# Analysis using Gephi on Biological Networks

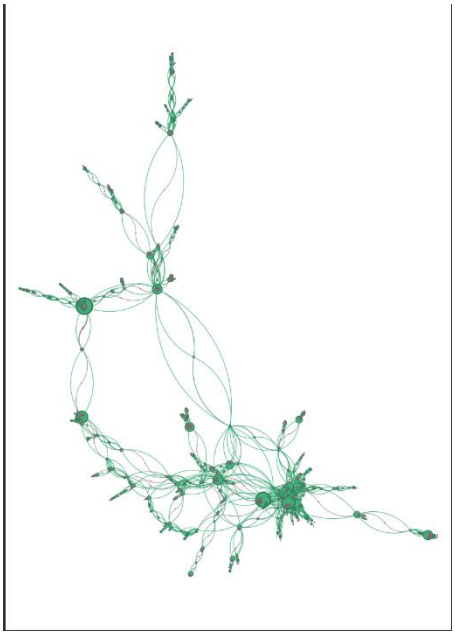
**Tharun Subramanian C**

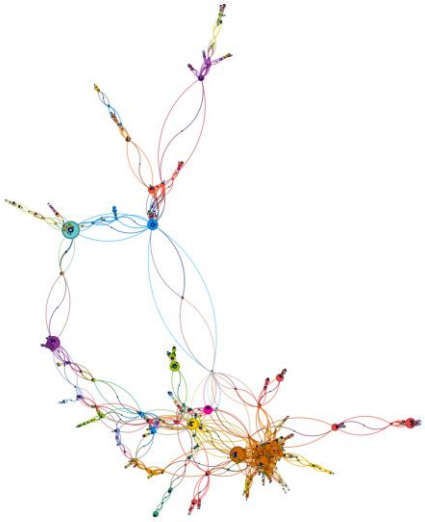
**RA2211003011187 – B2**

**BTech CSE CORE**

## 1. Dataset Overview

**Description**: GEXF. Diseasome: A network of disorders and disease genes linked by known disorder–gene associations, indicating the common genetic origin of many diseases. Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules. The original dataset can be found here: The Human Disease Network, Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007), Proc Natl Acad Sci USA 104:8685-8690



