Analysis using Gephi on Biological Networks

Tharun Subramanian C

RA2211003011187 – B2

1. Dataset Overview

Description: After analyzing the dataset, it was confirmed that the network represents biological interactions, likely within a yeast model system.

Format: The data was provided in `.gexf` format, a widely recognized format for exchanging graph structures, making it suitable for detailed analysis in tools like Gephi.

2. Degree Analysis

Degree Distribution: The degree of a node, representing the number of connections (edges) it has within the network, was calculated. The analysis showed that the degree varies across the network, with certain nodes displaying a high degree, indicating their significant connectivity and potential importance in the network structure.

3. Node Centrality

Centrality Measures: Several centrality measures were computed to identify the most influential nodes within the network:

> **Degree Centrality**: Nodes with the highest degree centrality were identified as critical hubs, having numerous connections within the network.
>
> **Betweenness Centrality**: Key nodes with high betweenness centrality were found to act as bridges, frequently appearing on the shortest paths between other nodes.
>
> **Closeness Centrality**: Nodes with high closeness centrality were highlighted for their efficiency in interacting with all other nodes in the network.
>
> **Eigenvector Centrality**: Influential nodes, identified through eigenvector centrality, were not only highly connected but also connected to other highly influential nodes.

4. Visualization and Analytical Steps

Dataset Loading: The `.gexf` file was successfully imported into Gephi for visualization and analysis.

Network Overview: The initial visualization provided a clear understanding of the overall structure, revealing key features of the biological network.

Degree Distribution Analysis: The degree distribution was analyzed to assess the spread of connectivity across the network. The results indicated a mix of highly connected hubs and sparsely connected nodes, suggesting a non-random network structure.

Centrality Visualization: Centrality measures were visualized, allowing for the identification of nodes that play crucial roles in the network's structure and function.

Modularity Analysis: Community detection using the modularity algorithm uncovered well-defined sub-communities within the network, highlighting areas of strong intra-community connections.

Visualization Refinement: The network visualization was refined by adjusting node sizes, colors, and layout, effectively emphasizing nodes with high degrees and centrality.

5. Insights from Data Laboratory

Highest Modularity: The community with the highest modularity score was identified, indicating the strongest and most distinct community structure within the network.

Largest Components: Nodes belonging to the largest connected components were analyzed, revealing their essential role in maintaining the network's overall connectivity and robustness.

6. Inference from Visualization

Degree Distribution: The analysis of node degrees confirmed that the network likely follows a scale-free model, characterized by a few highly connected hubs and many nodes with fewer connections.

Central Nodes: Nodes identified with high centrality measures were recognized as critical to the network's connectivity and influence, potentially representing key biological entities.

Modularity: The high modularity observed suggests that the network is organized into well-defined communities, which may correspond to functional clusters in the biological context.

Component Analysis: The analysis of the largest components underscored their importance in preserving the network's integrity, emphasizing their potential biological significance.