Investigation of Vertex Centralities in Human Gene-Disease Networks

by

© *Seyed Mehrzad Almasi*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Computer Science*

Memorial University of Newfoundland

May 2018

St. John's Newfoundland

Abstract

Studying associations among genes and diseases provides an important avenue for a better understanding of genetic-related disorders, phenotypes and other complex diseases. Research has shown that many complex human diseases cannot be attributed to a particular gene, but a set of interacting genes. The effect of a specific gene on multiple diseases is called *pleiotropy* and interactions among several genes to contribute to a specific disease is called *epistasis*. In addition, many human genetic disorders and diseases are known to be related to each other through frequently observed co-occurrences. Studying the correlations among multiple diseases helps us better understand the common genetic background of diseases and develop new drugs that can treat them more effectively and avoid side effects. Meanwhile, network science has seen an increase in applications to model complex biological systems, and can be a powerful tool to elucidate the correlations of multiple human diseases as well as interactions among associated genes. In this thesis, known disease-gene associations are represented using a weighted bipartite network. Subsequently, two new networks are extracted. One is the weighted human disease network to show the correlations of diseases, and the other is the weighted gene network to capture the interactions among genes. We propose two new centrality measures for the weighted human disease network and the weighted gene network. We evaluate our centrality measurements and compare them with the most commonly used centralities in biological networks including degree, closeness, and betweenness. The results show that our new centrality methods can find more important vertices since the removal of the top-ranked vertices leads to a higher decline rate of the network efficiency. Our

identified key diseases and genes hold the potential of helping better understand the genetic background and etiologies of complex human diseases.

Acknowledgements

This work wouldnt been accomplished without constant help of my supervisors Dr. Ting Hu and Dr. Wolfgang Banzhaf. During all of these two years, I bombarded their with my never-ending questions and they patiently guided me through them all. Therefore, my first thanks goes to them. Finally, I thank graduate students and my colleagues in the department of computer science for participating in informative talks that provided useful learning insights.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Genome-wide association study (GWAS) is one of the most important tools for investigating the genetic architecture of genetic-related phenotypes and human diseases [2]. Scientists have proposed many methods based on SNP (Single Nucleotide Polymorphisms) GWAS [3, 4] where SNP refers to a variation in a single DNA building block, called a nucleotide. To find the SNPs that are associated with a disease, a standard GWAS should investigate large number of people with the disease (case) and large number of healthy people (control). GWAS includes two steps of discovery and validation. In the first step, by investigating the cases and controls, researchers search for SNPs that can best discriminate cases and controls. All SNPs are analyzed at this stage and are ranked according to their significance levels. All SNPs that meet a p-value threshold can go forward to the next step, validation [5].

Some diseases can be caused by a single mutation at a single gene. However, most complex diseases are likely caused by interactions among a set of different genes, called epistasis [6, 7]. Hence, finding the interactions among genes is critical for us in order to better understand the genetic mechanisms of diseases in the human body [8]. Therefore, many researchers have changed their focus from considering the susceptibility of a single locus [9] to the interactions among a set of loci [10]. However, the huge number of SNPs cause extreme computational costs and subsequently some restrictions on the use of SNP based methods [11].

Nowadays, the most common approach for identifying the interactions among SNPs is based on statistical methods [12, 13, 14, 15, 16, 17], data mining approaches [18, 19, 20], machine learning [21, 22], and other methods [23, 24]. Considering the interactions among genes is not the only way to find genetic origins of complex diseases, but the study of the correlation of diseases is a good tool for such a purpose.

During the past decades, significant progress has been made on our understanding of human diseases [25]. However, the genetic architectures of complex diseases are still largely unclear. Many common diseases tend to be related to each other, and they may share a common genetic origin. Studying the correlations among multiple diseases provides an important avenue to investigate the common genetic background of diseases and has the potential for better elucidating the genotype to phenotype mapping [26, 27], as well as better predicting disease association genes [28, 29, 30, 31, 32]. Furthermore, learning which diseases are correlated can help use of existing drugs to treat multiple similar diseases [33, 34, 35, 36, 37].

Meanwhile, network-based analysis is a good way for utilizing and developing network-related metrics and measurements to perform advanced analysis of biomedical data. Network science is a rising field where entities and their complex relationships are studied on a global scale [38, 39, 40], and has seen increasing applications for performing advanced analysis on biomedical data [41, 42, 43, 44, 45, 46]. There are various cellular components in the human cells that interact with each other within the same cell or across different cells [39]. A network called the *human interactome* can be constructed according to the interactions of those different cellular components. Each component can be represented as a vertex in the network and interactions among them can be captured as links (or edges) connecting pairs of the cellular components. There are a few types of interactome networks such as a molecular network, which has been studied a lot in recent years. This type of interactome network is based on the interactions among proteins where the vertices are proteins and there is a link between two vertices if there is a physical interaction between corresponding proteins [47, 48, 49]. Another type of interactome network is called the metabolic network where vertices are metabolites and a link connects two metabolites if they participate in the same biochemical reactions [50, 51, 52]. There are some other types of interactome networks, such as regulatory [53, 54] and RNA networks [55, 56].

Considering genes and diseases as vertices in the interactome networks, as well as the links connecting the vertices in such networks help us address some features of the genes which are related to genetic phenotypes and complex diseases [39]. Theoretical tools including graph theory and the branch of mathematics that is related to networks such as probability and statistics can be used to analyze networks [57]. For example, the link weights in the gene networks can be interpreted as the strength of interaction among genes. The number of neighboring vertices connected to a vertex shows the importance of that vertex. By using some sophisticated centrality measures, the most important vertices in the network can be identified more precisely. The most common centrality measures include degree, closeness, and betweenness centralities [57].

In this thesis, we propose a new method for the construction of a weighted human disease network (WHDN) and a weighted gene network (WGN). In addition, we propose two new centrality measures to identify the most important diseases and genes. First, we use a large database of disease-gene associations to build a weighted bipartite disease-gene network and then construct the WHDN, where link weights capture the strength of the pairwise disease correlations. In the same manner, we construct the WGN where link weights refer to the strengths of the pairwise gene-gene interactions. After the backbone extraction of the WHDN and the WGN, we design a centrality measure specifically for the context of the WHDN that considers not only the degree of a vertex but also the importance of its incident edges. Then, we extend the proposed centrality measure upon application to the WGN. In both networks, we compare our new centrality measures with degree, closeness, and betweenness by evaluating the network efficiency decline rate with the removal of top-ranked vertices by each centrality measurement. From the WHDN, we find the most important and central diseases with their most correlated disease. From the WGN, we identify the most important genes with the gene that has the strongest interaction with them. Important vertices, in this study, refer to the vertices that by removing them from the networks, the network efficiency will be declined. We also find the diseases that have the strongest association with the top-ranked genes in the WGN.

The rest of thesis is organized as follows. In the next section, some related studies are discussed. In section 3, we describe the construction, reduction, and properties of the networks. In this section, we discuss the disease-gene associations that we used to construct the bipartite disease-gene network, the weighted human disease network, and the weighted gene network. Our new proposed centrality measurements, comparison, and evaluation are given in section 4, followed by discussion in section 5.

Chapter 2

Literature Review

2.1 Complex Network Analysis for Human Disease Studies

Biological networks have been studied extensively in recent years. In this section, we discuss some related studies in which authors use complex network analysis on the biomedical dataset for different purposes. In this section, we can see the role of complex networks analysis on interpreting biomedical concepts and extracting the meaningful information based on web-based methods and technologies.

Oti *et al.* [49] used a PPI (Protein-Protein Interaction) network for predicting disease genes. Wu *et al.* [58] integrated PPI networks with gene expression data in order to rank disease genes associated with various cancers. They showed that their method was able to find replicable high-rank genes using different datasets. Vazquez *et al.* [59] used the protein-protein interaction network to assign a function to a unassigned protein. The idea is assigning the most common function(s) belonging to the classified interacting proteins to an unclassified protein. Hu *et al.* [60], constructed a network based on interactions among SNPs. Then, they ranked all pairwise interactions to extract a statistical epistasis network. Only those interactions whose strength was higher than a certain threshold were included in the epistasis network. As a result, 36 new SNPs related to *bladder cancer* were found. They showed that the statistical epistasis network shows significant properties of the genetic architecture of *bladder cancer*. Özgür *et al.* [61] used network-based measures such as degree, closeness, and betweenness centralities to rank genes in a gene-gene interaction network. Based on the ranked genes, they identified gene-disease associations. The results showed that the top 20 important genes ranked by network centrality measures are related to *prostate cancer*. Some important genes selected by closeness and betweenness measurements, whose relation with the diseases is unknown, were interpreted as candidate genes for future experimental studies. Cho and Zhang [62] proposed a new algorithm for extracting a hidden hub-oriented tree structure from an interactome network by calculating functional similarities among proteins. The results showed that the selected hubs are significant proteins in the yeast protein network.

Some studies aimed at identifying the correlations among diseases through network analysis [39, 63, 64]. Lee *et al.* [65] constructed a network where the vertices are diseases; the goal of the study was to find disease comorbidity, which can help predict and prevent diseases. Goh *et al.* [66] constructed a human disease network (HDN) by connecting pairs of diseases when they share common association genes. Of 1,284 diseases in the HDN, 867 have at least one link to other diseases, and 516 form a giant component, suggesting that the genetic origins of most diseases, to some extent, are shared with other diseases. Moreover, the HDN naturally and visibly clustered according to major disease classes such as the cancer cluster and the neurological disease cluster. Basic network-based measures show that cancer and neurological disorders have high genetic locus heterogeneity which causes a similar phenotype by the mutation in different loci [67]. Human disease networks provide scientists with a genome-wide roadmap for future investigations of the correlations among diseases. Researchers can visually find the correlation among diseases and associations among disease genes and diseases. To test the robustness of results obtained in this article, the author expanded the disease genes from 1,777 to 2,765 genes in which the newly added disease genes have unidentified mutation links [68]. The results have confirmed that the backbone of previous findings is still preserved, which shows the robustness of the results. Zhou *et al.* [69] extracted over twenty million bibliographic records from PubMed [70] in order to obtain 147,978 connections between 322 symptoms and 4,219 diseases. A human symptoms-disease network (HSDN) was then constructed and could show the symptom similarity between all pairs of diseases (7,488,851 links) in the network. The weight of a link represented the similarity of symptoms between two diseases. They showed that the correlations among diseases were significantly related to the genetic associations that each pair of diseases had in common as well as the interactions between their related proteins. Lee *et al.* [65] built a disease metabolism network in order to study disease comorbidity for better disease prediction and prevention. In this study, two diseases are connected with each other if a mutated enzyme catalyzes a metabolic reaction between them. Their results show that diseases with higher degrees, i.e., connecting with many other diseases, have a higher rate of prevalence and mortality.

2.2 Centrality Measures

Finding the most important vertices in the networks is one of the most challenging problems. Different methods rank the vertices from different point of views. Some methods just use local information to rank vertices and some other use global information. Some methods are designed for weighted graph and some other is defined for unwaighted graphs. Reasearchres are trying to rank the vertices in the network based on differnet properties of networks. In this section, we discuss the most commonly used centrality measures followed by recently proposed vertex centrality measures.

2.2.1 The Most Commonly Used Centrality Measures

2.2.1.1 Degree

Degree centrality refers to the total number of connections that a vertex has in the networks. In the context of network science, the more connections a vertex has, the more important that vertex is. Although degree centrality is a good measure to quantify the importance of vertices in the networks according to the direct connections the vertices have, it still suffers from a lack of contribution in the context of large scale structures. In other words, a vertex may not have a high degree in the network, but because of its connection to other vertices with high degrees, it may be considered important as well.

2.2.1.2 Closeness Centrality

The next most important measure in the network theory is closeness centrality, which is a good measure to specify the importance of the vertices in terms of their mean distance to all other vertices in that network [57]. Most network-based methods used for the analysis of data are built on direct interactions among vertices. Recently, indirect connections have drawn increasing attention. These connections describe the closeness of vertices in a network rather than relying on direct interactions [2]. The closeness centrality is defined as follows [57]:

$$C_i = \frac{n}{\sum_j d_{ij}} \tag{2.1}$$

where d_{ij} is the length of shortest path between the vertex i and the vertex j, and n is the number of vertices in the network. Although closeness centrality is a good measure to compute the centralities of vertices, this measure suffers some drawbacks. It is only usable in a connected network because the minimum distance from a vertex to other vertices from independent components is infinite. Therefore, the minimum distance is meaningless in the networks with multiple disconnected components.

2.2.1.3 Betweenness Centrality

Another popular centrality measure is betweenness centrality, which measures the extent to which a vertex located on the shortest paths of all pairs of other vertices. Suppose we want to distribute a message to all persons in a social network. The goal is distributing the message from vertex to vertex through the shortest paths. The number of shortest paths a vertex lies on is called betweenness centrality, which is defined as follows [57]:

$$x_i = \sum_{st} \frac{n^i_{st}}{g_{st}} \tag{2.2}$$

where n_{st} is the number of the shortest paths from s to t that vertex i lies on, and g_{st} is the total number of the shortest paths from vertex s to vertex t. A drawback of betweenness centrality is that a vertex has a score of 0 if it does not lie on the shortest path between any vertex pair. Therefore, it is possible that a significant portion of vertices in a network have a betweenness centrality of 0, which means that it is impossible to distinguish their importance.

2.2.2 Recently Proposed Vertex Centrality Measures

Measuring the centrality of vertices helps identify important vertices in the network. The most common centrality measures include degree (the total number of neighbors), closeness (the total distance to all other vertices), and betweenness (the fraction of locating on the shortest paths of all pairs of vertices) [57]. Despite wide applications in biological networks, these centrality measures are rather general and may not be able to capture all the properties of vertices in the context of biological networks. Therefore, carefully tailored centrality measures are needed for specific networks of interest.

Köhler *et al.* [71] proposed a vertex importance measure for disease genes in the context of PPI networks. They used a random walk strategy to assess the distance between vertices in the network, and reported improved performance compared with conventional distance-based centrality measures.

Martinez *et al.* [72] proposed a generic vertex prioritization method using the idea of propagating information across data networks and measuring the correlation between the propagated values for a query and a target set of entities. The authors tested their method by ranking disease genes associated with *alzheimer's disease, diabetes mellitus type 2* and *breast cancer*. They reported some new high-rank association genes that could bring new insights into the study of the diseases.

Liu *et al.* [73] proposed a new method for finding bridge vertices in the network in terms of their importance. Their proposed method is based on calculating the line/edge importance. One of the most significant aspects of this method is using local information instead of gathering global information, which results in reducing computational costs and complexity compared to other common measures such as CC (Closeness Centrality) and BC (Betweenness Centrality). For evaluating the correctness of the proposed method, two different approaches were used. The first approach was the transmission dynamic and the second approach was based on the fault tolerance of the network in the absence of a vertex, meaning that the efficiency of the network was measured after removing a vertex. The lower efficiency of a network after removing a vertex, the more important the removed vertex is.

Nitsch *et al.* [74] provided a web-based tool for giving priority to genes in a genome-wide PPI network. The method is based on the differential expression which each vertex has with its neighbors. The idea is that a vertex with neighbors that have more differentially expressed genes is more likely to be an important vertex. The random walk method [75] was used in the proposed method. Both differentially expressed genes and the strength of the interactions contribute to the importance of a vertex. Therefore, a vertex with weak interactions may have a high importance because of the high differential expression the vertex has with its neighbors.

Hu *et al.* [76] determined the importance of vertices in an unweighted network. The proposed method called "vertex importance contribution correlation matrix (NICCM)" was compared with some basic vertex importance measures in network theory, such as degree centrality and betweenness centrality, as well as newly developed methods like “vertex importance contribution matrix (NICM)”, proposed by Xujing [77]. The most interesting point of the NICCM method is that both neighboring vertices and the neighbors of the neighbors in the network unevenly contribute to the importance of the vertices where the NICM method is just based on the directly connected vertices. Another point is that there is an initialized importance score for each vertex, which is calculated as the shortest distance between the vertex and all other vertices.

Nie *et al.* [78] used the information entropy concept to quantify vertex importance in complex networks. Their proposed method, called “Mapping Entropy (ME)”, specifies how much a vertex in the network correlates with its neighbors. One of the advantages of this method is that it uses the local information to find the correlation of a vertex with its neighbors instead of global information. Their results have shown that the proposed method outperforms both degree and betweenness centralities.

