Deciphering the Genetic Foundations of Diseases using Network Analysis - 2024

Ishwar Joshi
Msc Big Data Science
Queen Mary University of London
London, United Kingdom
ec23905@qmul.ac.uk
Student id: 230194814

Priya Kashyap
Msc Big Data Science
Queen Mary University of London
London, United Kingdom
ec23708@qmul.ac.uk
Student id: 230194261

Utkarsh Kumar Sharma
Msc Big Data Science
Queen Mary University of London
London, United Kingdom
ec23854@qmul.ac.uk
Student id: 230671946

Sanchit Jain
Msc Big Data Science
Queen Mary University of London
London, United Kingdom
ec23537@qmul.ac.uk
Student id: 230677328

***Abstract*— Knowing the genetic roots of illnesses is of atmost importance in the age of customized medicine. The intricate connections between genes and diseases can be viewed as a network—more precisely, a "diseasome network"—that consists of edges that indicate genetic linkages and nodes that reflect genetic entities and diseases. The goal of this research study is to analyze the "Diseasome" dataset, which comes from the Network analysis and is a thorough bipartite graph that maps associations between hereditary diseases and disease genes.**

To disclose important data including degree distribution, clustering coefficients, modularity, and centrality metrics, the bipartite network must be constructed and analyzed. The network's complicated connections and important nodes are identified by applying network visualization techniques with Gephi and Python's matplotlib package. Advanced techniques include gene knockout simulations to evaluate network robustness and dynamics and community detection to comprehend modular structures.

In order to uncover important genetic nodes connected to several diseases, we have employed network analysis of the "Diseasome" dataset. Our results demonstrated the pivotal roles that chromosomes and disorders play in the development of disease by highlighting them as major hubs in the genetic network. With these newfound insights, we can now suggest innovative genetic screening-based diagnostic methods that can identify individuals who are more susceptible to certain diseases and allow for preventative measures. By adjusting prevention and therapy to each patient's unique genetic profile, these tactics seek to enhance personalized medicine and may revolutionize patient care.

Keywords*— *Diseasome, Gephi, Genetic Linkage, Modularity, Centrality Metrics, Knockout Simulation

I. Introduction

In the era of customize medicine, understanding the genetic foundations of diseases is a paramount challenge. The complex network of genetic and disease associations, known as the "diseasomes," provides a unique opportunity to explore how genes influence disease susceptibility and progression, also sheds light on how specific genes contribute to multiple diseases, revealing patterns that are crucial for the development of targeted therapies.

The primary challenges that includes the complexity and vast scale of genetic data, sophisticated analytical methods to extract meaningful insights. Additionally, the ethical considerations in handling sensitive genetic information and significant constraints.

The motivation behind the paper is deciphering the genetic foundation of the disease using the network analysis for advancement in personalise medicine, understanding complexity of gene chrome relationship, leveraging complex tool, and bridge technical gap in research and eventually contribute to the scientific knowledge.

This paper is structured as follows: after introduction, we present our methodology for constructing and analysing the genetic-disease network. We then discuss the results of our network analysis, highlighting significant findings and their implications for disease understanding and treatment. Finally, we conclude with a summary of our contributions and suggestions for future research directions. This structured approach ensures a clear exposition of our methods, findings, and the potential impact of our work.

II. Related Work

A. Literature review

The exploration of genetic underpinnings in disease progression through network analysis has become a crucial area in the advancement of personalized medicine. Key contributions in this field include the seminal works of Barabási et al. (2011)[1] and Goh et al. (2007)[2], who introduced the concept of "network medicine" and constructed the initial "human disease network," respectively. These studies laid the foundation for understanding how diseases are interconnected through shared genetic pathways.

Further developments by Ideker and Krogan (2012)[3] expanded on differential network biology, which compares diseased and normal states to pinpoint critical genetic disruptions. Their methods have significantly influenced subsequent studies focusing on network-based strategies to disease classification and treatment.