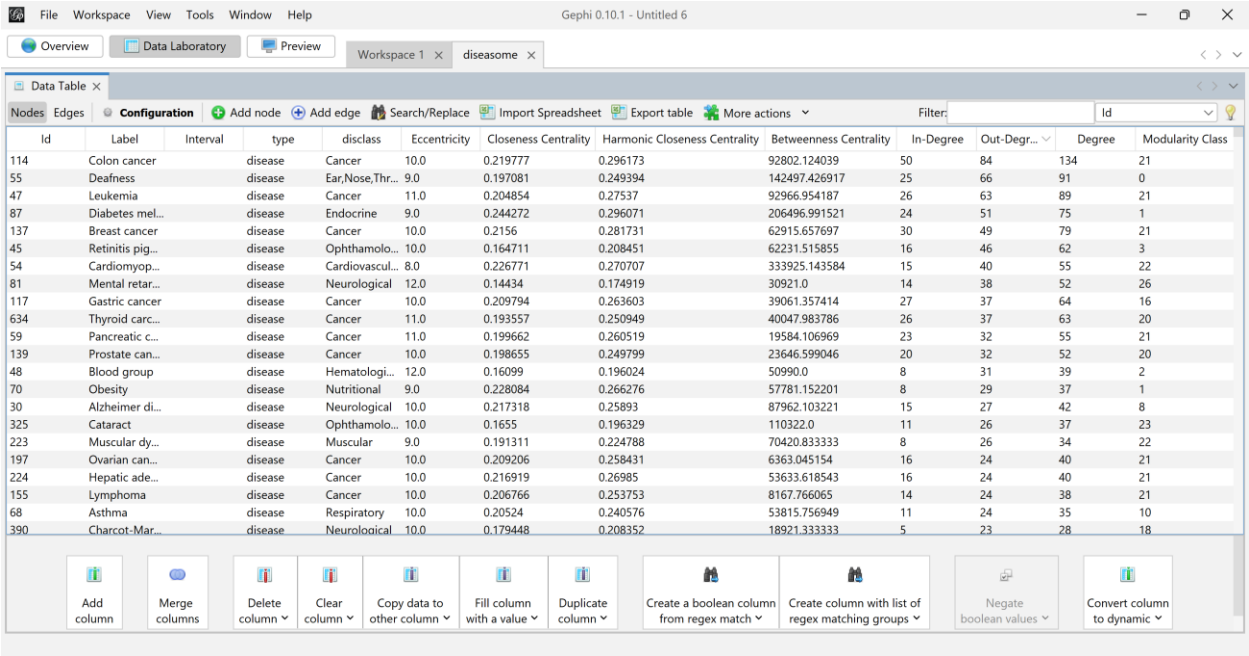


**Format**: The data was provided in `.gexf` format, a widely recognized format for exchanging graph structures, making it suitable for detailed analysis in tools like Gephi.

## 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

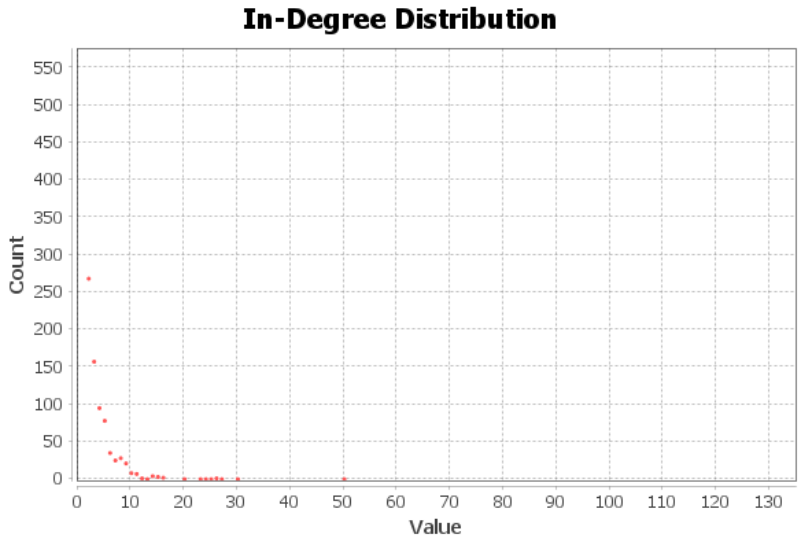
nodes displaying a high degree, indicating their significant connectivity and potential importance in the network structure.