Opsahl *et al.* [1] considered both vertex degree and the weight of edges together while calculating vertex importance. The focus of this method is not only on the number of links each vertex has but also on link weights. The idea can be given in an example where the score of the importance of a vertex is ten. This score can be given to a vertex with ten neighbors which have a link weight of one each, or for a vertex with one neighbor with a link weight of ten, or any combination of these states which results in ten. The goal of the article proposed by Opsahl [1] is to find the most important vertices in the networks based on making a balance between the number of links a vertex has and its link weights.

Yan *et al.* [79] proposed a new measure of vertex centrality in the weighted network. The method, called C-index, measures the collaboration competence of a vertex in such a network. The collaboration competence of a vertex depends on the number of neighbors of the vertex, the link weights incident to the vertex, and the importance of the neighbors. The basic idea of the C-index is based on the H-index, which was proposed by Hirsch [80]. The H-index gives an index, H, which shows the amount of achievement a scientist has. The number of a scientist's articles and the amount of citations in each paper contribute to the H-index. The most important aspect of the C-index method is that it uses some different factors to measure the collaboration competence of a vertex like the number of edges, edge weights, as well as the collaboration competence of collaborators themselves. In the proposed method, the total sum of link weights in a weighted network is defined as vertex strength, where the weight of edges captures the edge importance.

There are some different views for considering a vertex as the most important vertex in the network. One argument is that a vertex with the highest weighted links is the most important one, while another argument is that a vertex with the highest number of links is the most important. Alternatively, another view is that a vertex with the highest total sum of link weights is the most important vertex in the network. Many studies have been done to find the most important verticies based on different masures and techniquies by researchers. Specifically, in this study, the most important vertices refers to those which reduce the network efficiency after removing them from the networks.

Chapter 3

Network Construction

3.1 Dataset

The data used in this project contains disease-gene associations (DGAs) from multiple curated databases including UNIPROT, the comparative toxicogenomics database (CTD) (human subset), PsyGeNET, Orphanet, and human phenotype ontology (HPO). The disease-gene association data are conducted by DisGeNet group, available on DisGeNET v4.0 [81]. The current version of the dataset contains 130,821 DGAs, between 13,075 diseases and 8,949 genes. Each DGA is assigned a score a_i^k, for disease i and gene k, within the range of [0,1] based on its level of evidence, the number and the type of database sources supporting the DGA, and the number of publications verifying the association between the gene and the disease [81]. The formula to compute the GDA score is given below [82]:

$$a = C + M + \sum_{k=1}^{3} L_k \tag{3.1}$$

where:

$$C = \begin{cases} 0.6 & if N_{sources_i} > 2 \\ 0.4 & if N_{sources_i} = 2 \\ 0.2 & if N_{sources_i} = 1 \\ 0 & otherwise \end{cases}$$

where:

$N_{sources_i}$ is the number of CURATED sources supporting a DGA

$i \in$UNIPROT, CTD, PSYGENET, ORPHANET, HPO

$$M = \begin{cases} 0.16 & if N_{models} = 2 \\ 0.08 & if N_{models} = 1 \\ 0 & otherwise \end{cases}$$

where:

N_{models} is the number of animal models for a DGA $Models \in$ Rat, mouse from the rat genome database (RGD), the mouse genome database (MGD), and CTD

$$L = \begin{cases} 0.08 & if \frac{N_{gd} * 100}{N_{literature}} \geq 0.08 \\ \frac{N_{gd} * 100}{N_{literature}} & if \frac{N_{gd} * 100}{N_{literature}} < 0.08 \end{cases}$$

where:

N_{gd} is the number of publication supporting a DGA in the source k.

$N_{literature}$ is the total number of publication in the source k.

The first step in the project is to clean up the data in order to ensure that all diseases and genes in the dataset are unique and that there is no replication of disease-gene associations. Next, since some diseases and phenotypes overlap we only consider diseases in this study and remove all phenotypes from the dataset. We keep diseases and syndromes in the dataset for our analysis and remove injuries or poisonings, anatomical abnormalities, acquired abnormalities, mental or behavioral dysfunctions, signs or symptoms, findings, congenital abnormalities, neoplastic processes, and pathologic functions. We use DisGeNet web-based application [81] for this filtering.

3.2 Bipartite Disease-Gene Association Network

The best representation for depicting the associations among genes and diseases is a bipartite graph, which is called the disease-gene association network in this research. The bipartite graph contains two disjoint sets of vertices. One set represents diseases and another one represents genes. By definition, no edge is allowed to connect a pair of vertices in the same set of vertices in a bipartite graph. That is, there can be no link either between a pair of diseases or a pair of genes. There is an edge between a gene and a disease if there is an association between them. Link weights are given by

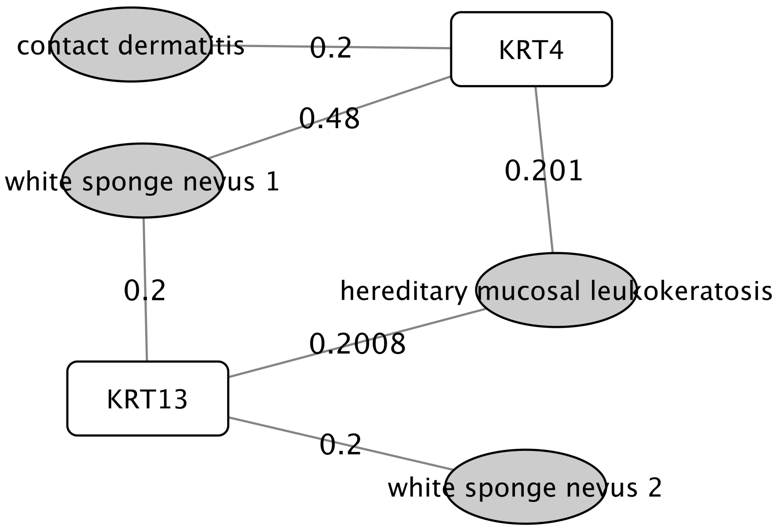
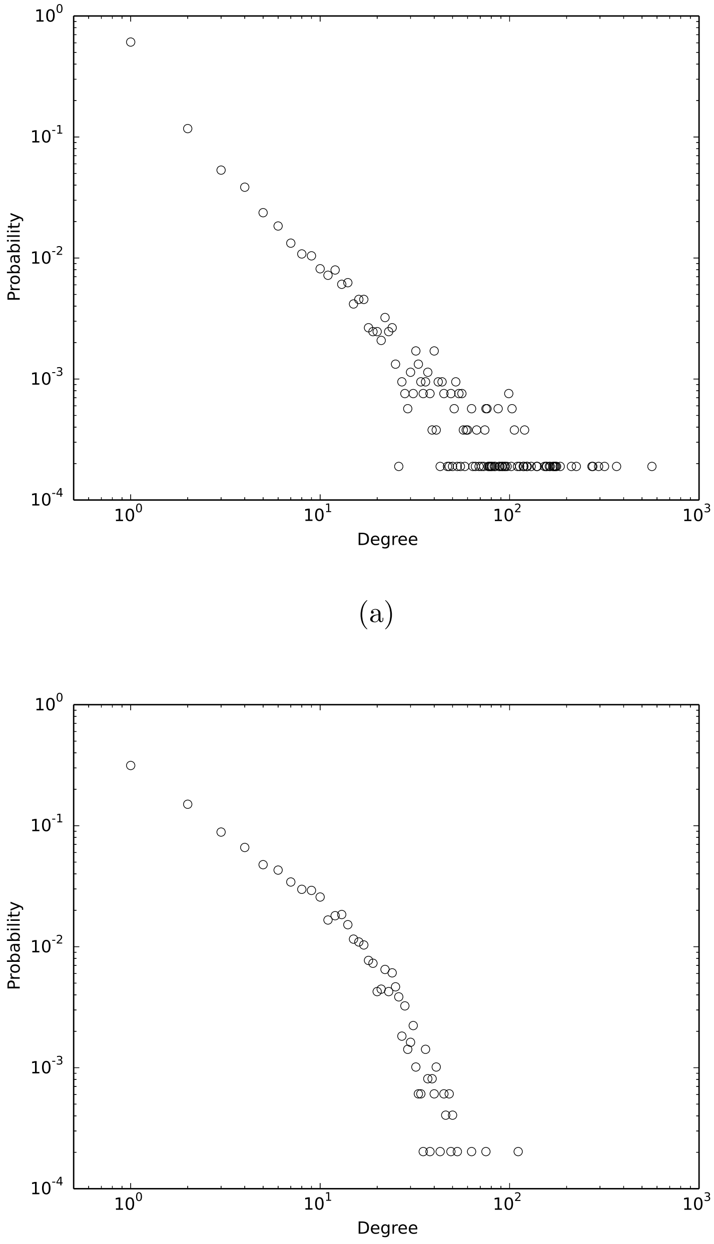


Figure 3.1: An example subgraph of the human disease-gene association network. The bipartite network has two sets of vertices, i.e., genes and diseases, represented by rectangle and gray ellipses, respectively. An edge connects a disease and a gene if there is a known association between them. The weight of an edge indicates the strength of the DGA a_i^k between disease i and gene k.

scores computed by the GDA method in the original dataset. A sample subgraph of the network is shown in figure 3.1.

Figure 3.2 depicts the degree distributions of diseases and genes in the bipartite disease-gene association network. For the set of diseases, the maximum degree is 564, of the disease *epilepsy*, and the average degree is 5.43. In Figure 3.2 a), the degree distribution of the diseases is right-skewed and approximately follows a power law distribution, indicated by the straight linear fit on a log-log scale. For the set of genes, the maximum degree is 111, of the gene LMNA, and the average degree is 5.81.

The bipartite network is comprised of multiple connected components with a single



(a)

(b)

Figure 3.2: Degree distribution of a) diseases and b) genes in the bipartite disease-gene association network. The distributions are shown on a log-log scale.

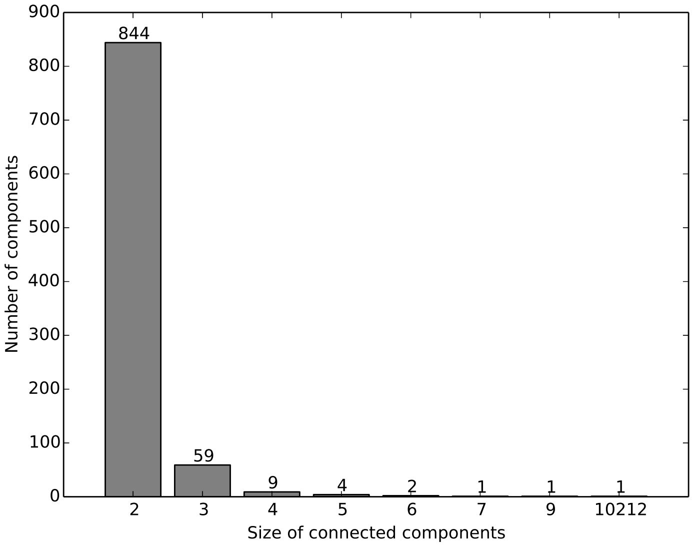


Figure 3.3: The size distribution of the connected components in the bipartite disease-gene network. The network has a single giant component with 10,212 vertices, and the majority of other connected components are of size two, i.e., consisting of only one disease and one gene.

giant component. Figure 3.3 shows its distribution of the size of connected components. The giant component has 10,212 vertices consisting of 5,278 diseases and 4,934 genes. Apart from the giant component, all other connected components are small with a size varying from two to nine, and most of them are only single pairs of one disease and one gene. Since we are interested in investigating the large-scale genetic correlations of human diseases as well as large-scale interactions among genes, we focus the giant component of the disease-gene bipartite network in the subsequent analyses.

3.3 Weighted Human Disease Network (WHDN)

We construct the weighted human disease network (WHDN) using the giant connected component of the bipartite disease-gene network. We use D and G to denote sets of 5,278 diseases and 4,934 genes respectively in the giant connected component. In the WHDN, an edge links two diseases i and j if they have at least one association gene in common, and the weight of the edge, w_{ij}, is computed based on the number of shared association genes, as well as the strengths of those associations.

Such a weight definition is inspired by Newman's study on scientific collaboration networks [38], where vertices are scientists and two scientists are connected by an unweighted edge if they have coauthored one or more scientific papers together. To define the strength of the tie between two connected scientists, two factors are considered. First, two scientists whose names appear on a paper together with many other coauthors know one another less well on average than two who are the sole author of a paper. Thus, the collaborative ties are weighted inversely according to the number of coauthors of a paper. Second, authors who have written many papers together will know one another better on average than those who have written few papers together. Thus, all coauthored papers are added up to account for the tie strength of two scientists.

Here, similarly, we consider that the correlation of two diseases through a gene is stronger when they are the only associated diseases with this gene than when there are many other diseases associated with the same gene. The correlation of two diseases is also considered stronger when they share more genes through stronger associations than fewer genes or weaker associations. Thus, we extend Newman's method to the weighted graph and define the weight of edge w_{ij} between two diseases i and j as

$$w_{ij} = \sum_{k \in G} \frac{\delta_i^k \delta_j^k (a_i^k + a_j^k)}{s_k}, \tag{3.2}$$

where δ_i^k is one if disease i and gene k have a DGA, and zero otherwise. a_i^k is the score of their DGA assessed by DisGeNET as discussed in the previous section, and s_k is the strength of gene k as a vertex in the bipartite disease-gene network, defined as the sum of the scores of the DGAs between gene k and its directly linked diseases,

$$s_k = \sum_{i \in D} a_i^k. \tag{3.3}$$

Such a weight definition indicates that the correlation strength of two diseases is weighted inversely according to the strengths of the genes they share, and is proportional to the total number of genes they share and the strengths of their DGAs.

For example, in Figure 3.1, the weight between diseases *contact dermatitis* (CD) and *white sponge nevus 1* (WSN1) is calculated as follows,

$$\begin{aligned} w_{CD,WSN1} &= \sum_{k \in G} \frac{\delta_{CD}^k \delta_{WSN1}^k (a_{CD}^k + a_{WSN1}^k)}{s_k} \\ &= \frac{a_{CD}^{KRT4} + a_{WSN1}^{KRT4}}{s_{KRT4}} \\ &= \frac{0.2 + 0.48}{0.881} \\ &= 0.7718. \end{aligned}$$

Note that the weight of two diseases can be greater than one when they share multiple genes. For example the weight between diseases WSN1 and *hereditary mucosal*

Leukokeratosis (HML) is calculated as follows,

$$\begin{aligned}
w_{WSN1,HML} &= \sum_{k \in G} \frac{\delta^k_{WSN1}\delta^k_{HML}(a^k_{WSN1} + a^k_{HML})}{s_k} \\
&= \frac{a^{KRT4}_{WSN1} + a^{KRT4}_{HML}}{s_{KRT4}} + \frac{a^{KRT13}_{WSN1} + a^{KRT13}_{HML}}{s_{KRT13}} \\
&= \frac{0.48 + 0.201}{0.881} + \frac{0.2 + 0.2008}{0.6008} \\
&= 0.7729 + 0.6671 \\
&= 1.44.
\end{aligned}$$

Since the WHDN is constructed using vertices from the giant component of the bipartite disease-gene association network, it only has a single connected component with all 5,278 vertices in the disease set D. Two vertices have an edge connecting them if the represented two diseases have at least one shared gene, and the edge weight is assessed as described above. The WHDN has 112,342 edges and an average vertex degree of 42.56. That is, a disease correlates with on average 42.56 other diseases with varying strengths.

3.4 Weighted Gene Network (WGN)

To construct the weighted gene network (WGN), we use the method proposed in the previous section. The idea of making a connection between a pair of genes and giving weight to the link is the same. The difference is in the definition of the network where in the WGN, an edge links two genes i and j if they are associated with at least one common disease, and the weight of the link, w_{ij}, is computed based on the number of shared association diseases, as well as the strengths of those associations.