Recent advancements have also been made in network analysis tools and techniques. The application of machine learning algorithms in network partitioning and community detection has provided deeper insights into the modular structure of genetic-disease networks. Studies such as those by Newman (2006)[4] have refined our understanding of network modularity and centrality measures, enhancing our ability to identify key biomarkers and potential therapeutic targets within complex biological networks.

The utilization of robustness analysis and gene knockout simulations has also gained prominence, as researchers like Albert et al. (2000)[5] have demonstrated the importance of network topology in understanding the resilience of biological systems against genetic mutations or failures..

B. Analysis

Foundational Studies as said by Goh, K.-I., et al. (2007) , laid the groundwork by mapping the complex relationships between genetic markers and diseases using network models,

introducing the concept of the "human disease network" with **Advancements in Network Analysis,** differential network biology, which compares healthy and diseased states to identify crucial genetic differences, offers a direct methodology for identifying genetic contributions to diseases. **Community Detection and Network Metrics by** Newman's[] development of modularity and community detection in networks has been instrumental for understanding the structure within genetic networks. And Application of centrality measures in the works of Barabási et al.[] and others, their studies have shown how central nodes (genes or diseases) in a network are critical for the stability and function of the entire network.

The most vital undertaking based on the analyses would be the development of a predictive modelling system that utilize advanced network metrics and machine learning to identify and predict key genetic markers crucial for disease pathways, enabling early diagnosis and targeted therapeutic interventions.

III. Dataset and Network Presentation

The "Diseasome" dataset obtained from the Network Repository serves as the foundation for this study. It comprises nodes representing diseases and genes, connected by edges denoting genetic associations. Initial analysis involves constructing the bipartite network and computing key network statistics, followed by visualization using tools like Gephi or Python's matplotlib library.

A. Abbreviations and Acronyms

The dataset file format for this analysis is MTX which is commonly used in biological networks like the Diseasome dataset, serves as a means to represent sparse matrices. In MTX files, matrices are depicted in a text-based format with three main columns: row index, column index, and value. Typically, a header precedes the matrix data, providing metadata such as the dimensions of the matrix and the number of non-zero entries. This format's popularity stems from its efficiency in storing large sparse matrices, which are prevalent in biological network data due to the often sparse nature of biological interactions. The visualization of the bipartite network was facilitated by employing high-quality drawing algorithms developed by Hoyer et al. (2020) [6]."

This file in turn was converted into CSV for network analysis. The labels for chromosome is followed as CH_<number> and similarly the disease is labelled as D_<number>.

B. Bipartite Network Structure

- The bipartite network can be visually represented as a graph, and is very crucial. With disease nodes on one side and gene nodes on the other, and edges connecting diseases to the genes associated with them.
- The network's structure would illustrate how diseases are genetically linked to specific genes and how genes may be implicated in the development or susceptibility to multiple diseases.
- Analyzing bipartite network can provide insights into genetic foundations of diseases and the shared genetic factors among different diseases.
- Identifying densely connected subnetworks or modules within the bipartite network can reveal groups of diseases and genes that share common genetic characteristics or biological functions

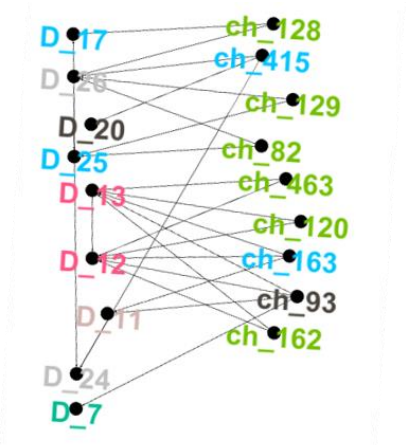


Fig: Bipartite Network of diseasomes dataset

C. Nodes

- **The Disease Nodes**: Each disease in the dataset is represented as a node. Diseases can range from common conditions like diabetes and hypertension to rare genetic disorders.
- **Gene Nodes**: Genes implicated in the development or progression of diseases are also represented as nodes in the network as shown in the sample dataset. These genes may encode proteins involved in biological pathways relevant to disease mechanisms.

