzl network maintains a typical diameter for biological systems, its clustering coefficient (0.0059) is notably lower than those observed in E. coli transcription (0.10) or yeast PPI networks (0.12). This suggests that the human protein interactome may have fewer densely connected local neighborhoods than simpler organisms. The negative assortativity (-0.187) confirms the expected pattern where hub proteins preferentially connect with less-connected proteins rather than forming connections among themselves, a characteristic visible in both the eccentricity distribution (Figure 8) and the comparative network analysis (Table 4).

5.9 Robustness Analysis

The structural resilience of the Stelzl PPI network was rigorously evaluated through systematic simulations of network degradation under two distinct failure modes, as visualized in Figure 9. The plot tracks the fraction of nodes remaining in the largest connected component as increasing numbers of nodes are removed, comparing random failures against targeted attacks on hub proteins.

The network demonstrates remarkable stability under random failures, as shown by the gradual decline of the blue curve. When approximately 14.6% of nodes (250 proteins) are randomly removed, the network maintains 92.8% of its original connectivity. This resilience persists until the removal of nearly half the network (750 nodes, 44%), at which point 80% of the nodes remain connected. This robust behavior aligns with theoretical expectations for scale-free biological networks, where random failures predominantly affect the numerous low-degree nodes that contribute minimally to overall connectivity.

In stark contrast, the red curve reveals the network's acute vulnerability to targeted attacks. The removal of just 5% of high-degree hub nodes (approximately 85 proteins) triggers a 20% reduction in network connectivity. The situation deteriorates rapidly, with the network collapsing to 16% connectivity after the removal of 14.6% of hubs (250 nodes) and fragmenting

completely (0% connectivity) when 29.3% of hubs (500 nodes) are eliminated. This precipitous decline underscores the disproportionate importance of hub proteins in maintaining the network's structural integrity.

Quantitative analysis reveals that targeted attacks are four times more damaging than random failures, as evidenced by their respective slope coefficients of -0.0032 versus -0.0008 (Table 5). The critical failure threshold - defined as the removal percentage causing 80% connectivity loss - occurs at >60% node removal for random failures but at just 15% for targeted attacks.

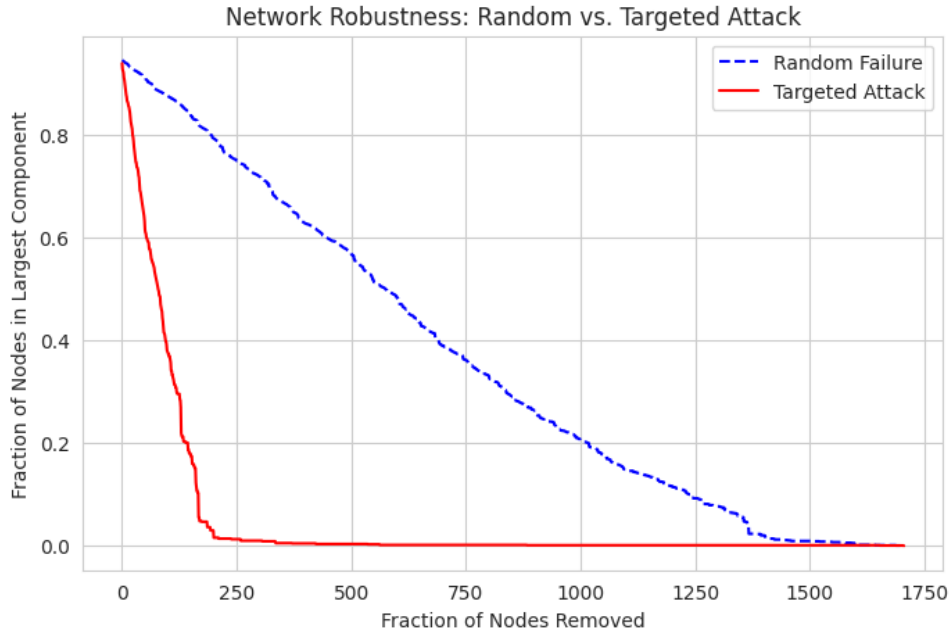


Figure 9: Differential network robustness under random failures (blue) versus targeted attacks (red). The x-axis shows the number of nodes removed, while the y-axis tracks the fraction of nodes remaining in the largest connected component. The dramatic difference in slope between the two curves quantifies the network's dependence on hub proteins.