The new extracted gene network is a single connected component because it is constructed from the giant component of the bipartite disease-gene association network. Therefore, the number of individual genes is the same as the number of genes in the gene set G of the bipartite graph, 4,934 vertices. Two vertices have a link connecting them if the represented two genes have at least one shared disease, and the link weight is assessed as described in the previous section. The WGN has 711,748 links and an average vertex degree of 288.5. That is, a gene interacts with on average 288.5 other genes with varying strengths.

3.5 The Multi-Scale Backbone of Networks

Figures 3.4 and 3.5 depict the distribution of all the edge weights in the WHDN and the WGN, respectively. As shown in this figure a large number of edge weights are of small values and may not be particularly interesting for subsequent analysis. Those weak edges not only add computational overhead to the network analysis, but also render the network difficult to interpret. So, we perform an edge reduction and only extract the most meaningful structure of the network.

3.5.1 Method

The most straightforward strategy for network reduction is to use a global weight threshold and remove all links that have weights lower than the threshold. However, such a global thresholding strategy is somewhat arbitrary and may overlook the network information present below the cutoff scale. To preserve the multi-scale backbone of the weighted human disease network (WHDN) and the weighted gene network (WGN) while removing less relevant and meaningful edges, we use a multi-

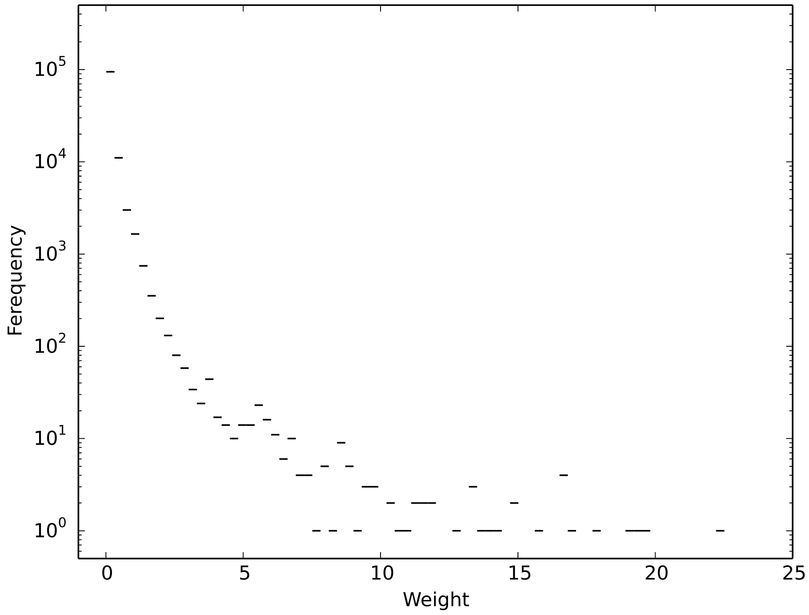


Figure 3.4: Distribution of edge weights in the WHDN. The weight of an edge quantifies the shared genetic background of two connected diseases. There are 112,342 edges in the graph with weights ranging from 0.0152 to 22.4506.

scale filtering method proposed by Serrano *et al.* [83]. The backbone of networks means the overall structure and topology of the networks.

First, the weights of edges linking vertex i with its neighbors can be normalized as

$$p_{ij} = \frac{w_{ij}}{s_i} \tag{3.4}$$

where s_i is the vertex strength, i.e., the sum of weights incident to vertex i, given by:

$$s_i = \sum_{j \in \Gamma(i)} w_{ij} \tag{3.5}$$

where $\Gamma(i)$ is the set of vertex i's neighbors. Therefore, there are two different normalized values for a link e_{ij} using the strengths of its two end vertices s_i and s_j as

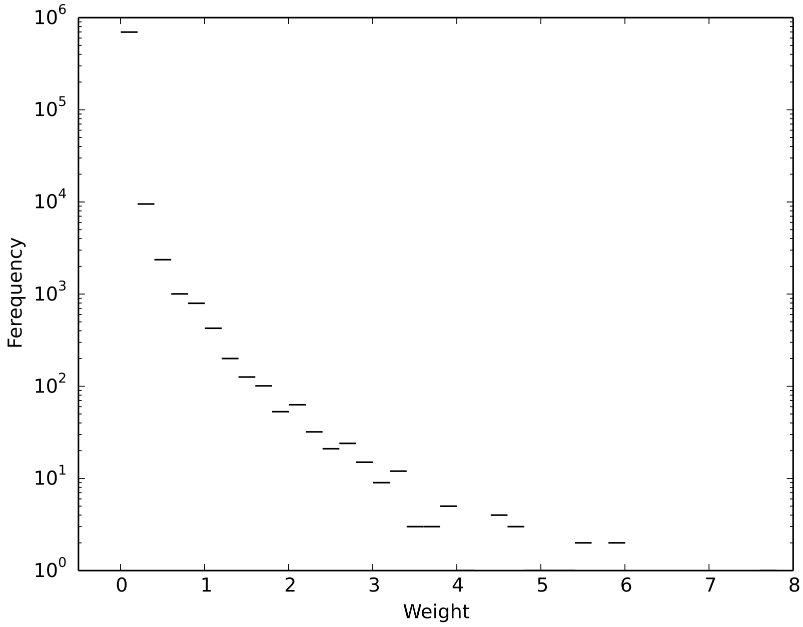


Figure 3.5: Distribution of edge weights in the WGN. The weight of an edge quantifies the strength of interaction between two genes. There are 711,748 edges in the graph with weights ranging from 0.0062 to 7.7856.

the denominator.

Second, a null model is introduced to inform us about the random expectation for the distribution of weights associated with the connections of a particular vertex. That is, the normalized weights p_{ij} that correspond to the connections of a certain vertex of degree k are produced by a random assignment from a uniform distribution. Thus the probability density function for one of these variables taking a particular value x is

$$p(x)dx = (k-1)(1-x)^{k-2}dx. \tag{3.6}$$

Then, Formula (3.7) is used to identify whether the probability, β_{ij}, of link weight

p_{ij} is compatible with the null model with a threshold β.

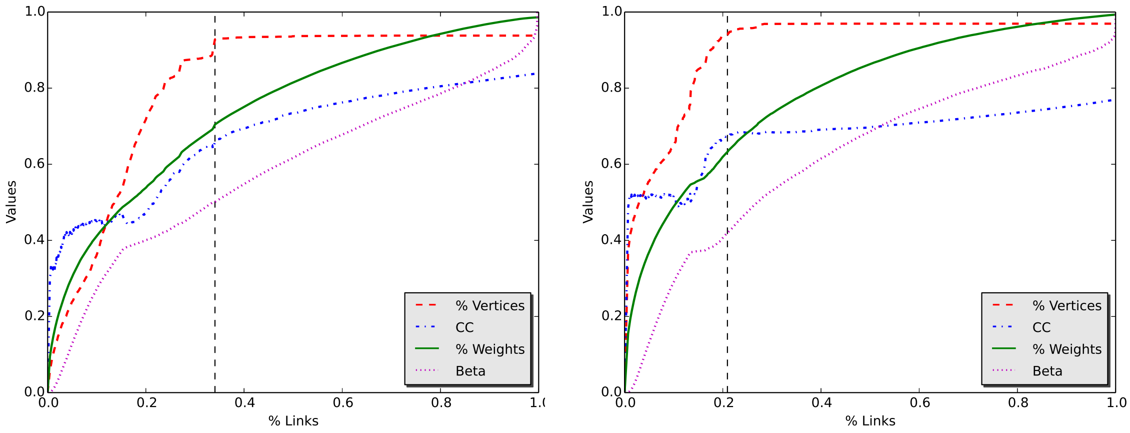
$$\beta_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2} dx < \beta \tag{3.7}$$

All links with β_{ij} lower than β are preserved in the network. Note that each edge has two different values β_{ij} and β_{ji}. For solving this problem, OR and AND rules can be used. Under the first rule, if either β_{ij} and β_{ji} is lower than β, the link will be preserved. In the second case, an edge is preserved if both β_{ij} and β_{ji} are lower than β. Darabos *et al.* [84] empirically found that the AND rule preserves the network features better than using the OR rule in the context of human phenotype networks. In this project, the AND rule is adopted to reduce the size of the networks by removing the links which are less relevant.

To find the best cutoff for β, we calculate clustering coefficient, percentage of remaining vertices and links, and total weight of the networks after applying a β cutoff while β changes from 0 to 1.

3.5.2 Results

Figure 3.6 show the results as a function of the percentage of remaining links in the WHDN and the WGN since we aim at removing as many links as possible while preserving the multi-scale backbone of the original weighted networks. We choose a β cutoff when the clustering coefficient and the remaining vertices and weights are maximally preserved while as many links are removed as possible. Accordingly, the cutoff $\beta = 0.501$ and $\beta = 0.42$ can be determined for the WHDN and the WGN, respectively, shown as the intersection of the vertical dashed line and the β curve in the figure.

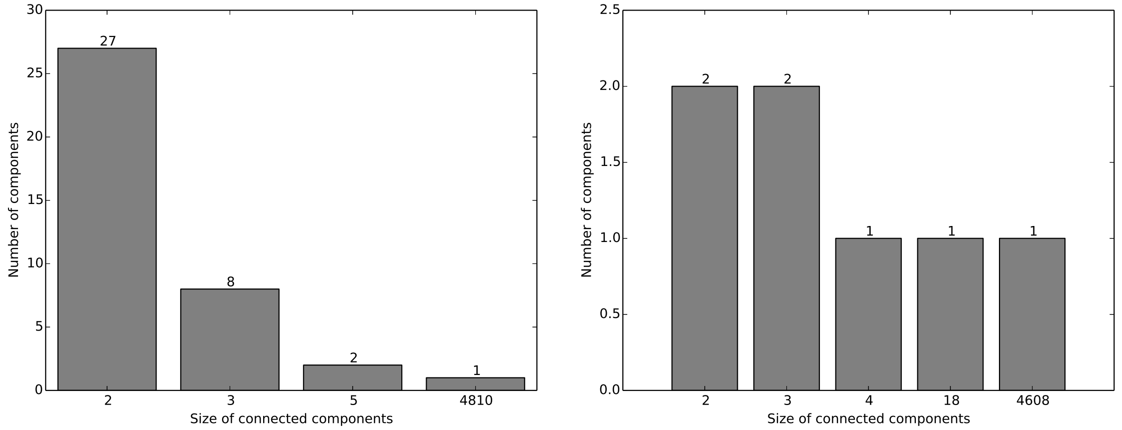


(a) (b)

Figure 3.6: Choosing the β value for a) disease network and b) gene network. CC represents clustering coefficient, %Vertices is the percentage of remaining vertices, %Weights is the percentage of weights left after removing links, and %Links is the percentage of remaining links.

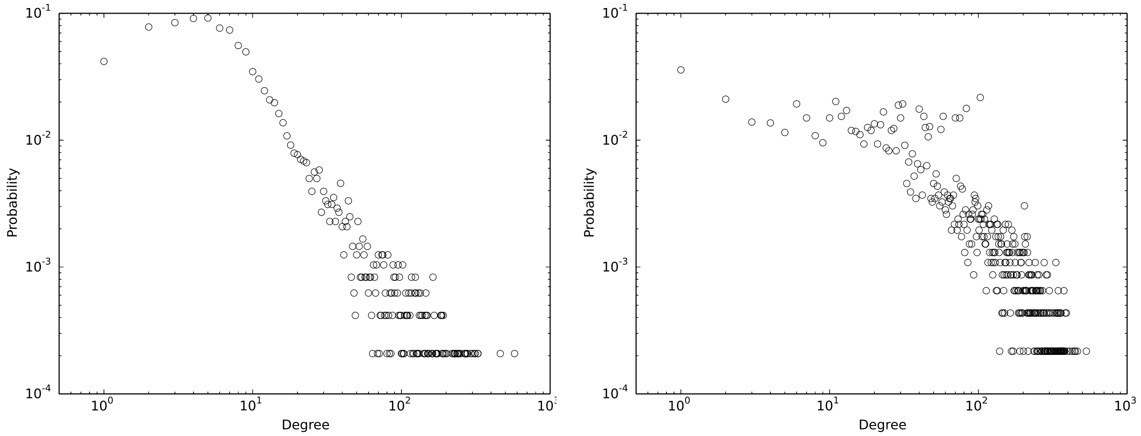
After the backbone extraction, the WHDN has 4,898 vertices and 38,275 edges and there are 4,640 vertices and 149,063 edges in the WGN. Those vertices are no longer connected by a single component.

Figure 3.7 shows the size distribution of connected components in the WHDN and the WGN. For the WHDN, there is a giant component with 4,810 vertices while the giant component in the WGN has 4,608 vertices. Degree distribution of giant components in the networks are shown in Figure 3.8. Again the degree distributions are heavy-tailed and resembles a power-law relationship. The vertex *epilepsy* has the highest degree of 576 in the WHDN and the highest degree 532 belongs to vertex *ERCC6* in the WGN. These giant components will be the focus for our next step analysis, i.e., measuring vertex importance in order to find the most central diseases in terms of correlating with other diseases and the most central genes in terms of interacting with other genes.



(a) (b)

Figure 3.7: The size distribution of connected components in the extracted backbone of the a) the WHDN and b) the WGN. The WHDN has a single giant component with 4,810 vertices while there are 4,608 vertices in the giant component of the WGN.



(a) (b)

Figure 3.8: Degree distribution of vertices in the giant component of the extracted backbone of a) the WHDN and b) the WGN. The distributions are shown on a log-log scale.

Chapter 4

Measuring Vertex Importance in Networks

4.1 Measuring Vertex Importance in WHDN

4.1.1 Proposed Method (DIL-W)

We introduce a vertex importance measure for the weighted human disease network (WHDN) by extending a centrality measure for unweighted networks proposed by Liu *et al.* [73]. This measure assesses the centrality of a vertex based on both its degree and the importance of its incident links (DIL centrality). For its extension on weighted graphs, we name it the DIL-W centrality.

First, in the context of unweighted graph, the importance of a link e_{ij} that connects vertex v_i and v_j can be calculated as follows:

$$I_{e_{ij}} = \frac{U_{e_{ij}}}{\lambda_{e_{ij}}}, \tag{4.1}$$

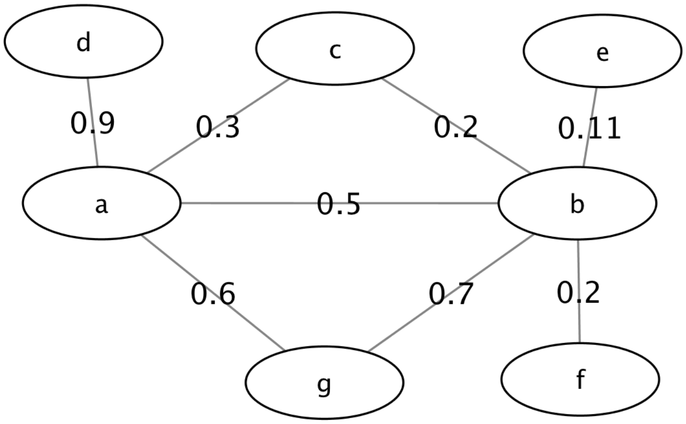


Figure 4.1: An example weighted graph.

where $U_{e_{ij}} = (k_i - t - 1)(k_j - t - 1)$ and $\lambda_{e_{ij}} = \frac{t}{2} + 1$. Following the convention, k_i and k_j are the degrees of vertex v_i and v_j, respectively, and t is the number of triangles with one edge being e_{ij}. The contribution that vertex v_i makes to the importance of e_{ij} is computed as

$$C_{v_i v_j} = I_{e_{ij}} \times \frac{k_i - 1}{k_i + k_j - 2}, \tag{4.2}$$

where $j \in \Gamma_i$, and Γ_i is the neighborhood of vertex i.

Then, the DIL centrality of vertex v_i is calculated by combining both its degree and the importance of its incident links,

$$\mathrm{DIL}_{v_i} = k_i + \sum_{v_j \in \Gamma_i} C_{v_i v_j}. \tag{4.3}$$

For weighted networks, we modify the computation of U in Equation (4.1) as

$$U_{e_{ij}} = (s_i - p_i) \times (s_j - p_j), \tag{4.4}$$

where s_i is the strength of vertex v_i, calculated by Formula (3.5), and p_i is the sum of link weights incident to vertex v_i that form triangles with e_{ij}. This follows the intuition that first an edge is considered more important when its two end vertices have higher strengths. Second, the importance of an edge is reduced when it has alternative two-hop paths connecting the same set of end vertices. Therefore, we subtract p_i from s_i in Equation (4.4).