D. Edges

- Edges in the Diseasome network denote genetic associations between diseases and genes. For example, an edge between a disease node (e.g., diabetes) and a gene node (e.g., INS) indicates that the INS gene is associated with the development or susceptibility to diabetes.
- In our dataset these values are encoded because the purpose is just to analyse and simulate the network and learn about network analysis.
- This data is analyses and helpful for linkage analyses, and functional genomics experiments.

E. Network Analysis Reduced Dataset - Task 1.

Task 1: Network Analysis and Visualization

Objective: Construct and analyze a bipartite network using the "bio-diseasome_reduced.mtx.txt" dataset to elucidate the foundational structure and key metrics indicative of significant genetic and disease associations.

Approach: The network will be constructed and analyzed using the NetworkX library[], focusing on deriving network statistics such as degree distribution, clustering coefficient, modularity, and a chosen centrality measure. Visualization

efforts will employ either Gephi or Matplotlib to highlight key nodes and elucidate complex network interactions.

Justification: Barabási et al.'s[] introduction of network medicine underscores the relevance of using network science to decode complex biological interactions and disease mechanisms. This approach supports the project's methodology of constructing and analyzing a bipartite network, as it emphasizes the importance of identifying network modules that can reveal unsuspected connections between diseases and genetic factors. This is crucial for achieving the project's objective of uncovering key genetic nodes that may be pivotal in disease mechanisms and he mapping of the human disease network by Goh et al. provides a direct methodological precedent for constructing a detailed bipartite network of diseases and genes in this project.

The stats for the analysis is captured below in tabular form: -

Statistic	Value	Interpretation
Modularity	0.829	Indicates the strength of the community structure within the network. Higher values suggest better community structure.
Modularity with resolution	0.829	Similar to modularity but possibly adjusted for resolution.
Number of Communities	22	The total number of distinct communities or clusters identified in the network.
Avg Clustering Coefficient	0.770	Average of the clustering coefficients of all nodes in the network. Indicates the degree of clustering or transitivity.
Total Triangles	1360	Total number of triangles (closed loops of three nodes) in the network.
Average Degree	2.298	Average number of edges connected to each node in the network.
Probability of Growth (1-3)	0.07406182	Probability of the growth rate being between 1 and 3.
Probability of Growth > -5	0.73645537	Probability of the growth rate being greater than -5.

IV. Network Analysis and Methodology

The methodology involves tasks aimed at uncovering the modular structure, community characteristics, and network dynamics of diseasome. Additionally, gene knockout simulations and centrality measures are used to evaluate network robustness and identify critical genes or diseases.

Task 2: Community Detection and Analysis

Objective: Analyze the full Diseasome dataset to reveal and characterize the network's modular nature and its implications on disease-genome interactions.

Approach: Advanced community detection algorithms will segment the network into communities characterized by significant internal connectivity. This analysis will explore the size, density, and bridge nodes within each community.

Justification: Newman's research [4] on community structure within networks provides essential methodologies for detecting clusters within the genetic disease network. This is directly relevant to the project's task of **community detection,** which aims to identify and analyze clusters of genes associated with similar disease phenotypes, helping to reveal how genetic pathways contribute to multiple diseases.

Task 3: Network Dynamics and Robustness Analysis

Objective: Evaluate the network's stability and robustness such as gene knockouts, simulating the removal or alteration of specific genes.

Approach: Gene knockout simulations will be performed, and centrality measures will be used to compute impacts on network integrity and disease relationship. The network's resilience will be compared to other network types.