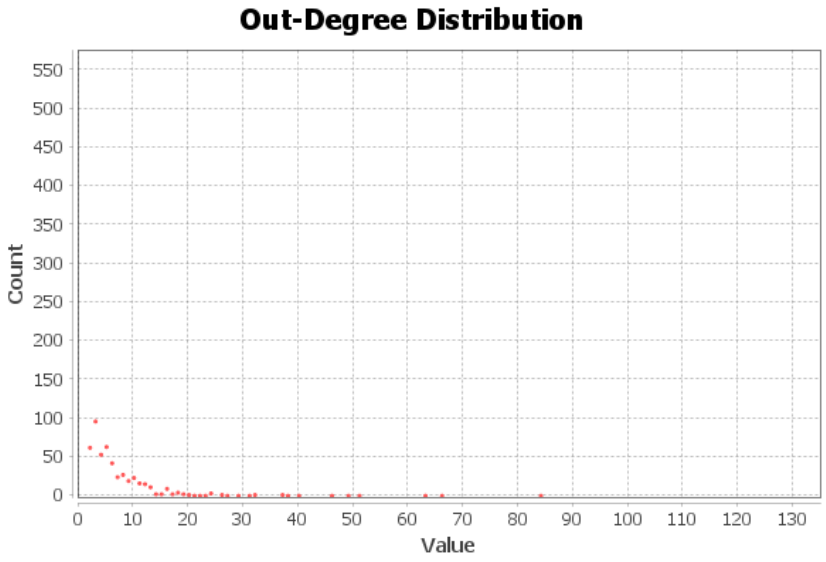


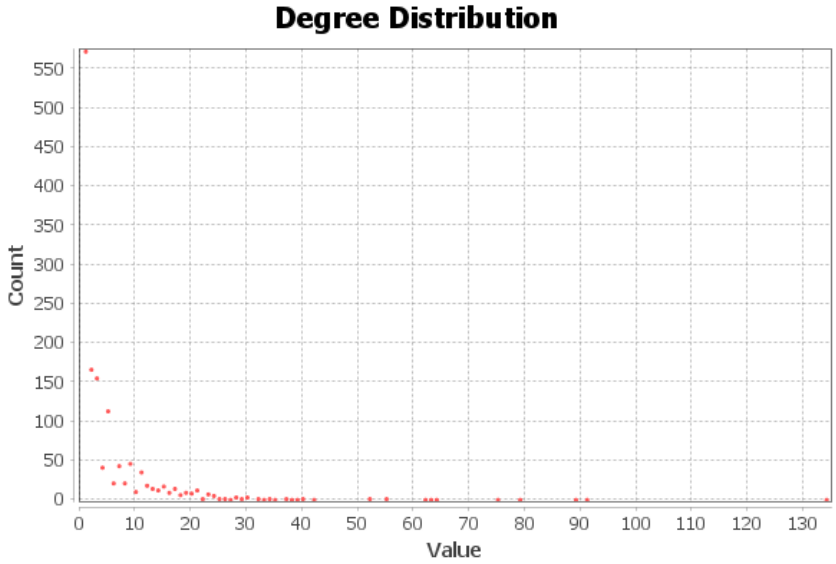

| Id | Label | Interval | type | disclass | Eccentricity | Closeness Centrality | Harmonic Closeness Centrality | Betweenness Centrality | In-Degree | Out-Degr... | Degree | Modularity Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 114 | Colon cancer | | disease | Cancer | 10.0 | 0.219777 | 0.296173 | 92802.124039 | 50 | 84 | 134 | 21 |
| 55 | Deafness | | disease | Ear,Nose,Thr... | 9.0 | 0.197081 | 0.249394 | 142497.426917 | 25 | 66 | 91 | 0 |
| 47 | Leukemia | | disease | Cancer | 11.0 | 0.204854 | 0.27537 | 92966.954187 | 26 | 63 | 89 | 21 |
| 87 | Diabetes mel... | | disease | Endocrine | 9.0 | 0.244272 | 0.296071 | 206496.991521 | 24 | 51 | 75 | 1 |
| 137 | Breast cancer | | disease | Cancer | 10.0 | 0.2156 | 0.281731 | 62915.657697 | 30 | 49 | 79 | 21 |
| 45 | Retinitis pig... | | disease | Ophthamolo... | 10.0 | 0.164711 | 0.208451 | 62231.515855 | 16 | 46 | 62 | 3 |
| 54 | Cardiomyop... | | disease | Cardiovascul... | 8.0 | 0.226771 | 0.270707 | 333925.143584 | 15 | 40 | 55 | 22 |
| 81 | Mental retar... | | disease | Neurological | 12.0 | 0.14434 | 0.174919 | 30921.0 | 14 | 38 | 52 | 26 |
| 117 | Gastric cancer | | disease | Cancer | 10.0 | 0.209794 | 0.263603 | 39061.357414 | 27 | 37 | 64 | 16 |
| 634 | Thyroid carc... | | disease | Cancer | 11.0 | 0.193557 | 0.250949 | 40047.983786 | 26 | 37 | 63 | 20 |
| 59 | Pancreatic c... | | disease | Cancer | 11.0 | 0.199662 | 0.260519 | 19584.106969 | 23 | 32 | 55 | 21 |
| 139 | Prostate can... | | disease | Cancer | 10.0 | 0.198655 | 0.249799 | 23646.599046 | 20 | 32 | 52 | 20 |
| 48 | Blood group | | disease | Hematologi... | 12.0 | 0.16099 | 0.196024 | 50990.0 | 8 | 31 | 39 | 2 |
| 70 | Obesity | | disease | Nutritional | 9.0 | 0.228084 | 0.266276 | 57781.152201 | 8 | 29 | 37 | 1 |
| 30 | Alzheimer di... | | disease | Neurological | 10.0 | 0.217318 | 0.25893 | 87962.103221 | 15 | 27 | 42 | 8 |
| 325 | Cataract | | disease | Ophthamolo... | 10.0 | 0.1655 | 0.196329 | 110322.0 | 11 | 26 | 37 | 23 |
| 223 | Muscular dy... | | disease | Muscular | 9.0 | 0.191311 | 0.224788 | 70420.833333 | 8 | 26 | 34 | 22 |
| 197 | Ovarian can... | | disease | Cancer | 10.0 | 0.209206 | 0.258431 | 6363.045154 | 16 | 24 | 40 | 21 |
| 224 | Hepatic ade... | | disease | Cancer | 10.0 | 0.216919 | 0.26985 | 53633.618543 | 16 | 24 | 40 | 21 |
| 155 | Lymphoma | | disease | Cancer | 10.0 | 0.206766 | 0.253753 | 8167.766065 | 14 | 24 | 38 | 21 |
| 68 | Asthma | | disease | Respiratory | 10.0 | 0.20524 | 0.240576 | 53815.756949 | 11 | 24 | 35 | 10 |
| 390 | Charcot-Mar... | | disease | Neurological | 10.0 | 0.179448 | 0.208352 | 18921.333333 | 5 | 23 | 28 | 18 |