We define λ for weighted graphs as

$$\lambda_{e_{ij}} = \frac{p_i + p_j}{2} + 1. \tag{4.5}$$

Finally, the importance of a vertex can be measured by

$$\text{DIL-W}_{v_i} = s_i + \sum_{v_j \in \Gamma_i} C_{v_i v_j}, \tag{4.6}$$

where $C_{v_i v_j}$ is defined as

$$C_{v_i v_j} = I_{e_{ij}} \times \frac{s_i}{s_i + s_j}. \tag{4.7}$$

As an example, in the weighted graph given in Figure 4.1, vertex a has a higher strength but a lower degree than vertex b. We compute their DIL-W centralities and investigate which one is more central when both factors are considered.

First we have their strength values $s_a = 0.9 + 0.3 + 0.5 + 0.6 = 2.3$, and $s_b = 0.2 + 0.11 + 0.2 + 0.7 + 0.5 = 1.71$. Their neighborhoods are $\Gamma_a = \{b, c, d, g\}$ and $\Gamma_b = \{a, c, e, f, g\}$. For vertex a,

$$\sum_{v_j \in \Gamma_a} C_{a v_j} = C_{ab} + C_{ac} + C_{ad} + C_{ag},$$

where

$$C_{ab} = I_{e_{ab}} \times \frac{s_a}{s_a + s_b},$$

and

$$I_{e_{ab}} = \frac{U_{e_{ab}}}{\lambda_{e_{ab}}} = \frac{(s_a - p_a) \times (s_b - p_b)}{\frac{p_a + p_b}{2} + 1}.$$

We have

$$p_a = w_{ac} + w_{ag} = 0.3 + 0.6 = 0.9,$$

and

$$p_b = w_{bc} + w_{bg} = 0.2 + 0.7 = 0.9.$$

So

$$\begin{aligned} C_{ab} &= \frac{(s_a - p_a) \times (s_b - p_b)}{\frac{p_a + p_b}{2} + 1} \times \frac{s_a}{s_a + s_b} \\ &= \frac{(2.3 - 0.9) \times (1.71 - 0.9)}{\frac{0.9+0.9}{2} + 1} \times \frac{2.3}{2.3 + 1.71} \\ &= 0.3423 \end{aligned}$$

We can also have

$$C_{ac} = 0.3285, \quad C_{ad} = 1.4878, \quad \text{and} \quad C_{ag} = 0.4312.$$

Then

$$\begin{aligned} \text{DIL-W}_a &= s_a + \sum_{v_j \in \Gamma_a} C_{av_j} \\ &= 2.3 + (0.3423 + 0.3285 + 1.4878 + 0.4312) \\ &= 4.8898. \end{aligned}$$

Similarly, we can compute the DIL-W centrality of vertex b as DIL-W$_b = 2.8916$. Based on both the degree and importance of incident edges, vertex a is considered more important than vertex b.

We apply the DIL-W centrality measurement to the giant component of the backbone of the WHDN, the distribution is shown in Figure 4.2. The DIL-W scores have a high dynamic range, from 0.0610 to 80688.1129. The majority of the vertices have low scores and a few number of vertices have scores that are greater by orders of magnitude.

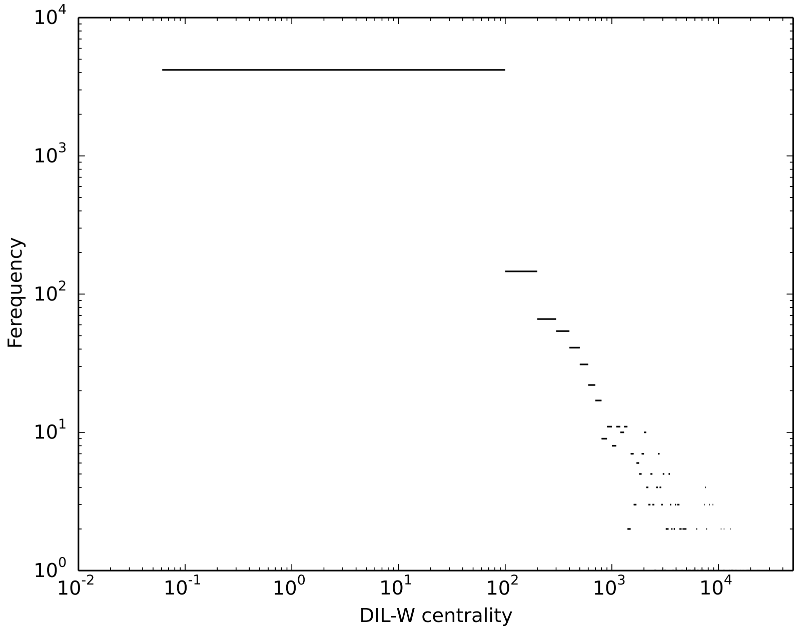


Figure 4.2: Distribution of the DIL-W centrality in the giant component of the WHDN on a log-log scale.

4.1.2 Results

We compare our DIL-W measurement with three most commonly used centralities, degree, closeness, and betweenness, when applied to the giant component of the backbone of the WHDN. For weighted graphs, degree centrality is calculated as vertex strength given by Equation (3.5). Closeness and betweenness are shortest-path-based centralities. Shortest path computation can be extended for the weighted graph as follows,

$$d_{ij}^{w} = min(\frac{1}{w_{ih}} + ... + \frac{1}{w_{hj}}). \tag{4.8}$$

Here d_{ij}^{w} denotes the weighted distance between vertex i and j, and w_{ih} is the weight of the edge linking vertex i and h. Since in our WHDN edge weight is strength, the distance between two vertices is the minimum sum of the inverse of edge weight along the path connecting them. Once the weighted distance is defined, closeness and betweenness can be calculated by their original definitions.

Figure 4.3 shows the correlation of DIL-W scores with a) degree, b) closeness, and c) betweenness centralities. As we can see, there is a positive correlation between the DIL-W measure and all three vertex centrality measures. The Spearman's rank correlation coefficient is 0.672 comparing DIL-W with closeness, is 0.71 comparing DIL-W with betweenness, and is 0.947 comparing DIL-W with degree.

To evaluate our new vertex importance quantification method, DIL-W, we measure the network efficiency before and after we remove the most important vertices in the WHDN. We calculate the decline rate of network efficiency after removing m top-rank vertices. The network efficiency [85] is computed based on the connectivity of a network. A higher connectivity suggests a higher network efficiency. The network efficiency is defined by

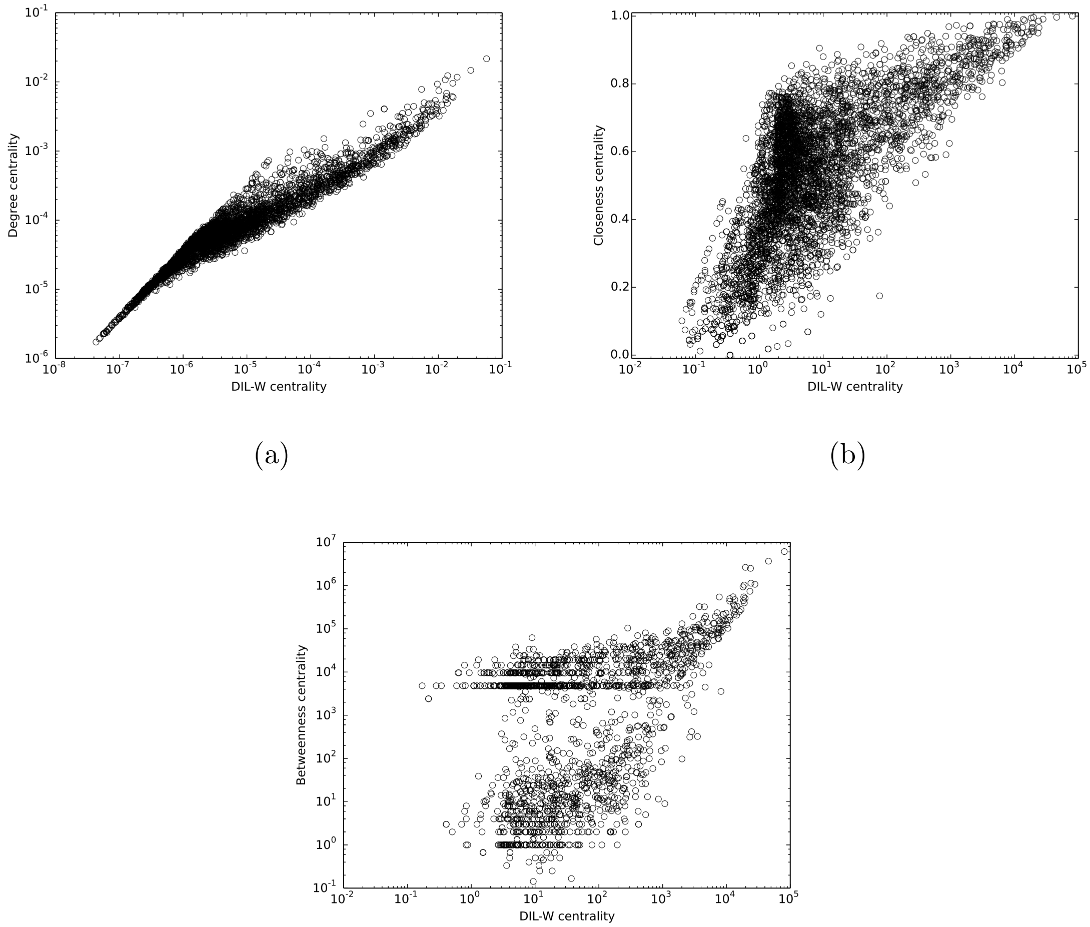
$$\eta = \frac{1}{n(n-1)} \sum_{v_i \neq v_j \in V} \frac{1}{d_{ij}}, \tag{4.9}$$

where n is the total number of vertices in the network, V is the vertex set, and d_{ij} is the distance, i.e., shortest path length, between vertex v_i and v_j. Thus, the decline rate of the network efficiency is calculated as

$$\mu = 1 - \frac{\eta}{\eta_0}, \tag{4.10}$$

where η_0 is the efficiency of the original network, and η is the network efficiency after some vertices are removed.

When a more important vertex is removed, we expect to see a greater decline rate of the network efficiency. Thus we can use μ as a indicator for the actual impact of



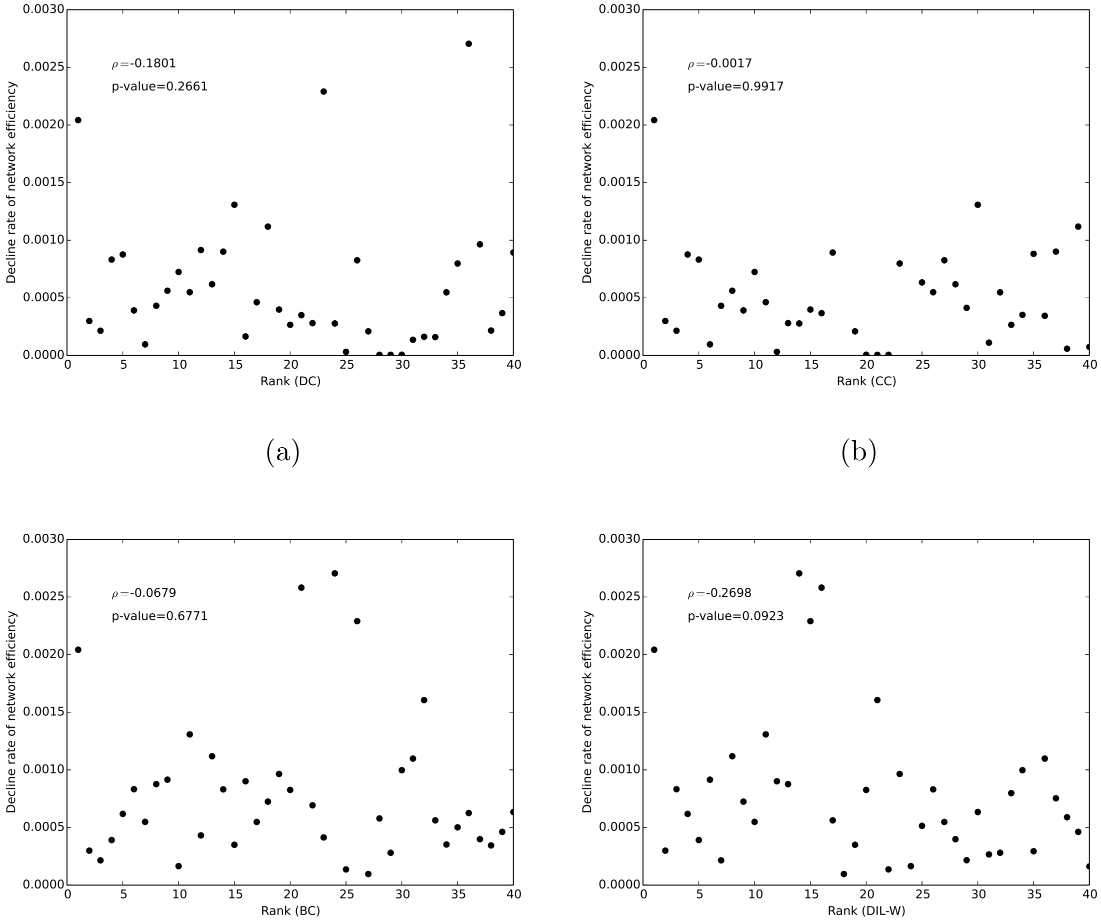
(a)

(b)

(c)

Figure 4.3: Correlation of DIL-W scores with a) degree centrality, b) closeness centrality, and c) betweenness centrality in the WHDN.

removing a vertex in the network. Figure 4.4 shows the decline rate of the network efficiency when we remove each of the top 40 vertices ranked by a) degree (DC), b) closeness (CC), c) betweenness (BC), and d) DIL-W. Further removal of top ranked vertices could be investigated but was not included in the current study given the high computational demand.



(a)

(b)

(c)

(d)

Figure 4.4: Decline rate of network efficiency after removing a single vertex ranked by a) degree centrality (DC), b) closeness centrality (CC), c) betweenness centrality (BC), and d) DIL-W.

As shown in the Figure 4.4, we do not observe a monotonic relationship across all four centrality methods. However, the correlation analysis shows that our method, DIL-W, has a slightly stronger negative correlation between the decline rate and the rank of the removed vertex than the other three. The Spearman's rank correlation coefficient for degree, closeness, and betweenness is -0.18, -0.001, and -0.06, re-

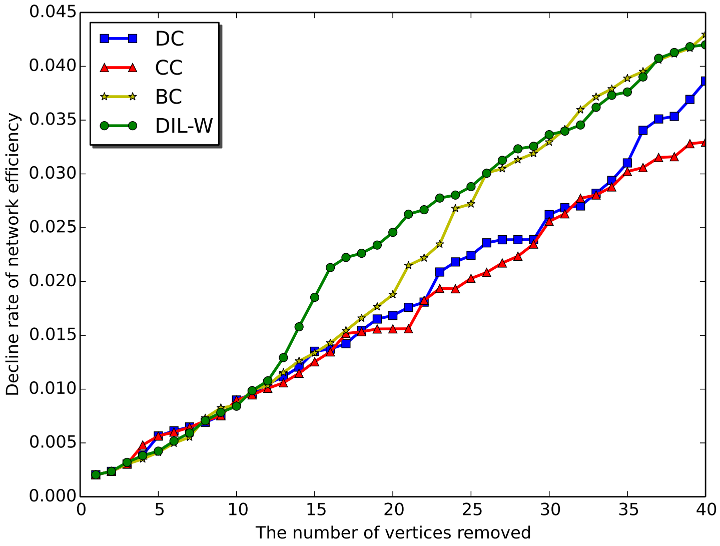


Figure 4.5: The decline rate of the network efficiency as a function of removing the top m vertices ranked by degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), and DIL-W.

spectively. In comparison, DIL-W has a negative correlation coefficient -0.26.