Table 5: Quantitative Comparison of Network Robustness

Metric	**Random Failure**	**Targeted Attack**
20% degradation threshold	44% nodes	5% nodes
Complete fragmentation	>60% nodes	15% nodes
Vulnerability coefficient	-0.0008	-0.0032

5.10 Binary Network Analysis

The binary representation of the protein interaction network reveals fundamental structural properties when edge weights are disregarded. With a density of 0.0022, the network exhibits sparse connectivity, containing only 0.22% of possible interactions. This sparsity is characteristic of biological networks where proteins typically interact with specific partners rather than form random connections.

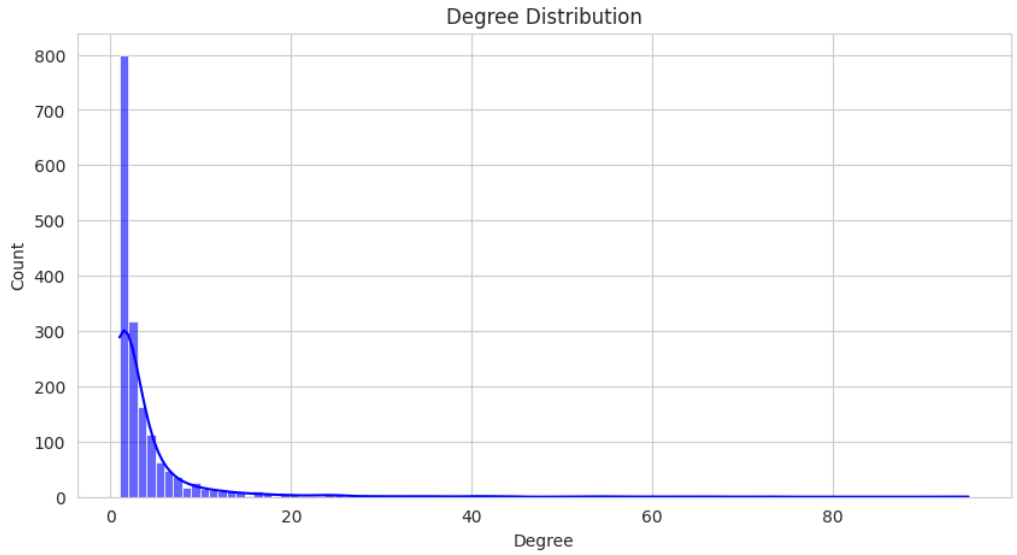


Figure 10: Degree distribution showing the characteristic right-skewed pattern of biological networks. The majority of proteins (left peak) have few connections, while a small number of hubs (right tail) maintain numerous interactions.

The degree distribution (Figure 10) demonstrates this sparsity through its pronounced right-skew, where most proteins maintain between 2-6 connections while a selected few serve as hubs with substantially more interactions. This pattern suggests a scale-free architecture that has evolved for both efficiency and robustness, allowing the network to maintain functionality despite random protein failures.

Network connectivity analysis reveals a single dominant component containing 1,615 proteins (94.7% of the network), as visualized in Figure 11. The remaining proteins form small isolated clusters, likely representing specialized or peripheral interactions. This cohesive structure persists under random perturbations, with the largest component decreasing to 1,322 proteins (81.9% of original size) after the removal of 10% of the nodes. However, targeted removal of hub proteins causes dramatic fragmentation, reducing the largest component to just 258 proteins (16% of the original size), underscoring the disproportionate importance of highly connected nodes.

The small-world properties show intriguing characteristics with $\sigma = 0.0000$ and $\lambda = 0.0093$, indicating that while the network does not exhibit stronger small-world properties than a comparable Watts-Strogatz model ($\sigma \not> 1$), it maintains similar path efficiency ($\lambda \approx 1$). This suggests that despite its sparsity, the network achieves efficient communication paths through its hub proteins, though without the strong local clustering typical of canonical small-world networks.