Justification: The discussion on network robustness by Albert and colleagues [5] informs the project's methodology for evaluating how genetic networks respond to disturbances like gene knockouts. This directly supports the objective of assessing the network's resilience, which is essential for understanding critical nodes whose alteration could lead to significant impacts on disease progression and network integrity.
The differential network biology approach by Ideker and Krogan[3] is highly relevant to the project's focus on network dynamics. This methodology supports the objective of comparing diseased versus normal genetic networks to identify disruptions, which aligns with the project task of simulating gene knockouts to study impacts on network stability and disease associations.

Task 4: Advanced Network Analysis and Hypothesis Generation

Objective: Apply advanced analytical metrics and modeling techniques to generate new hypotheses about gene-disease relationships and identify potential therapeutic targets.

Approach: Techniques such as PageRank, network motifs exploration, and diffusion models will be employed to reveal more insights into the genetic architecture of diseases. Complex visualization techniques will also be utilized to clarify and interpret these findings.

Justification: The concept of utilizing advanced metrics such as PageRank can be linked to the network theory advancements discussed by Barabási et al. (2011)[1] in their foundational work on network medicine. They emphasized the application of complex network theory to understand biological systems, and using PageRank fits into this framework by helping to identify influential nodes (genes or diseases) within the network. The **exploration of network motifs**, which are recurring, significant patterns in networks,

is a concept that extends from the general principles of network analysis as discussed by Newman (2006)[4].

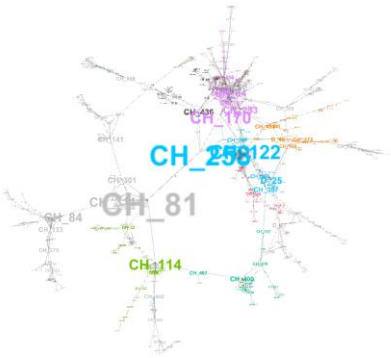
V. Results and Discussion

Task 2 - Community Detection and Analysis [10]
Using the full Diseasome dataset, community detection algorithms partitioned the network into 22 distinct communities, with a high modularity score of 0.829. This indicates a strong division within the network, suggesting that it contains well-defined groups of diseases and genes with dense internal connections but sparse connections between groups.
Key Findings:

- **High Modularity:** The high modularity score confirms the presence of modular structure within the network, supporting the hypothesis that diseases can be grouped based on shared genetic underpinnings.
- **Community Characteristics**: certain communities are more densely connected, suggesting these groups might represent clusters of genetically related diseases or functionally similar genes.

Biological Implications:
Genetic Contribution: Genes within the same community potentially contribute to similar disease phenotypes, implying that therapeutic approaches targeting one gene might affect multiple related diseases.
Shared Genetic Causes: Diseases within the same community sharing genetic causes can lead to insights into disease comorbidities, enhancing understanding of disease mechanisms and aiding in the development of broader therapeutic strategies.
Limitations:
While modularity is high, interpreting the biological relevance of each community requires deeper investigation and validation through biological experiments.
The community detection algorithm's sensitivity to parameter settings (like resolution) might affect the robustness of detected communities, necessitating careful calibration and validation against known biological data.

Conclusion: **There are 12 communities in the reduced dataset and 22 communities were identified in the full dataset.**