We also consider removing all m top-rank vertices at once to see how this accumulative removal affects the efficiency of the network. Figure 4.5 shows the decline rate of the network efficiency after removing top m vertices ranked by different centrality measures. The graph shows that the proposed method, DIL-W, has the highest decline rate of network efficiency for 57.5% of the data points, while betweenness, closeness, and degree have 27.5%, 10%, and 5%, respectively. This suggests that DIL-W is able to select a set of more important vertices comparing with the other three centrality measures. As seen in Figure 4.5 , the four methods are very competitive until the top 11 diseases are removed from the network. Then DIL-W has a significant higher network efficiency decline rate than the rest. Betweenness centrality catches up at point 31 and becomes very competitive afterwards.

We take a closer look at the top 31 diseases ranked by DIL-W since this is the most important set of diseases that resulted from the comparative study. Table 4.1 shows the top 31 diseases ranked by each centrality measure. Diseases that appear in multiple columns are shown with colors.

Table 4.1: Top 31 vertices in the WHDN ranked by different centrality measurements.

Rank	**DIL-W**	**CC**	**DC**	**BC**
1	Epilepsy	Epilepsy	Epilepsy	Epilepsy
2	Pediatric failure to thrive	Pediatric failure to thrive	Pediatric failure to thrive	Pediatric failure to thrive
3	Sensorineural hearing loss	Nystagmus	Nystagmus	Nystagmus
4	Anemia	Strabismus	Sensorineural Hearing loss	Obesity
5	Obesity	Sensorineural hearing loss	Strabismus	Anemia
6	Osteoporosis	Optic atrophy	Obesity	Sensorineural hearing loss
7	Nystagmus	Retinitis pigmentosa	Optic atrophy	Heart failure
8	Liver cirrhosis	Cerebral atrophy	Retinitis pigmentosa	Strabismus
9	Low vision	Obesity	Cerebral atrophy	Osteoporosis

Table 4.1: Top 31 vertices in the WHDN ranked by different centrality measurements.

Rank	DIL-W	CC	DC	BC
10	Heart failure	Low vision	Low vision	Chemical and drug induced liver injury
11	Muscle degeneration	Hypogonadism	Heart failure	Muscle degeneration
12	Diabetes mellitus, non-insulin-dependent	Developmental regression	Osteoporosis	Retinitis pigmentosa
13	Strabismus	Glaucoma	Anemia	Liver cirrhosis
14	Exophthalmos	Blindness, legal	Diabetes mellitus, non-insulin-dependent	Endometriosis
15	Myopia	Conductive hearing loss	Muscle degeneration	Rheumatoid arthritis
16	Degenerative polyarthritis	Supratentorial atrophy	Chemical and drug induced liver injury	Diabetes mellitus, non-insulin-dependent
17	Cerebral atrophy	Hyperinsulinism	Hypogonadism	Hypertrophic cardiomyopathy
18	Optic atrophy	Night blindness	Liver cirrhosis	Low vision

Table 4.1: Top 31 vertices in the WHDN ranked by different centrality measurements.

Rank	DIL-W	CC	DC	BC
19	Rheumatoid arthritis	Dystonic disease	Conductive hearing loss	Myocardial infarction
20	Hydrocephalus	Atrophy of cerebellum	Anxiety disease	Hydrocephalus
21	Alopecia	Cerebellar degeneration	Rheumatoid arthritis	Degenerative polyarthritis
22	Myocardial ischemia	Infratentorial atrophy	Glaucoma	Amyotrophic lateral sclerosis
23	Myocardial infarction	Hypoglycemia	Myopia	Neonatal hypotonia
24	Chemical and drug induced liver injury	Keratoconus	Blindness	Exophthalmos
25	Asthma	Gastroesophageal reflux disease	Developmental regression	Myocardial ischemia
26	Endometriosis	Heart failure	Hydrocephalus	Myopia
27	Hypertrophic cardiomyopathy	Hydrocephalus	Dystonic disease	Optic atrophy
28	Conductive hearing loss	Anemia	Atrophy of cerebellum	Coronary artery disease

Table 4.1: Top 31 vertices in the WHDN ranked by different centrality measurements.

Rank	DIL-W	CC	DC	BC
29	Brain ischemia	Neonatal hypotonia	Cerebellar degeneration	Glaucoma
30	Gastroesophageal reflux disease	Muscle degeneration	Infratentorial atrophy	Alzheimer's disease
31	Anxiety disease	Spastic quadriplegia	Myocardial ischemia	Polycystic ovary syndrome

Diseases that appear in multiple columns are shown with colors.

4.2 Measuring Vertex Importance in WGN

4.2.1 A special case in the DIL-W Method

We extend the DIL-W method proposed in the previous section to rank the vertices in the WGN. Although the DIL-W method provides a better result in terms of finding the most central vertices, there is a special case in which the DIL-W cannot distinguish between two links. The DIL-W does not take into account the number of triangles, t, into calculating the link importance. Recall Equation 4.1 which is used to calculate the link importance. In this equation, the $U_{e_{ij}}$ can be calculated by:

$U_{e_{ij}} = (s_i - p_i)(s_j - p_j)$ and $\lambda_{e_{ij}} = \frac{p_i+p_j}{2} + 1$

where s_i is the strength of vertex v_i, calculated by Formula (3.5), and p_i is the sum of link weights incident to vertex v_i that form triangles with e_{ij}.

As shown in the equation above, only p_i can contribute to calculate the link importance, while the number of triangles t should be considered as well.

In a situation where the values of s_i, s_j, p_i, and p_j are the same for two different links, while the number of triangles is different, the final value of link importance will be the same. In such a case, the number of triangles can make a difference between the values of link importance.

An example can be illustrated by looking at Figure 4.6 where both links e_{ab} in network (a) and e_{AB} in network (b) have the same value 10.1538 where their link importance can be calculated as follows:

$I_{e_{ab}} = \frac{U_{e_{ab}}}{\lambda_{e_{ab}}}$

For network (a) we have

$s_a = 12$, $p_a = 6$, $s_b = 16$, $p_b = 5$

So

$I_{e_{ab}} = \frac{((12-6)*(16-5))}{\frac{11}{2}+1} = 10.1538$

There are similar values for the parameters in the network (b).

$s_A = 12$, $p_A = 6$, $s_B = 16$, $p_B = 5$

Then

$I_{e_{AB}} = 10.1538$

By considering the number of triangles, we can distinguish between the links of the same importance. In the given example, link e_{ab} consists of two triangles with vertices c and d, while link e_{AB} can only make a triangle with vertex C.

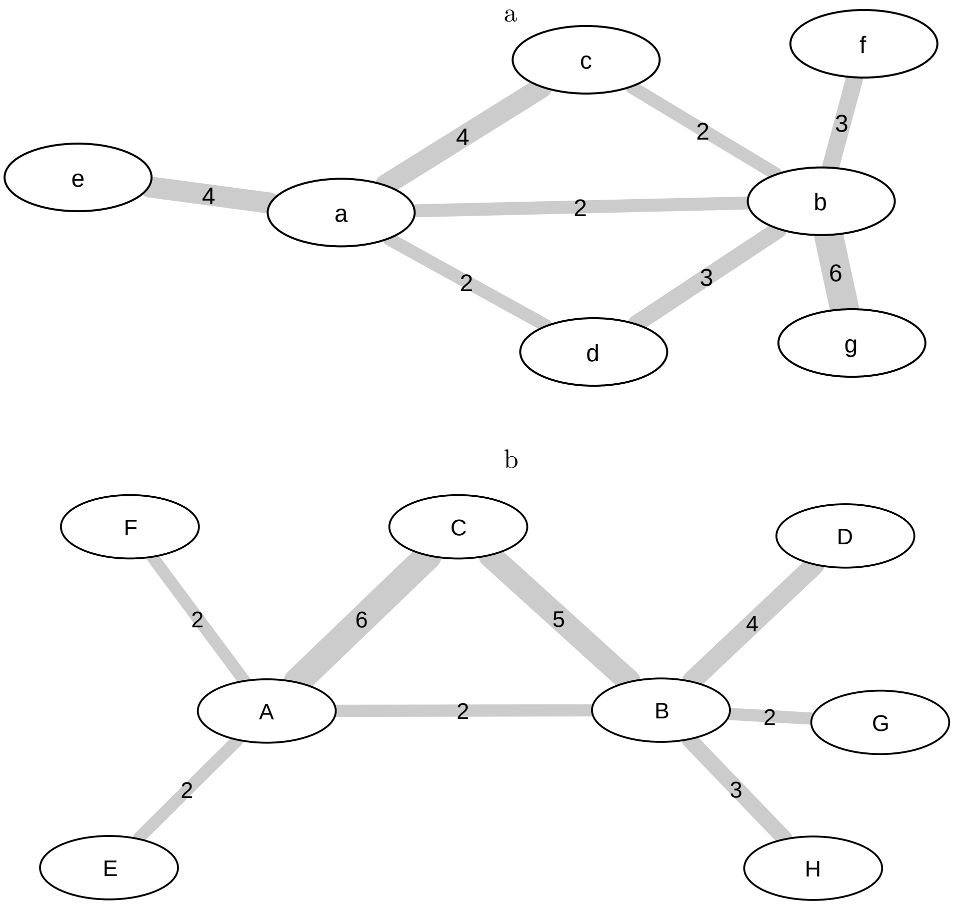


Figure 4.6: Two example networks.

4.2.2 Opsahl Method

The common methods for calculating degree centrality (DC) in unweighted and weighted networks is counting the number of links of vertex i and the strength, i.e., the total sum of the weights of links connected to vertex i, respectively. The number of links is neglected in computing DC in weighted networks and only the link weights contribute to DC. The main purpose of the Opsahl's [1] method is to consider both vertex degree and link weights together while calculating vertex importance. As shown in Figure 4.7, the strength of vertices A and B are the same. In this network,

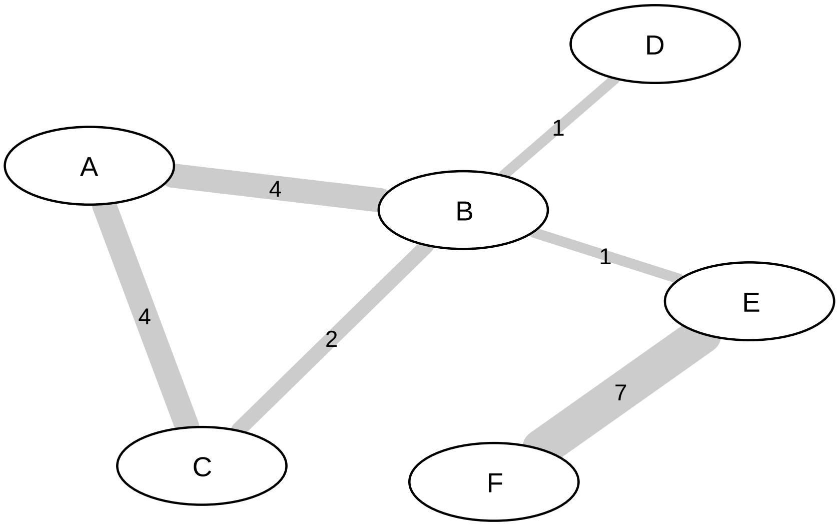


Figure 4.7: A network with 6 vertices and 6 weighted links. The size of links correspond to the link weights [1].

vertex A has two links while there are four links for vertex B. Since the number of links does not contribute to vertex strength, both A and B have the same score.

To distinguish between vertices A and B, a tuning parameter α is used to balance the number of links and the link weights. The formula for calculating degree centrality is as follows:

$$C_D^{w\alpha}(i) = k_i \times (\frac{s_i}{k_i})^{\alpha} = k_i^{(1-\alpha)} \times s_i^{\alpha} \tag{4.11}$$

where α is a positive parameter and s_i is the sum of link weights which can be calculated by Equation 3.5, and k_i is the number of links connected to vertex i:

$$k_i = \sum_{j}^{N} x_{ij} \tag{4.12}$$

where x is an adjacency matrix in which x_{ij} is 1 if there is a link between vertex i and vertex j, and 0 otherwise.

Table 4.2 shows the effect of α on the value of degree centrality. When α is between

Table 4.2: Degree centrality of vertices with different values of α

Vertex	k_i	s_i	$C_D^{w\alpha=0}$	$C_D^{w\alpha=0.5}$	$C_D^{w\alpha=1}$	$C_D^{w\alpha=1.5}$
A	2	8	2	4	8	16
B	4	8	4	5.7	8	11.3
C	2	6	2	3.5	6	10.4
D	1	1	1	1	1	1
E	2	8	2	4	8	16
F	1	7	1	2.6	7	18.5

0 and 1, the higher degree causes a higher score. For example, in the case of $\alpha = 0.5$, B has a higher score than A because vertex B has a higher degree. When α is greater than 1, the fewer number of vertex degree is favorable. For instance, when $\alpha = 1.5$, vertex A has a higher score compared with vertex B. $C_D^{w\alpha}$ is equal to k_i when α is 0. It means the $C_D^{w\alpha}$ score is equal to the vertex degree. $C_D^{w\alpha}$ is equal to s_i when α is 1. That is, the $C_D^{w\alpha}$ score is equal to the total sum of link weights.

By applying tuning parameter α, the closeness centrality can be defined as follows:

$$C_C^{w\alpha}(i) = \left[\sum_{j}^{N} d^{w\alpha}(i,j) \right]^{-1} \tag{4.13}$$

where the shortest distance between vertices i and j can be calculated as follows:

$$d^{w\alpha}(i,j) = min(\frac{1}{(w_{ih})^{\alpha}} + ... + \frac{1}{(w_{hj})^{\alpha}}) \tag{4.14}$$

Finally, the betweenness centrality is defined as follows:

$$C_B^{w\alpha}(i) = \frac{g_{jk}^{w\alpha}(i)}{g_{jk}^{w\alpha}} \tag{4.15}$$

where $g_{jk}^{w\alpha}(i)$ is the number of intermediary vertices and $g_{jk}^{w\alpha}$ refers to the link weights.

4.2.3 Proposed Method (EDIL-W)

The proposed method is called EDIL-W, which is an extension of the DIL-W method discussed in Section 4.1.1. There is a crucial question about the role of the number of triangles, t, and the sum of link weights, p_i, incident to vertex v_i that form triangles with e_{ij}. One can view the number of triangles as more important than p_i. That is, the presence of many triangles with any p_i might be considered more significant than p_i. On the other hand, p_i can be considered as a more important factor compared with the number of triangles in a weighted network. This trade-off is the most important reason for extending the DIL-W method where EDIL-W takes into account both the number of triangles, t, and p_i in calculating link importance.

The EDIL-W is defined as follows:

$$I_{e_{ij}}^{\alpha} = \frac{U_{e_{ij}}^{E}}{\lambda_{e_{ij}}^{E}}, \tag{4.16}$$

We modify the computation of U in Equation (4.4) as

$$U_{e_{ij}}^{E} = (C_i^{w\alpha} - P_i^{w\alpha}) \times (C_j^{w\alpha} - P_j^{w\alpha}) \tag{4.17}$$

where $C_i^{w\alpha}$ is the measure to compute centrality of vertex v_i, calculated by Formula (4.11), and $P_i^{w\alpha}$ can be defined as follows:

$$P_i^{w\alpha} = t^{(1-\alpha)} \times p_i^{\alpha} \tag{4.18}$$

where t refers to the number of triangles that include link e_{ij} as one of three edges, and p_i is the sum of link weights incident to vertex v_i that forms triangles with e_{ij}. From Equation 4.17, we know that, first, the importance of a link, e_{ij}, is dependent on both degrees of two end vertices, v_i and v_j, and their link weights in the case of calculating the strength of vertices $C_i^{w\alpha}$ and $C_j^{w\alpha}$. When α is between 0 and 1, high degree is favorable. When α is greater than 1, low degree is favorable and link weights contribute more to calculate the strength of vertices. Second, the importance of the link, e_{ij}, is reduced when there is an alternative two-hop path that connects the same set of end vertices ($P_i^{w\alpha}$ and $P_j^{w\alpha}$). In this case, both the number of triangles, t, and the sum of link weights, p_i, connecting the same set of end vertices contribute to reducing the link importance. When α is between 0 and 1, a fewer number of triangles is favorable. When α is greater than 1, a greater number of triangles increases link importance and p_i contribute more to link importance. A link is considered more important when its two end vertices have a higher centrality score $C_i^{w\alpha}$ and a lower $P_i^{w\alpha}$. Therefore, we subtract $P_i^{w\alpha}$ from $C_i^{w\alpha}$ in Equation (4.17).