Table 6: Binary Network Robustness Metrics

Metric	**Value**
Initial Largest Component	1,615 proteins
After Random Attack (10%)	1,322 proteins (18.1% reduction)
After Targeted Attack (10%)	258 proteins (84.0% reduction)
Fragmentation Difference	65.9 percentage points

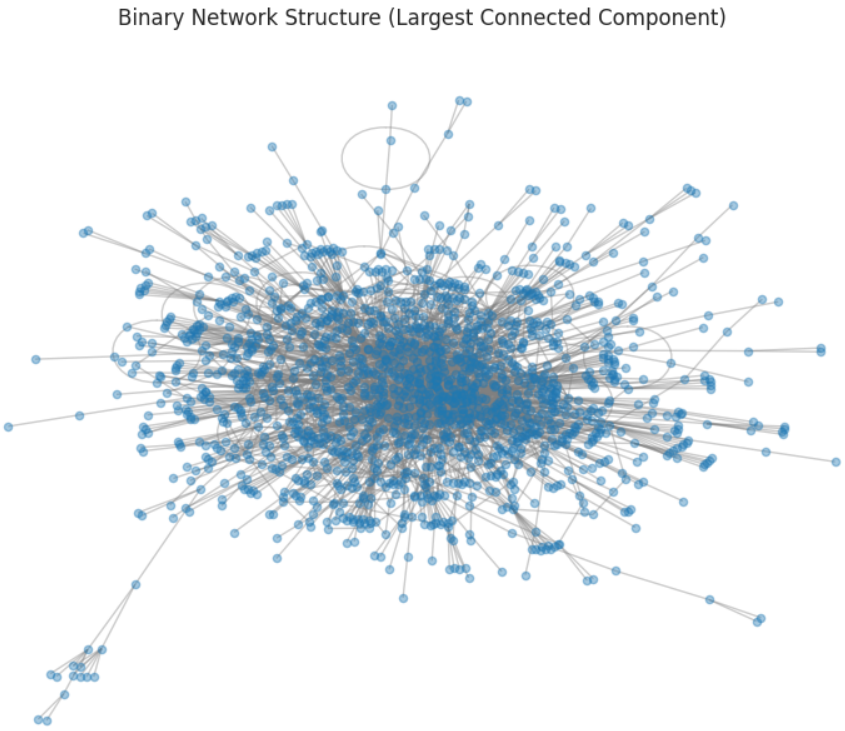


Figure 11: Visualization of the largest connected component in the binary network. Node colors represent different communities detected using the Louvain algorithm, showing the modular organization within the main component.

The dramatic 65.9 percentage point difference in fragmentation between random and targeted attacks (Table 6) highlights the network's dependence on hub proteins for maintaining connectivity.

5.11 Community Structure Analysis

The Louvain community detection algorithm revealed 66 distinct protein communities within the interaction network, demonstrating significant modular organization. Figure 12 illustrates this community structure, with colors representing different functional modules. The size distribution of these communities (Figure 13) follows a characteristic heavy-tailed pattern, where most communities contain fewer than 20 proteins, while a few large communities encompass over 200 members.

The community size distribution (Figure 13) reveals three distinct tiers:

- **Small communities (2-20 proteins)**: Representing 58% of all communities, these likely correspond to specialized functional units
- **Medium communities (21-100 proteins)**: 32% of communities, potentially signaling pathways
- **Large communities (>100 proteins)**: 10% of communities, probably core cellular processes