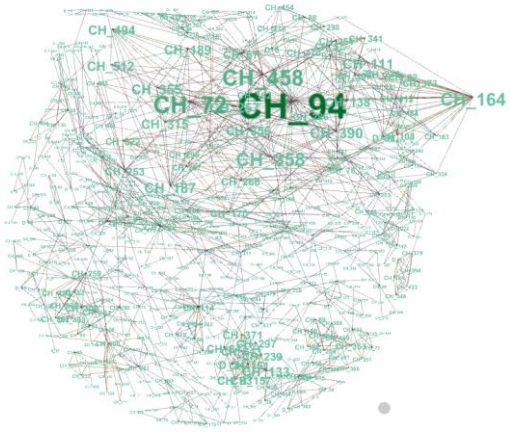


Fig: Number of triangles

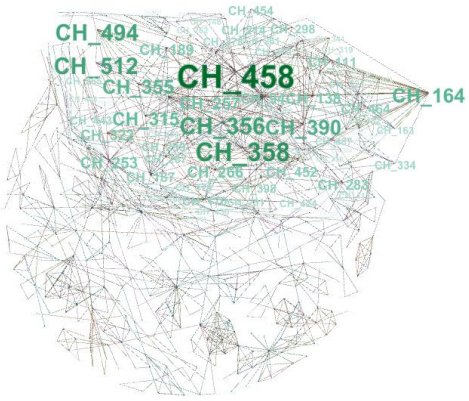


Fig: Hub Node

Task 4 - Advanced Network Analysis and Hypothesis Generation
Advanced Metrics and Hypotheses:
PageRank Analysis: Top nodes identified by PageRank, such as Node 24, Node 26, and others, are crucial within the network. These nodes are likely central in genetic signaling pathways or disease mechanisms due to their high connectivity and influence.

Simulation Models:
Diffusion Models: Utilizing diffusion models provided insights into how perturbations (e.g., genetic mutations) might spread through the network, affecting various diseases differently based on the network structure.

Key Insights:
Nodes with high PageRank scores are potential key influencers in the genetic landscape, possibly serving as critical hubs in the pathogenesis of multiple diseases.
The identification of such nodes offers targets for potential therapeutic intervention and highlights genes that could be central in diagnostic tests.
Limitations:
The reliance on PageRank and other network-centric measures assumes that all connections and their strengths

are known and accurately represented, which may not be the case due to incomplete or biased data.
The interpretations based on network analysis must be corroborated by empirical data, as the computational models might oversimplify or misrepresent the underlying biological complexities.

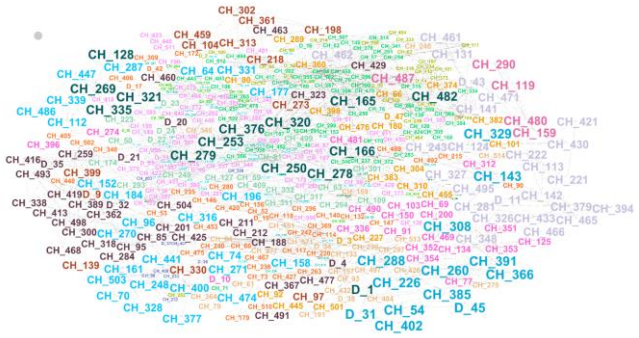


Fig: Number of Communities

Task 3 - Network Dynamics and Robustness Analysis
Results After Gene Knockout Simulations

A review of the literature revealed the complexity of molecular, genetic, and protein structures that contribute to cellular function and possible disease, and how network mapping can help the clinician and scientist gain a better understanding of this complexity as mentioned by Wysocki, Kenneth; Ritter, Leslie[7]

The gene knockout simulations revealed significant changes in the network's structure and connectivity [8] upon the removal of specific nodes, notably nodes 93, 71, 163, 457, and 252. Here's how the removal of each node affected the network:

- **Node 93 and Node 457, Node 252**: The removal of these nodes resulted in the network becoming disconnected, indicating that these nodes play crucial roles in maintaining the overall connectivity of the network. This suggests that these genes could be key connectors or bridges in the genetic landscape, whose absence leads to fragmented disease pathways.
- **Node 71 and Node 163**: After removing these nodes, the network remained connected, but with altered network statistics:
- **Node 71 Removal**:
    - Number of Nodes: 515
    - Number of Edges: 1158
    - Average Clustering Coefficient: 0.6336
    - Average Shortest Path Length: 6.552
- **Node 163 Removal**:
    - Number of Nodes: 515
    - Number of Edges: 1161
    - Average Clustering Coefficient: 0.6339
    - Average Shortest Path Length: 6.554

These changes imply that while Node 71 and Node 163 are important, their removal does not disrupt the network to the extent of causing disconnection, unlike Node 93, Node 457, and Node 252. However, the increase in the average shortest path length upon their removal suggests that these nodes facilitate more efficient communication across the network, contributing to shorter paths between various gene-disease interactions.