We define λ for EDIL-W as

$$\lambda_{e_{ij}}^{E} = \frac{P_i^{w\alpha} + P_j^{w\alpha}}{2} + 1. \tag{4.19}$$

Finally, the importance of a vertex can be measured by

$$\text{EDIL-W}_{v_i} = C_i^{w\alpha} + \sum_{v_j \in \Gamma_i} C_{v_i v_j}, \tag{4.20}$$

where $C_{v_i v_j}$ is defined as

$$C_{v_i v_j} = I_{e_{ij}}^{\alpha} \times \frac{C_i^{w\alpha}}{C_i^{w\alpha} i + C_j^{w\alpha}}. \tag{4.21}$$

For example, in Figure 4.6, the importance of links e_{ab} in the network (a) and e_{AB} in the network (b) is the same if the DIL-W method proposed in the previous section is applied. By applying EDIL-W on network (a), we will have

when $\alpha = 0.5$

for the network (a)

$C_a^{w\alpha} = 4^{0.5} \times 12^{0.5} = 6.9282$, $C_b^{w\alpha} = 5^{0.5} \times 16^{0.5} = 8.9442$,

$P_a^{w\alpha} = 2^{0.5} \times 6^{0.5} = 3.4641$, and $P_b^{w\alpha} = 1^{0.5} \times 5^{0.5} = 3.1622$

then

$I_{e_{ab}}^{0.5} = \frac{(6.9282-3.4641)\times(8.9442-3.1622)}{\frac{3.4641+3.1622}{2}+1} = 4.6438$

For the network (b) we have

$C_A^{w\alpha} = 6.9282$, $C_B^{w\alpha} = 8.9442$, $P_A^{w\alpha} = 2.4494$, and $P_B^{w\alpha} = 2.2360$

then

$I_{e_{AB}}^{0.5} = 8.9881$

When $\alpha = 1.5$, for the network (a), we have

$C_a^{w\alpha} = 4^{-0.5} \times 12^{1.5} = 20.7846$, $C_b^{w\alpha} = 5^{-0.5} \times 16^{1.5} = 28.6216$,

$P_a^{w\alpha} = 2^{-0.5} \times 6^{1.5} = 10.3923$, and $P_b^{w\alpha} = 1^{-0.5} \times 5^{1.5} = 7.9056$

then

$I_{e_{ab}}^{1.5} = \frac{(20.7846-10.3923)\times(28.6216-7.9056)}{\frac{10.3923+7.9056}{2}+1} = 21.2127$

For the network (b) we have

$C_A^{w\alpha} = 20.7846$, $C_B^{w\alpha} = 28.6216$, $P_A^{w\alpha} = 14.6969$, and $P_B^{w\alpha} = 11.1803$

then

$I_{e_{AB}}^{1.5} = 7.6175$

Since all parameters including vertex strength, the number of links, and the value of parameter p_i are the same for both links e_{ab} and e_{AB}, we cannot distinguish between these two links by applying the DIL-W method. The EDIL-W method can easily solve this issue by adding a tuning parameter α. This parameter can create a balance between the number of triangles, t, and p_i.

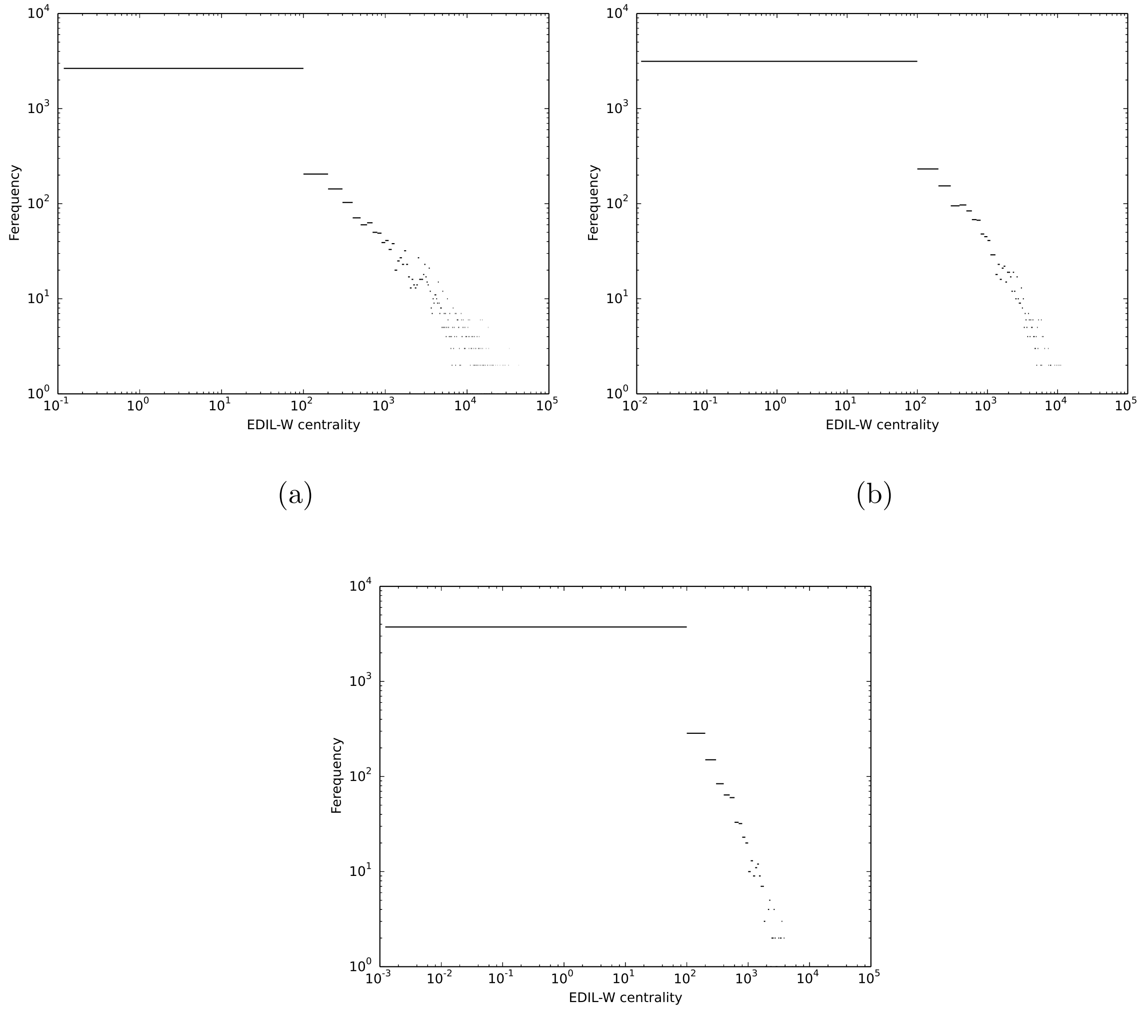
When $\alpha = 0.5$, a smaller number of triangles is favorable. That is, less triangles increases the link importance score. For example, in Figure 4.6, link e_{ab} in the network (a) creates two triangles. Therefore, the amount of $P_a^{w\alpha}$ will be increased when $\alpha = 0.5$, which results in a decreased link importance score. That is the reason that link e_{AB} seems more important than e_{ab}. The results are different by setting α to 1.5. In this case, a larger number of triangles is favorable. As result shown, the link importance score for e_{ab} is higher than e_{AB}, which means link e_{AB} is more important.

Table 4.3 shows the importance of all links of Figure 4.6 obtained by the EDIL-W method. The scores change when changing the value of α. For example, both links e_{ac} and e_{AC} have the same importance value when $\alpha = 1$. As shown in Figure 4.6, both links create one triangle, where p_C is greater than p_c and s_C is greater than s_c. The difference between these two parameters changes the link importance by changing the α value. When $\alpha = 0.5$, e_{AC} is more important than e_{ac}, and when $\alpha = 1.5$, the link e_{ac} is more important than e_{AC}.

In the case of an absence of triangles, there is still a possibility of having the same score for different links when $\alpha = 1$. For example, e_{ae} and e_{BH} have the same score since there is no triangle and the result of equation 4.17 for both links is the same. EDIL-W can distinguish between links in such a situation. When α is between 0 and 1, the higher number of links is favorable.

Table 4.3: The importance of links in the networks of figure 4.6 obtained by EDIL-W.

Link	$I_{e_{ij}}^{\alpha=0.5}$	$I_{e_{ij}}^{\alpha=1}$	$I_{e_{ij}}^{\alpha=1.5}$
bf	15.4919	48.0	148.7225
cb	4.0725	7.0	9.62
bg	21.9089	96.0	420.650
ae	13.8564	48.0	166.2768
ad	3.0648	5.7142	9.7067
ab	4.6438	10.1538	21.2127
ac	4.6818	13.3333	35.4762
bd	5.45230	14.0	34.2070
AF	9.7979	24.0	58.7877
BH	15.4919	48.0	148.7225
CB	5.7555	14.0	29.3273
BG	12.6491	32.0	80.9543
AE	9.7979	24.0	58.78775
AB	8.9881	10.1538	7.6175
AC	4.7902	13.3333	32.7901
BD	17.8885	64.0	228.9733



(a)

(b)

(c)

Figure 4.8: Distribution of EDIL-W centrality in the giant component of the WGN on a log-log scale with a) $\alpha = 0.5$, b) $\alpha = 1$, and c) $\alpha = 1.5$

When alpha is greater than 1 the less number of links is favorable. When $\alpha = 0.5$, e_{BH} is more important than e_{ae} because the degree of vertex B ($k_B = 5$) is more than degree of vertex a ($k_a = 4$). As a result, link e_{BH} is more important than e_{ae}. When $\alpha = 1.5$, link e_{ae} is more important than e_{BH}.

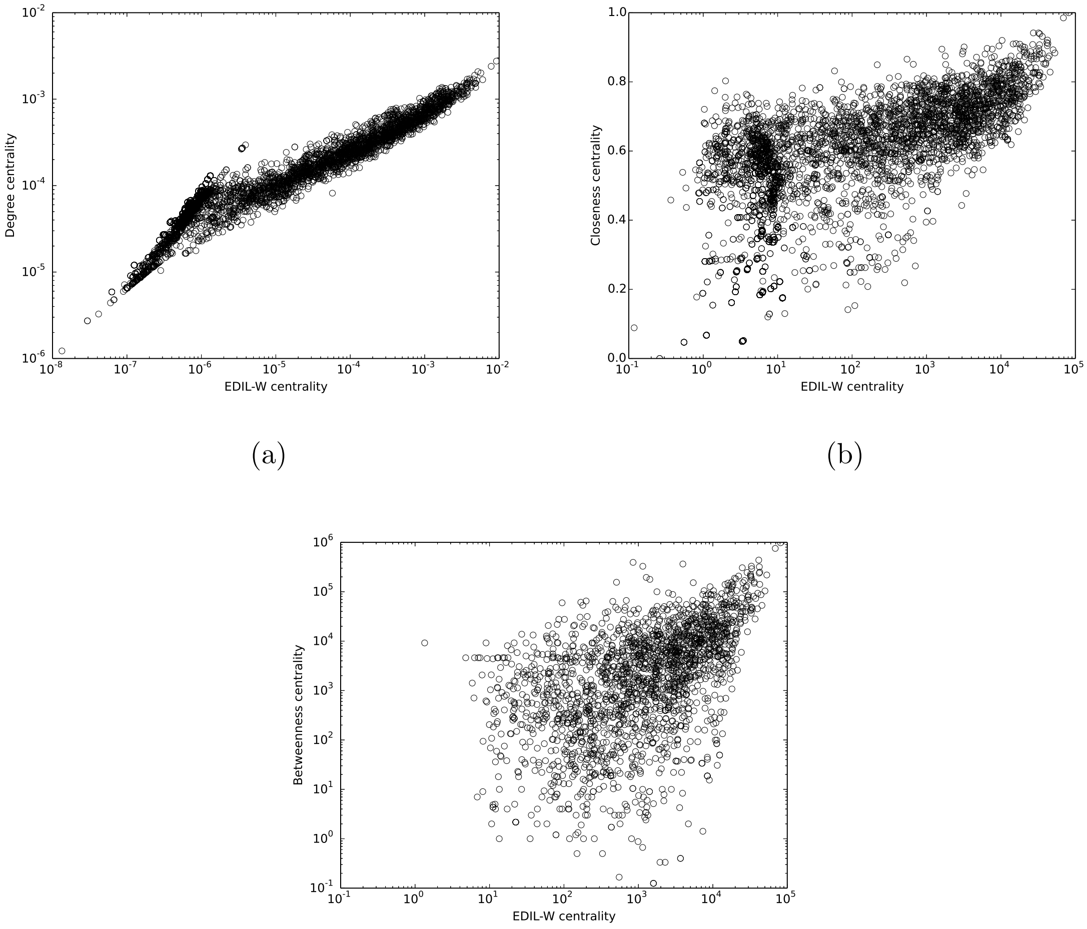
4.2.4 Results

We apply the EDIL-W measurement to the giant component of the backbone of WGN, the distribution is shown in Figure 4.8. The EDIL-W scores, with $\alpha = 0.5$ have a high dynamic range from 0.1180 to 81220.2734. By setting α to 1 and 1.5, the dynamic ranges are from 0.0115 to 29743.2460 and from 0.0012 to 10254.1827, respectively. The majority of vertices in each case with different α values have low scores and a few number of vertices have scores that are greater by orders of magnitude.

For the next step, we compare our newly extended method, EDIL-W, with three of the most commonly used centralities, i.e., degree, closeness, and betweenness, when applied to the giant component of the WGN. We use a generalized form of DC, CC, and BC proposed by [1] which can be calculated by Equations 4.11, 4.13, and 4.15, respectively.

We set the tuning parameter α to three different values 0.5, 1, and 1.5. Figure 4.9 to 4.11 show the correlation of EDIL-W with three other measures. As we can see there is a positive correlation between the EDIL-W measure and the other measures. When $\alpha = 0.5$, the Spearman's rank correlation coefficient is 0.95 comparing EDIL-W with degree centrality, 0.67 comparing EDIL-W with closeness centrality, and 0.82 comparing EDIL-W with betweenness centrality. When $\alpha = 1$, the Spearman's rank correlation coefficient is 0.97, 0.77, and 0.78 comparing EDIL-W with degree, closeness, and betweenness centralities. Finally, the Spearman's rank correlation coefficient is 0.97, 0.85, and 0.75 comparing EDIL-W with degree, closeness, and betweenness centralities when $\alpha = 1.5$.