From the analysis and the provided screen-shot we can see conclude that ***Colon cancer*** has the highest In-Degree(50) and Out-Degree(84) with a degree of 134.

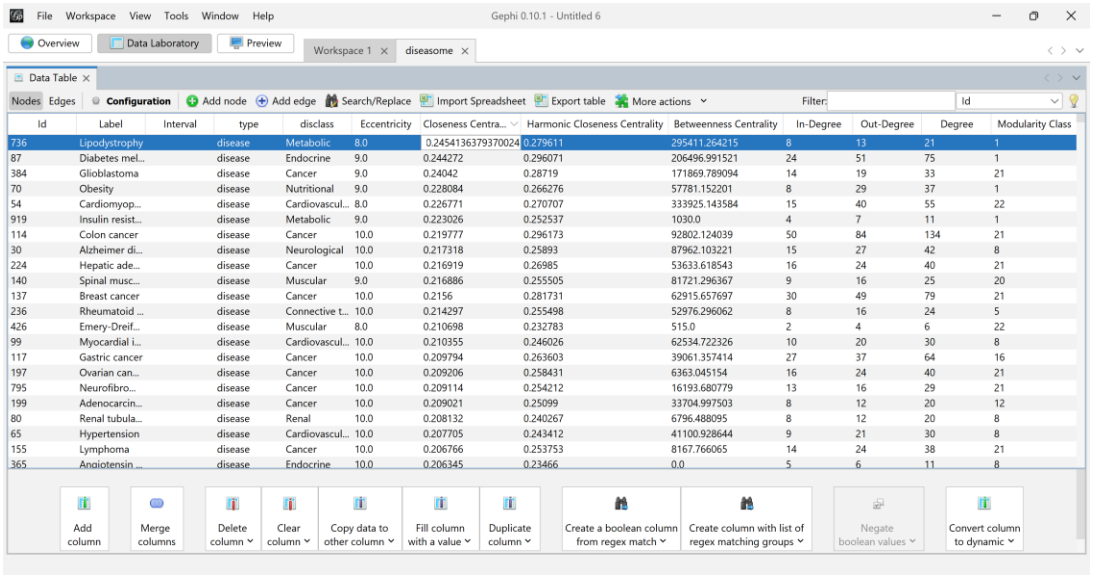




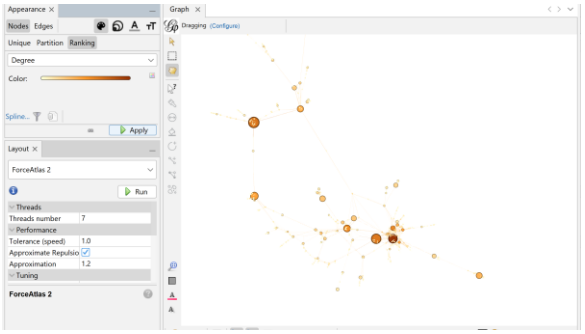




### 3. Node Centrality

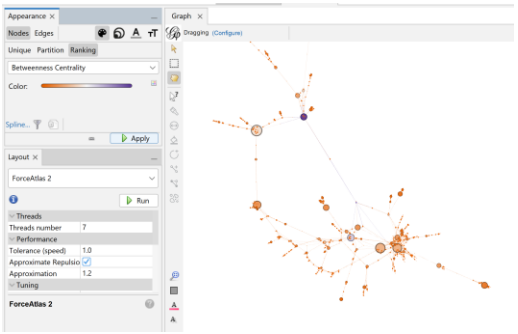
**Centrality Measures**: Several centrality measures were computed to identify the most influential nodes within the network:



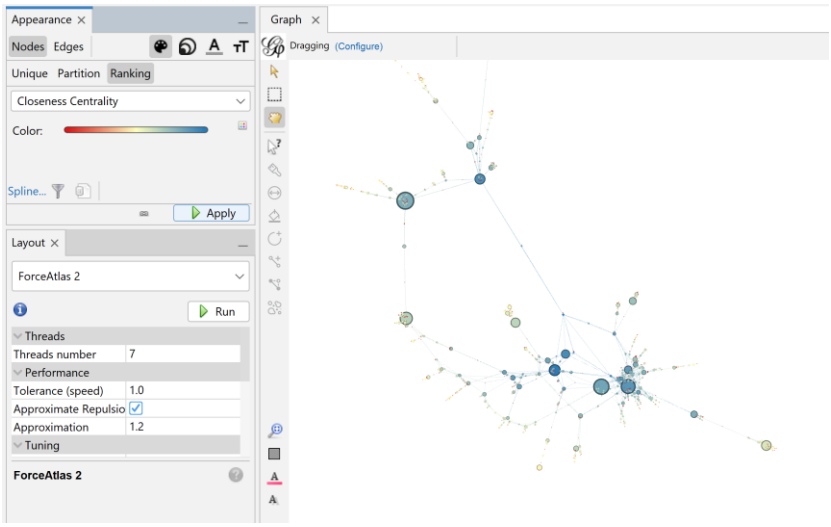

**Degree Centrality**: Nodes with the highest degree centrality were identified as critical hubs, having numerous connections within the network.




**Betweenness Centrality**: Key nodes with high betweenness centrality were found to act as bridges, frequently appearing on the shortest paths between other nodes.

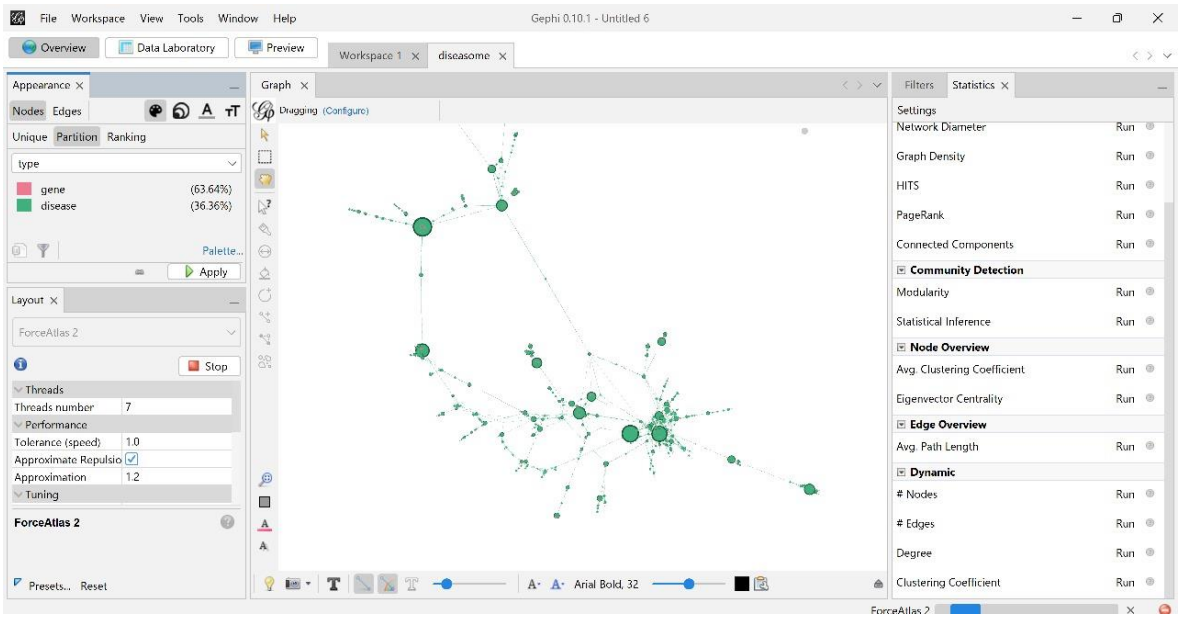



**Closeness Centrality**: Nodes with high closeness centrality were highlighted for their efficiency in interacting with all other nodes in the network.