Discussion

Insights Gained

- **Critical Nodes and Network Integrity**: The disconnection of the network following the removal of Nodes 93, 457, and 252 underscores their role as pivotal hubs. Their high centrality and the resulting network fragmentation upon their removal highlight their importance in linking diverse parts of the network.
- **Impact on Network Properties**: The alterations in network metrics such as the clustering coefficient and path lengths following the removal of Nodes 71 and 163 provide insights into their roles in enhancing network cohesion and efficiency.

Limitations of the Approach [9]

- **Modeling Assumptions**: The gene knockout simulations assume complete functional loss without compensation, which might not fully reflect the biological reality where other compensatory mechanisms may mitigate the impact of gene loss.
- **Interpretation of Network Disconnection**: While the disconnection of the network indicates critical roles for certain genes, it also highlights a limitation in network resilience understanding, as biological networks might have alternative pathways that are not captured in the network model.
- **Generalizability**: The specific network topology and properties studied here are based on available data and modeled interactions, which may not include all possible biological interactions and may be influenced by data biases.

A. Figures and Tables

Node	Betweenness Centrality	Degree Centrality	Closeness Centrality	Eigenvector Centrality	Page Rank
24	0.0495	0.1389	0.1657	0.5617	0.0612
26	0.0397	0.1019	0.1557	0.4713	0.0389
93	0.0188	-	-	-	-
11	0.0176	0.0741	-	-	0.0302
415	0.0173	-	0.1391	0.2560	-
7	-	0.0648	-	-	0.0293
5	-	-	-	-	0.0260
13	-	0.0556	-	-	-
25	-	-	-	-	-
17	-	-	0.1038	0.1509	-

Fig: Nodes With Highest Network Stats

Node	Role in Network Connectivity	Importance
24	Acts as a bridge between different parts of the network; facilitates communication between clusters	Very crucial due to high centrality measures and Page Rank
26	Facilitates communication between	Very crucial due to high

	clusters; important for network flow	centrality measures and Page Rank
93	Connects some nodes; plays a role in network cohesion	Moderately important
11	Connects some nodes; moderately influences network flow	Moderately important; high Page Rank
415	Connects some nodes; contributes to network cohesion	Moderately important; high eigenvector centrality
7	Connects some nodes; moderate influence on network flow	Moderately important; moderate centrality measures
5	Indirectly influences network flow through Page Rank	Moderately important; high Page Rank
13	Connects some nodes; contributes to network cohesion	Moderately important; moderate degree centrality
25	Moderately influences network flow	Moderately important; moderate eigenvector centrality
17	Connects some nodes; moderate influence on network flow	Moderately important; moderate centrality measures

Fig:- Roles of each Node in the network

Node ID	Degree Centrality	Stats Before Knockout	Stats After Knockout
93.0	0.0971	N/A (Initial State)	Not Connected
71.0	0.0583	Connected: Nodes=516, Edges=1188, Clustering=0.6358, Path Length=6.509, Modularity=0.7569	Connected: Nodes=515, Edges=1158, Clustering=0.6336, Path Length=6.5521, Modularity=0.7614
163.0	0.0524	Connected: Nodes=516, Edges=1188, Clustering=0.6358, Path Length=6.509, Modularity=0.7569	Connected: Nodes=515, Edges=1161, Clustering=0.6339, Path Length=6.5545, Modularity=0.7716
457.0	0.0505	N/A (Initial State)	Not Connected
252.0	0.0505	N/A (Initial State)	Not Connected