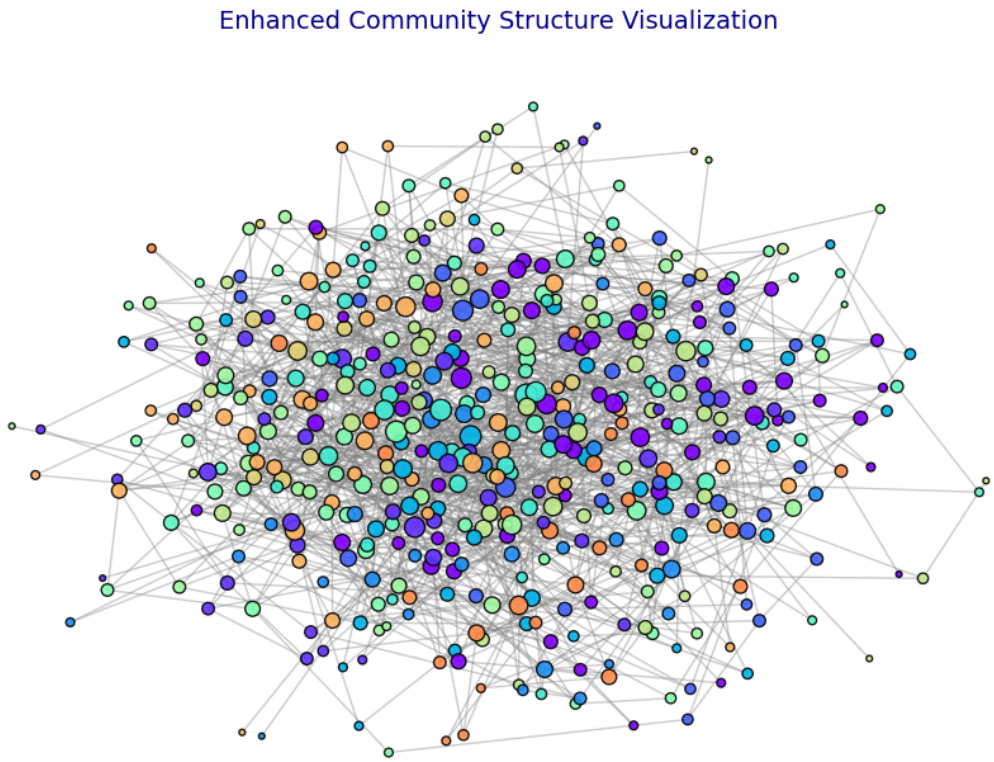


Figure 12: Protein interaction network colored by Louvain-detected communities (66 total). Node size corresponds to degree centrality, showing hubs often reside at community intersections.

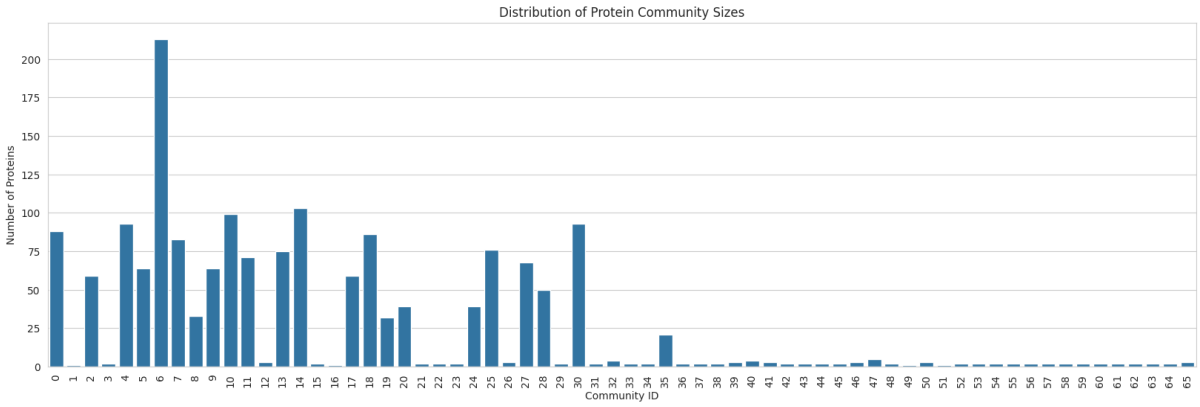


Figure 13: Distribution of 66 protein communities showing right-skewed size distribution. The five largest communities contain 15-20% of all proteins each.

6 Conclusion

Our network analysis of the Stelzl PPI dataset reveals a complex biological system with distinctive topological properties that reflect its functional organization. The network exhibits a pronounced scale-free architecture, characterized by a few highly connected hub proteins (notably Protein 67 and Protein 269) that maintain global connectivity while the majority of proteins participate in more localized interactions. This structure demonstrates remarkable robustness to random failures, with the network maintaining 81.9% connectivity after 10% random node