## 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.

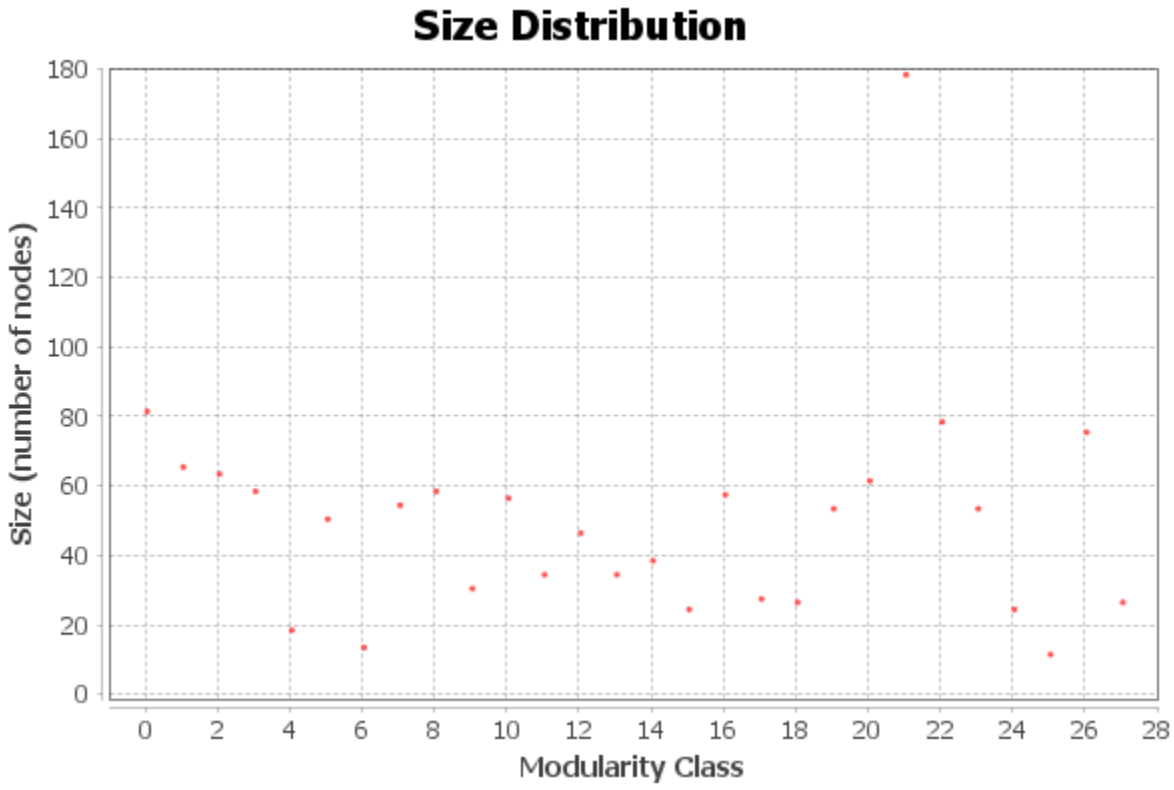
**Centrality Visualization**: Centrality measures were visualized, allowing for the identification of nodes that play crucial roles in the network's structure and function.

**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

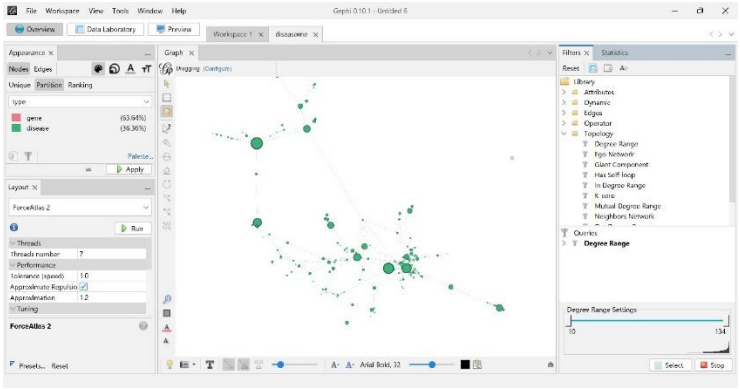
**Highest Modularity**: The community with the highest modularity score was identified, indicating the strongest and most distinct community structure within the network.




**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.