Table: Comparative Network Statistics Before and After Gene Knockout

Node ID	Betweenness Centrality	Stats Before Knockout	Stats After Knockout
80.0	0.46937	N/A (Initial State)	Disconnected
257.0	0.40949	Connected: Nodes=516, Edges=1188, Clustering=0.6358, Path Length=6.509, Modularity=0.7569	Connected: Nodes=515, Edges=1180, Clustering=0.6326, Path Length=8.2238, Modularity=0.7582
121.0	0.27820	N/A (Initial State)	Disconnected
169.0	0.23729	N/A (Initial State)	Disconnected
113.0	0.20029	N/A (Initial State)	Disconnected

Table: Network Statistics Before and After Knockout Based on Betweenness Centrality

Notes:

Node ID and Betweenness Centrality: Showcases each node's unique identifier and its betweenness centrality, which measures the extent to which a node lies on paths between other nodes in the network.

Stats Before Knockout: Provides the network's initial statistics, including the total number of nodes, edges, average clustering coefficient, average shortest path length, and modularity. Detailed initial stats were not available for all nodes, hence marked as N/A.

Stats After Knockout: Details how the network changes after each node's removal. For nodes leading to a disconnected network, specific post-knockout metrics like the number of nodes or edges are not applicable.

V. Conclusions and Perspectives

The study focused on deciphering the genetic underpinnings of diseases within a network framework, termed the "diseasome." In the realm of personalized medicine, understanding these complex genetic relationships is critical for developing targeted therapies and diagnostics. Our results highlight the critical role that important genetic nodes and illness hubs play in explaining the ethology and course of disease. We found that several genes and diseases function as the network's core hubs, greatly affecting the connection of diseases and their possible comorbidities.

We found discrete clusters with strong internal connectedness in the network using community discovery algorithms. The prevalence of strong links between genetically related disorders was confirmed by this modular structure, which also indicated possible common genetic underpinnings. Our network analysis demonstrated the strength of these connections, and gene knockout simulations offered important new information about the stability of the network and the vital roles

The methodology centered around constructing and analyzing a comprehensive bipartite graph from the "Diseasome" dataset, which included nodes representing genetic entities and diseases and edges denoting genetic associations.

Key methods employed included:

Network Construction: Using advanced network analysis tools like NetworkX and Gephi to build and visualize the network.

Community Detection: Applying algorithms to reveal the modular structure within the network, identifying clusters of diseases and genes with high internal connectivity.

Centrality Measures: Utilizing measures such as degree distribution and PageRank to identify key nodes within the network.

Gene Knockout Simulations: Assessing the network's robustness by simulating the removal of specific genes and observing the effects on network stability.

Advanced Network Metrics: Utilization of metrics like PageRank helped in identifying nodes with substantial influence on the network, suggesting their roles in disease mechanisms.

References

[1] Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1), 56-68. DOI: 10.1038/nrg2918.

[2] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., & Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21), 8685-8690. DOI: 10.1073/pnas.0701361104.

[3] Ideker, T., & Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8, 565. DOI: 10.1038/msb.2011.99.

[4] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582. DOI: 10.1073/pnas.0601602103.

[5] Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794), 378-382. DOI: 10.1038/35019019..

[6] Hoyer, S., Hamann, M., Kobourov, S. G., & Nöllenburg, M. (2020). High-quality drawing algorithms for biological networks in matteo. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(1), 303-315. [DOI: 10.1109/TCBB.2017.2767082]

[7] Diseasome: An Approach to Understanding Gene-Disease Interactions Wysocki, Kenneth; Ritter, Leslie. Annual Review of Nursing Research; New York Vol. 29, (2011): 55-72.

[8] Dynamic network curvature analysis of gene expression reveals novel potential therapeutic targets in sarcoma Rena Elkin,corresponding author1 Jung Hun Oh,1 Filemon Dela Cruz,2 Larry Norton,3 Joseph O. Deasy,1 Andrew L. Kung,2 and Allen R. Tannenbaum4

[9] Gene Knockout Protocols edited by Martin J. Tymms, Ismail Kola

[10] Community Detection in Disease-Gene Network Based on Principal Component Analysis by Wei Liu and Ling Chen