To evaluate our new vertex importance qualification method, EDIL-W, we calcu-



(a)

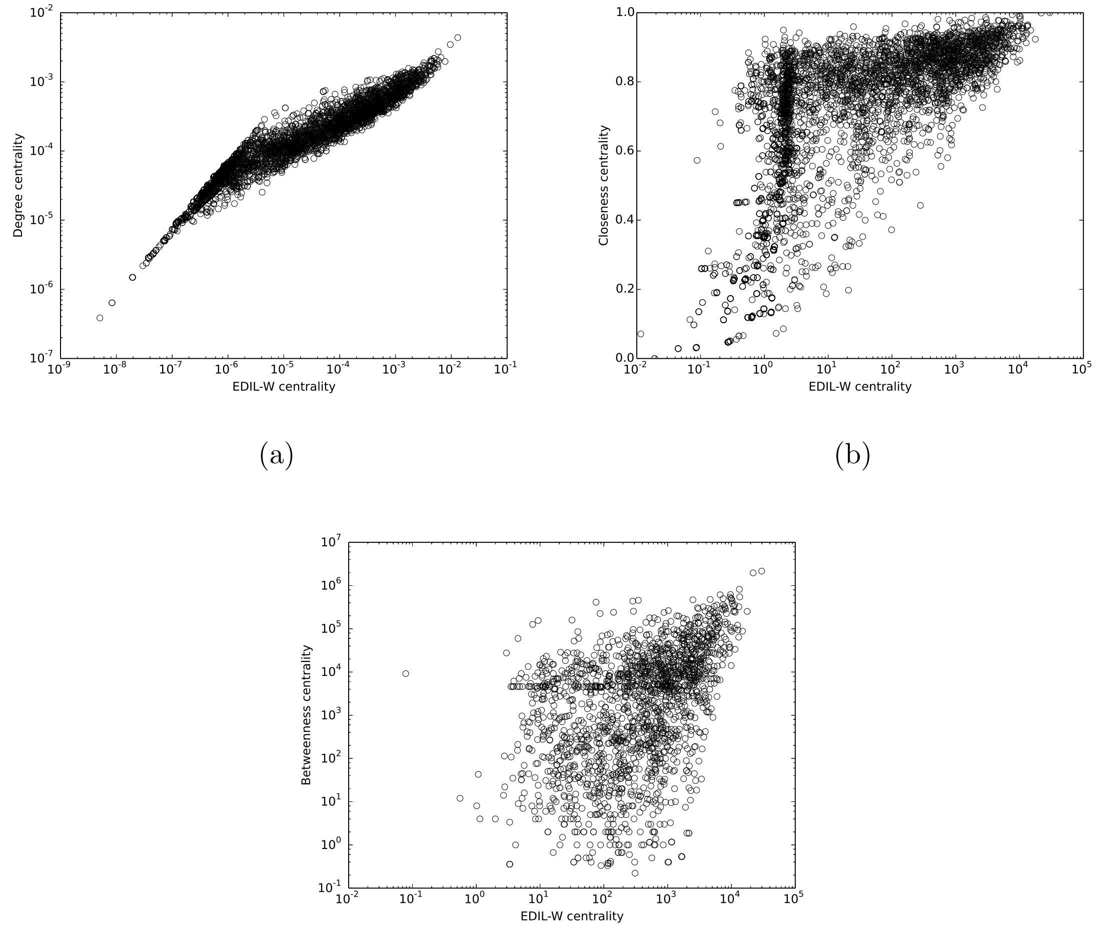
(b)

(c)

Figure 4.9: Correlation of EDIL-W scores with a) degree centrality, b) closeness centrality, and c) betweenness centrality when $\alpha = 0.5$ in the WGN.

late the decline rate of network efficiency after removing top m vertices selected by different measures to see how this accumulative removal affects the efficiency of the network. Figure 4.12 shows the decline rate of the network efficiency after removing top m vertices ranked by different measures.

When $\alpha = 0.5$, the fewer number of triangles is favorable. As we can see in



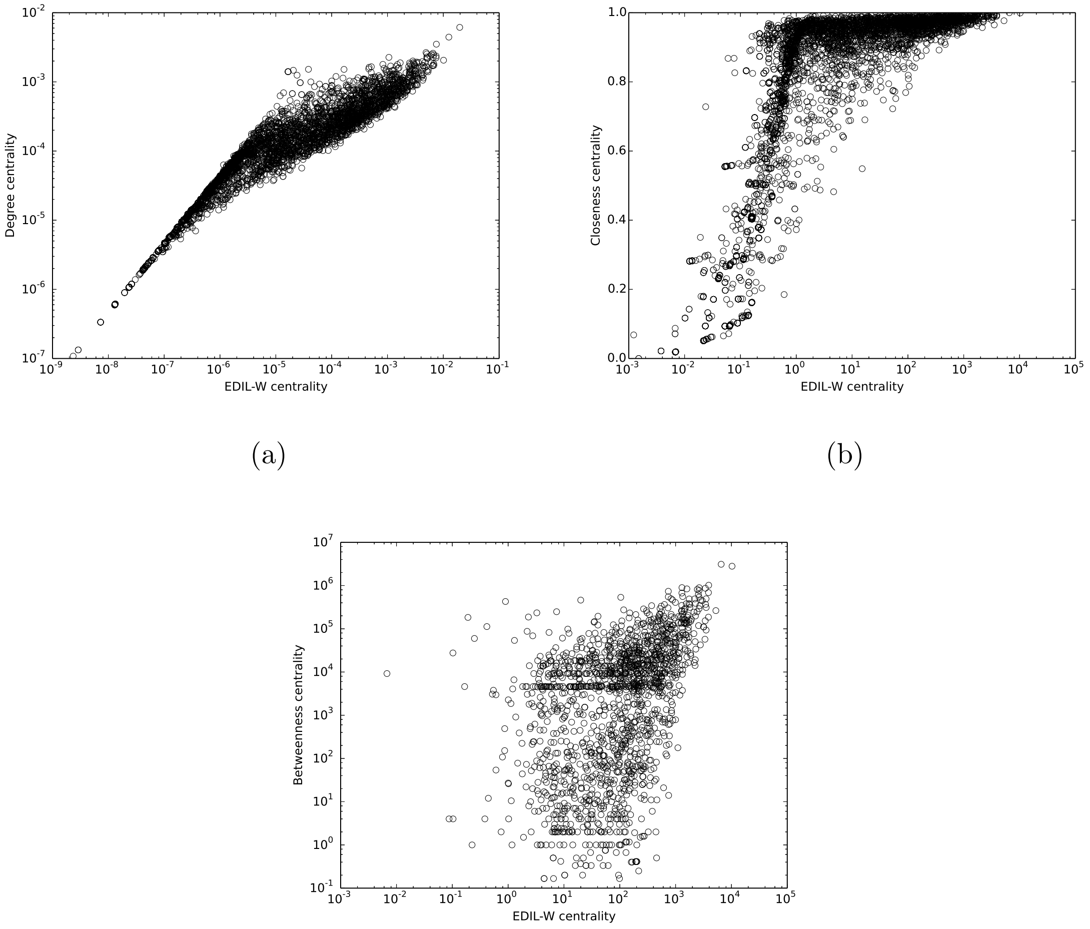
(a)

(b)

(c)

Figure 4.10: Correlation of EDIL-W scores with a) degree centrality, b) closeness centrality, and c) betweenness centrality when $\alpha = 1$ in the WGN.

Figure 4.12a, the new proposed method, EDIL-W, performs better than the degree and closeness centralities, but it is less effective than betweenness centrality. When $\alpha = 1$, the method will be the same as the DIL-W method proposed in Section 4.1.1. In this case, EDIL-W still performs better compared with degree and closeness centralities. Based on the decline rate of network efficiency given in Figure 4.12b, it



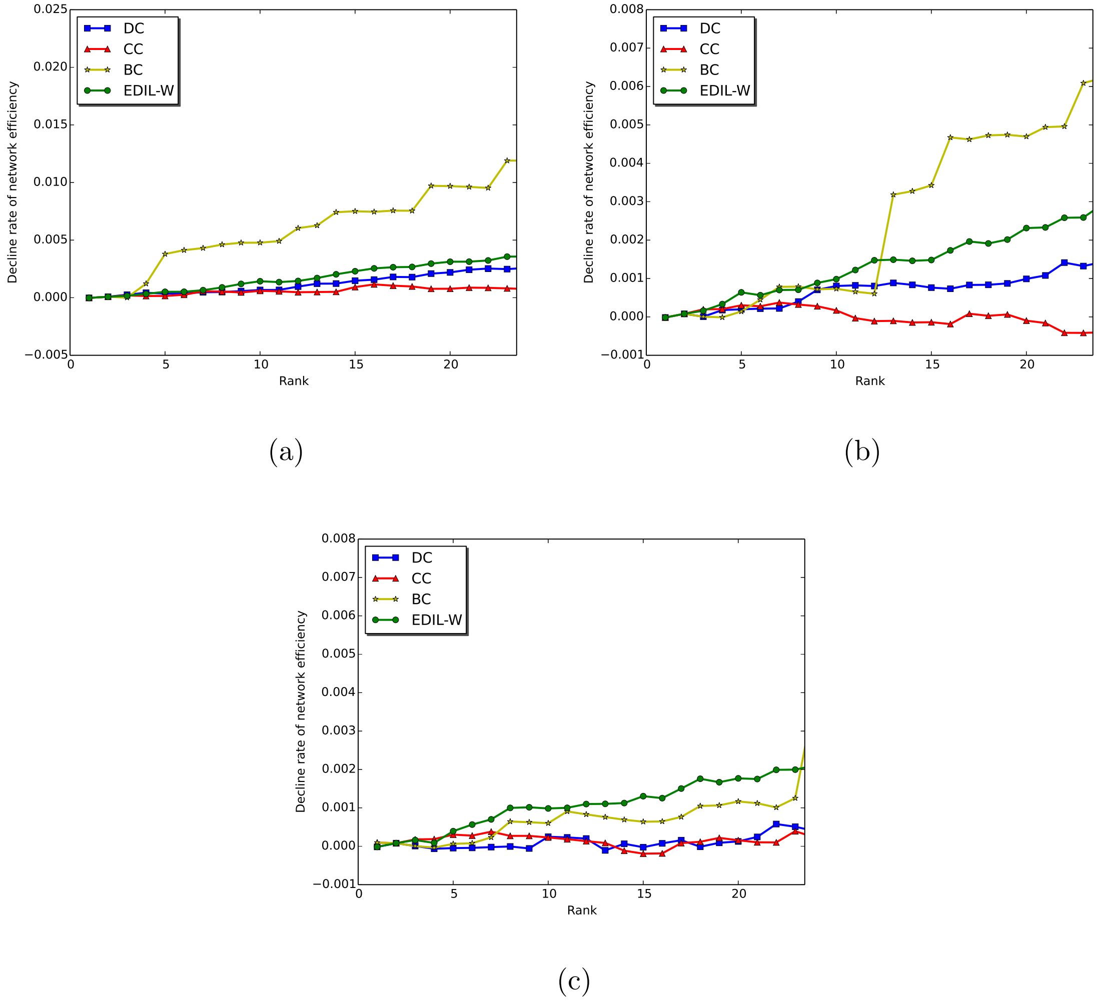
(a)

(b)

(c)

Figure 4.11: Correlation of EDIL-W scores with a) degree centrality, b) closeness centrality, and c) betweenness centrality when $\alpha = 1.5$ in the WGN.

seems that the top 12 vertices ranked by EDIL-W are more important than the top 12 vertices selected by betweenness centrality, which means that there may be stronger interactions among corresponding genes selected by EDIL-W than those selected by betweenness centrality in the network. Finally, Figure 4.12c, when $\alpha = 1.5$, indicates our proposed method outperforms other three centrality measures to find the most



(a)

(b)

(c)

Figure 4.12: The decline rate of the network efficiency as a function of removing the top m vertices ranked by applying different values of a) $\alpha = 0.5$ b) $\alpha = 1$, and c) $\alpha = 1.5$ on degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), and EDIL-W.

important top 23 vertices. That is, the top 23 genes selected by the EDIL-W method may have strong interaction with each other. In conclusion, the new proposed method under any value of alpha, identifies more important vertices compared with degree and closeness centralities. In comparison with betweenness centrality, the EDIL-W works less effectively when α is 0.5, but is more effective at finding some top vertices

Table 4.4: Top 12 genes in the WGN ranked by different centrality measurements when $\alpha = 1$.

Rank	**DC**	**CC**	**BC**	**EDIL-W**
1	LMNA	LMNA	LMNA	LMNA
2	TNF	TNF	TNF	TNF
3	FGFR2	ZMPSTE24	FGFR2	GBA
4	BRAF	IL1B	FGFR3	BRAF
5	ELN	IL6	PIK3CA	PTEN
6	FGFR1	MMP9	PTEN	FGFR2
7	IL1B	WRN	SOD1	POMC
8	ERCC6	TGFB1	IL1B	ALMS1
9	PTEN	HGD	FAS	ERCC6
10	IL6	IL10	FBN1	POLG
11	FBN1	LBR	GJB2	INS
12	FGFR3	GJB2	LDLR	GNAS

Diseases that appear in multiple columns are shown with colors.

when α is 1 and at finding more important vertices when α is 1.5.

The point is that both closeness and betweenness centralities are based on the shortest distance among vertices, which are not applicable measures to identify the most important vertices in a network with disconnected components. Betweenness centrality has another disadvantage where some vertices do not lie on the shortest paths between a pair of vertices. In such a case, betweenness centrality gives the vertices a value of zero. For example, the betweenness centrality gives positive value to only about 45 percent of the vertices in the gene network when α is 0. This is 40 and 35 percent when α value is 1 and 1.5, respectively. Therefore, the EDIL-W method outperforms the other three measures in finding the most important vertices.

Tables 4.4 and 4.5 show the top 12 and 23 most important genes selected by EDIL-W when α is 1 and 1.5, respectively. Genes that appear in multiple columns are shown with colors.

Table 4.5: Top 23 genes in the WGN ranked by different centrality measurements when $\alpha = 1.5$.

Rank	**DC**	**CC**	**BC**	**EDIL-W**
1	LMNA	TNF	TNF	LMNA
2	TNF	LMNA	LMNA	TNF
3	FGFR2	IL6	FGFR2	GBA
4	FAS	IL1B	NOS2	FGFR2
5	FBN1	ZMPSTE24	IL6	PTEN
6	IL1B	MMP9	FBN1	BRAF
7	ELN	WRN	PIK3CA	POMC
8	FGFR1	IL1A	SOD1	TBX1
9	TGFB1	IL10	FGFR3	FGFR1

Table 4.5: Top 23 genes in the WGN ranked by different centrality measurements when $\alpha = 1.5$.

Rank	**DC**	**CC**	**BC**	**EDIL-W**
10	PTEN	NOS2	MMP9	APOE
11	FGFR3	HGD	PTEN	FBN1
12	APOE	TGFB1	GJB2	POLG
13	SCN5A	IFNG	FAS	ALMS1
14	BRAF	LBR	SCN9A	ELN
15	COL2A1	GJB2	LDLR	ERCC6
16	IL6	PPARG	IL1B	TGFB1
17	GBA	BSCL2	SCN10A	FLNA
18	HBB	AGPAT2	CYP19A1	GNAS
19	HLA-DRB1	HLA-DRB1	FGFR1	COL2A1
20	TRNL1	GJB6	HLA-DRB1	IL6
21	ZMPSTE24	LDLR	IFNG	FGFR3
22	SOD1	CSF3	IL10	INS
23	COL1A1	ALB	FLNA	IL1B

Chapter 5

Discussion

In this project, we use a network-based analysis to identify important human diseases that share a genetic background with many other diseases through strong associations. In addition, we identify important disease genes associated with different diseases. We collect a large number of known disease-gene associations (DGAs) using the DisGeNET database in order to construct a bipartite disease-gene network. Subsequently, a weighted human disease network (WHDN) is built by connecting pairs of diseases that share associated genes. The edge weights reflect the number of genes they share as well as the strength of the DGAs. In a similar way, we construct a weighted gene network (WGN) in which link weights show the strength of interactions between gene pairs. Then, two methods are proposed to rank the vertices based on their centralities in the networks. To evaluate the proposed methods, results are compared with three commonly used centrality measures.

5.1 The Most Important Diseases

To identify the most important diseases, we propose a new vertex centrality measure, DIL-W, that considers both the degree of a vertex and the importance of its incident edges in weighted graphs. Upon application to the WHDN, DIL-W is shown to outperform degree, closeness and betweenness to find the important diseases. The DIL-W method is able to identify a set of 31 important diseases including *epilepsy*, *anemia*, and *obesity*.

As shown in Table 4.1, all four methods rank *epilepsy* and *pediatric failure to thrive* as the two most important diseases in the network. *Sensorineural hearing loss* is found as the third important disease by DIL-W whereas all other methods pick *nystagmus*. As we can see in Figure 4.4 d), *sensorineural hearing loss* (rank 3) has a higher decline rate of network efficiency than *nystagmus* (rank 7).

We show a visualization of the subnetwork including the top 31 diseases ranked by DIL-W in Figure 5.1. The 31 top-rank diseases not only appear very central on the entire WHDN but also have a large number of interconnections among themselves. They may be regarded as a dense core of the WHDN. In addition, *epilepsy*, *pediatric failure to thrive*, *obesity*, *heart failure*, and *osteoporosis* correlate with many other diseases in this subgraph, reflected by their vertex size. Some neurological diseases also tend to form a cluster.