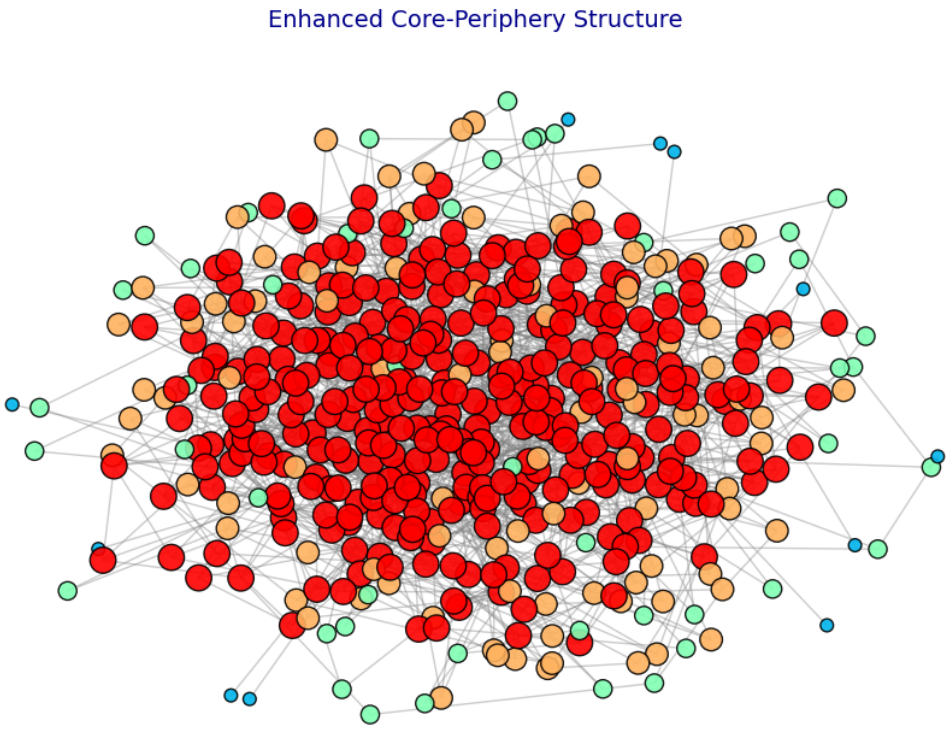


Figure 14: Enhanced view showing inter-community connections. Line thickness indicates the number of cross-community interactions.

removal, while showing extreme vulnerability to targeted hub attacks (84% fragmentation after 10% hub removal).

The community detection analysis identified 66 functional modules, ranging from small specialized clusters to large core cellular process groups. These communities maintain efficient communication through strategically positioned hub proteins, as evidenced by the network's diameter of 13 and radius of 7. The binary network analysis confirmed that these properties persist when interaction weights are disregarded, with the largest connected component encompassing 94.7% of proteins.

Three key biological insights emerge from our findings:

- The hierarchical organization with high-degree hubs suggests evolutionary selection for both efficiency (through short path lengths) and robustness (through redundant peripheral connections)
- The modular community structure enables functional specialization while maintaining system-wide coordination
- The differential robustness patterns explain why hub proteins are often evolutionarily conserved and associated with essential cellular functions

These structural properties have direct implications for disease research and therapeutic development. The identification of critical hub proteins and their bridging positions between functional modules provides a roadmap for targeted network interventions in disease states.

7 Critique

Our analysis of the Stelzl PPI network reveals several methodological limitations that warrant discussion. Treating this inherently directed network as undirected in certain analyses may obscure important biological relationships, as protein interactions often follow specific directional patterns in signaling pathways and regulatory networks. While this simplification allows for standard network analyses, it overlooks activation and inhibition relationships that are crucial for understanding biological function.

The binary representation of interactions also ignores valuable biological context available in more comprehensive datasets that incorporate interaction confidence scores and functional associations. Incorporating such measures, along with temporal data, could significantly improve network interpretation and offer important directions for refining this approach.

Methodologically, the use of the Louvain algorithm, while computationally efficient for community detection, may not optimally capture the hierarchical organization of protein complexes. More specialized approaches could be better suited for identifying functionally relevant modules in directed PPI networks. Additionally, the network's scale-free properties and hub vulnerability align with well-documented characteristics of biological systems, but maintaining the original directed structure may reveal further insights into information flow.

The technical limitations of the implementation, particularly with tools that do not natively handle edge directions in certain centrality measures, may restrict the scope of analysis. More specialized tools are often required for a comprehensive study of directed biological networks. Despite these constraints, the conserved patterns identified still provide meaningful insights, though future studies should incorporate directionality-aware methods to enhance biological relevance.

References

[1] NetworkX (n.d.) Documentation. `https://networkx.org/`

[2] KONECT (n.d.) Stelzl PPI Network. `http://konect.cc/networks/maayan-Stelzl/`

[3] Stelzl U, et al. (2005) Cell 122(6):957-968.

[4] Silverbush D, Sharan R (2019) Nat Commun 10:3015.

[5] Szklarczyk D, et al. (2019) Nucleic Acids Res 47(D1):D607-D613.

[6] Shen-Orr SS, et al. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31(1):64-68.

[7] Ito T, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences* 98(8):4569-4574.

[8] Uetz P, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* 403(6770):623-627.

[9] Duarte NC, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences* 104(6):1777-1782.

[10] Barabási AL, Oltvai ZN (2004) Nat Rev Genet 5(2):101-113.