Table 5.1 shows the degree in the WHDN and the most correlated disease of those 31 top-rank diseases. We are also able to find previous publications that verify almost all the correlations of the pairs of diseases, which are referenced in the table. Besides some very well-known correlations such as *heart failure* - *obesity* and *diabetes*

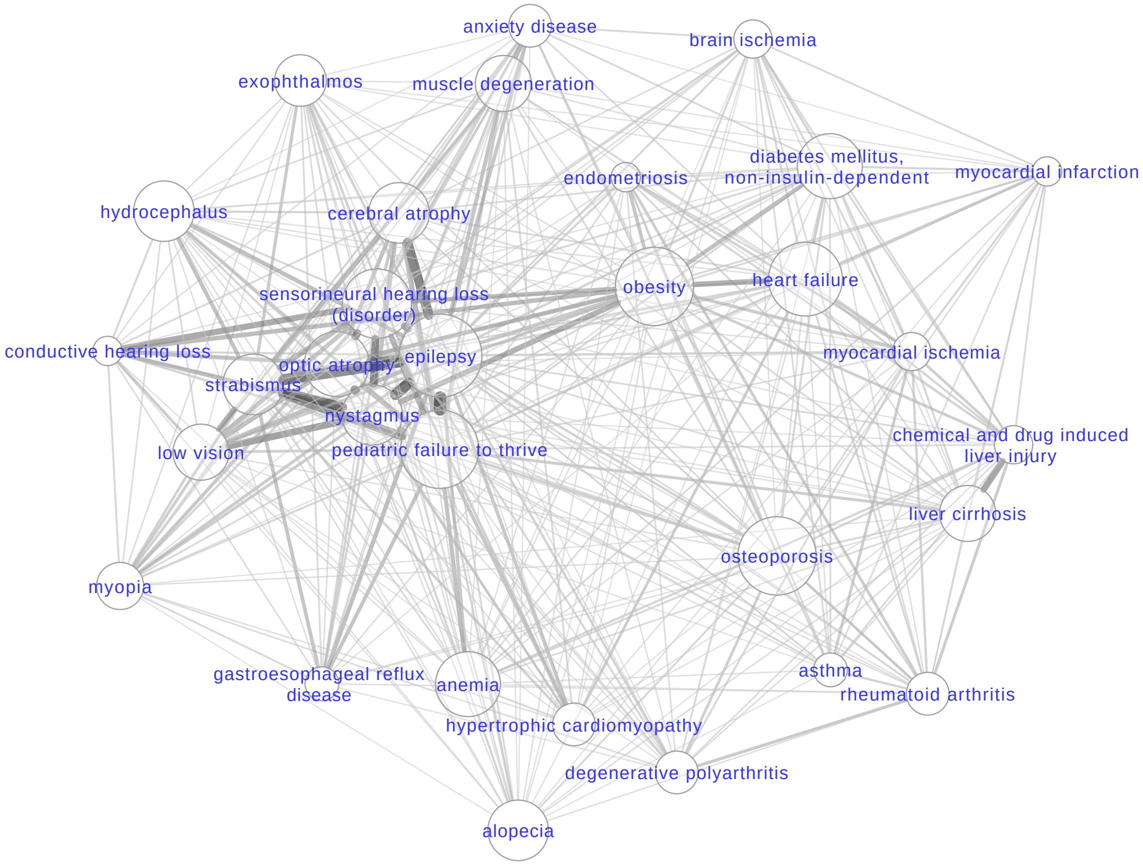


Figure 5.1: Subgraph of the WHDN that includes the 31 top diseases ranked by DIL-W. The size of a vertex is proportional to its degree. The width of an edge is proportional to its weight. Thicker edges indicate stronger disease correlations based on our weighting method defined in section 3.3. The gray scale of the edges is also proportional to their weights for visualization purposes. The network is shown using a force-directed layout such that vertices with stronger links appear closer.

- *obesity*, the table also reports some less known but interesting correlations. For instance, Savin [86] showed that *atypical retinitis pigmentosa* is correlated with *obesity*. Moreover, the correlation between *anemia* and *pediatric failure to thrive* had not been reported in the literature until recently Dimmock *et al.* [87] suggested *anemia* as one of the novel causes of *failure to thrive* in children. Zimmerman [88] studied the cause of different types of *cirrhosis* resulting from different drug-induced injuries. This supports our finding on the correlation between *cirrhosis* and *chemical and drug induced liver injury*.

Table 5.1: The diseases that have the most correlation with top 31 ranked diseases.

Rank	**Disease**	**Degree**	**The most correlated disease**	**Reference**
1	Epilepsy	576	Pediatric failure to thrive	–
2	Pediatric failure to thrive	462	Epilepsy	–
3	Sensorineural hearing loss (disorder)	313	Retinitis pigmentosa	[89]
4	Anemia	327	Pediatric failure to thrive	[87]
5	Obesity	268	Retinitis Pigmentosa	[86]
6	Osteoporosis	326	Osteopenia	[90]
7	Nystagmus	276	Epilepsy	[91]
8	Liver cirrhosis	278	Chemical and drug induced liver injury	[88]
9	Low vision	270	Nystagmus	[92]
10	Heart failure	311	Obesity	[93]

Table 5.1: The diseases that have the most correlation with top 31 ranked diseases.

Rank	Disease	Degree	The most correlated disease	Reference
11	Muscle degeneration	277	Amyotrophic lateral sclerosis	[94]
12	Diabetes mellitus, non-insulin-dependent	245	Obesity	[95]
13	Strabismus	293	Epilepsy	[96]
14	Exophthalmos	302	Strabismus	[97]
15	Myopia	266	Sensorineural hearing loss (disorder)	[98]
16	Degenerative polyarthritis	239	Rheumatoid arthritis	[99]
17	Cerebral atrophy	267	Epilepsy	[100]
18	Optic atrophy	236	Nystagmus	–
19	Rheumatoid arthritis	188	Lupus erythematosus, systemic	[101]
20	Hydrocephalus	250	Epilepsy	[102]
21	Alopecia	241	Dystrophia unguium	–
22	Myocardial ischemia	166	Obesity	–
23	Myocardial infarction	228	Coronary artery disease	[103]

Table 5.1: The diseases that have the most correlation with top 31 ranked diseases.

Rank	Disease	Degree	The most correlated disease	Reference
24	Chemical and drug induced liver injury	174	Cholestasis	[104]
25	Asthma	198	Dermatitis, atopic	[105]
26	Endometriosis	135	Obesity	[106]
27	Hypertrophic cardiomyopathy	187	Pediatric failure to thrive	[107]
28	Conductive hearing loss	163	Sensorineural hearing loss (disorder)	[108]
29	Brain ischemia	191	Diabetes mellitus, non-insulin-dependent	–
30	Gastroesophageal reflux disease	190	Epilepsy	[109]
31	Anxiety disease	185	Amyotrophic Lateral Sclerosis	[110]

5.2 The Most Important Genes

To identify the important disease-associated genes in the WGN, we extend the proposed method, DIL-W, to the method called EDIL-W. The purpose of this extension is to consider both the number of triangles and the sum of link weights incident to vertex v_i, which form a triangle with e_{ij} when calculating link importance. Recall that v_i is one of the end vertices of link e_{ij}. In addition, we balance the number of links and link weights when calculating vertex strength by adding a tuning parameter α. Table 4.4 shows the top 12 genes in WGN ranked by different centrality measures when $\alpha = 1$. *LMNA* and *TNF* are the first and second most central genes in the WGN, respectively, which are introduced by all centrality measures. Closeness centrality and then EDIL-W methods have the highest number of unique genes among all four methods. From Figure 4.12b, closeness centrality performs the least effectively among all four methods, while EDIL-W outperforms other three centrality measures in terms of finding the most important genes. Identifying the genes as the more important ones, which can not be found by other measures, can be considered as an important aspect of the proposed measure.

Table 5.2 shows the top 12 genes selected by EDIL-W when $\alpha = 1$. This table shows the degree of selected genes in the WGN and the genes that have the strongest interaction with them. In addition, the disease that has the strongest association with the corresponding gene is shown in the table.

We show a visualization of the subnetwork including the top 12 genes ranked by EDIL-W in Figure 5.2. This figure shows the interconnection among the top 12 genes, not the other genes connected to them. *ALMS1* and *LMNA* have the highest degree in the subgraph. From Table 5.2, we know the disease that is directly connected to the subgraph through one of the genes. Other genes in the subgraph that are not directly connected to the disease can be considered as candidate genes. Candidate genes are the genes that probably have an association with the diseases, but we are not sure

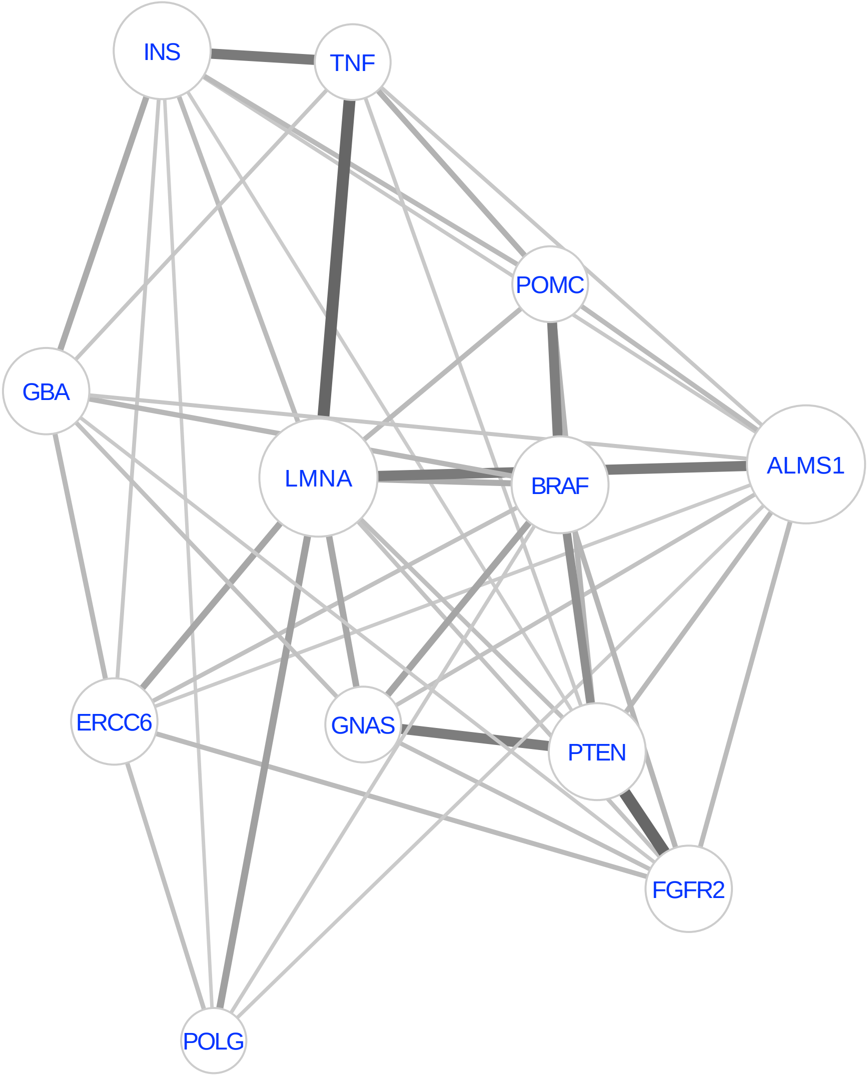


Figure 5.2: Subgraph of the WGN that includes the 12 top genes ranked by EDIL-W when $\alpha = 1$. The size of a vertex is proportional to its degree. The width of an edge is proportional to its weight. Thicker edges indicate stronger gene interactions based on our weighting method defined in section 3.4. The gray scale of the edges is also proportional to their weights for visualization purposes. The network is shown using a force-directed layout such that vertices with stronger links appear closer.

about that. For example, disease *progeria* is connected to subgraph through gene *LMNA*. Gene *LMNA* has 10 links in the subgraph, where its strongest interaction is with *ZMPSTE24* according to Table 5.2, which is not part of the subgraph. Then, all other 11 genes can be considered as potential candidate genes if no association has been already recognized between disease *progeria* and these 11 genes from the dataset.

Table 5.2: The genes that have the strongest interaction as well as the diseases that have the strongest association with top 12 ranked genes selected by EDIL-W when $\alpha = 1$.

Rank	**Gene**	**Degree**	**The most interacted gene**	**The most associated disease**
1	LMNA	449	ZMPSTE24	progeria
2	TNF	429	IL1B	rheumatoid arthritis
3	GBA	369	SCARB2	gaucher disease, type 1
4	BRAF	463	PTPN11	noonan syndrome
5	PTEN	299	AKT1	macrocephaly/autism syndrome
6	FGFR2	338	FGFR3	cutis gyrata syndrome of beare and stevenson
7	POMC	296	COL2A1	obesity
8	ALMS1	391	DICER1	alstrom syndrome
9	ERCC6	532	ERCC8	cockayne syndrome, type ii
10	POLG	412	TYMP	alpers syndrome (disorder)
11	INS	384	KCNJ11	diabetes mellitus, permanent neonatal

Table 5.2: The genes that have the strongest interaction as well as the diseases that have the strongest association with top 12 ranked genes selected by EDIL-W when $\alpha = 1$.

Rank	**Gene**	**Degree**	**The most interacted gene**	**The most associated disease**
12	GNAS	397	STX16	pseudohypoparathyroidism, type Ia

To evaluate the importance of the genes selected by the EDIL-W, we use DAVID [111, 112] (Database for Annotation, Visualization and Integrated Discovery) application v.6.8 to extract biological meanings from top 12 genes selected by EDIL-W. Table 5.3 shows a functional annotation chart of the 12 gene list. In this chart, 'Category' refers to original databased/resources where the terms orient, 'Term' lists enriched terms associated with the top 12 genes, 'Genes' refers to those genes involved in the term, and 'p-value' shows the Modified Fisher Exact p-value, Ease Score. The nineteen most significant terms based on p-value out of 35 are shown in this table which are sorted by p-value.

Table 5.3: Functional annotation chart of top 12 ranked genes selected by EDIL-W with $\alpha = 1$

Category	**Term**	**Genes**	**p-value**
KEGG-PATHWAY	prostate cancer	FGFR2, INS, PTEN, BRAF	3.7E-4

Table 5.3: Functional annotation chart of top 12 ranked genes selected by EDIL-W with $\alpha = 1$

Category	Term	Genes	p-value
GOTERM-BP-FAT	positive regulation of developmental process	GNAS, FGFR2, TNF	3.3E-3
SP-PIR-KEYWORDS	cleavage on pair of basic residues	GNAS, INS, POMC	3.7E-3
KEGG-PATHWAY	dilated cardiomyopathy	GNAS, LMNA, TNF	8.0E-3
GOTERM-BP-FAT	regulation of osteoclast differentiation	GNAS, TNF	8.7E-3
UP-SEQ-FEATURE	disulfide bond	GBA, INS, POMC, TNF	8.9E-3
SP-PIR-KEYWORDS	myristylation	GNAS, TNF	1.2E-2
GOTERM-BP-FAT	regulation of myeloid leukocyte differentiation	GNAS, TNF	1.5E-2
GOTERM-BP-FAT	regulation of cell proliferation	FGFR2, PTEN, TNF	1.8E-2
SP-PIR-KEYWORDS	disulfide bond	GBA, INS, POMC, TNF	2.0E-2
GOTERM-BP-FAT	regulation of myeloid cell differentiation	GNAS, TNF	2.5E-2

Table 5.3: Functional annotation chart of top 12 ranked genes selected by EDIL-W with $\alpha = 1$

Category	**Term**	**Genes**	**p-value**
GOTERM-BP-FAT	embryonic organ morphogenesis	GNAS, FGFR2,	3.7E-2
SP-PIR-KEYWORDS	glycoprotein	GNAS, GBA, POMC, TNF	4.1E-2
GOTERM-BP-FAT	bone development	GNAS, FGFR2	4.2E-2
GOTERM-BP-FAT	ossification	GNAS, FGFR2	4.2E-2
KEGG-PATHWAY	regulation of actin cytoskeleton	FGFR2, INS, BRAF,	4.4E-2
GOTERM-BP-FAT	embryonic organ development	GNAS, FGFR2	4.8E-2
GOTERM-BP-FAT	egulation of cell adhesion	PTEN, TNF	4.9E-2
SP-PIR-KEYWORDS	hormone	INS, POMC	5.0E-2

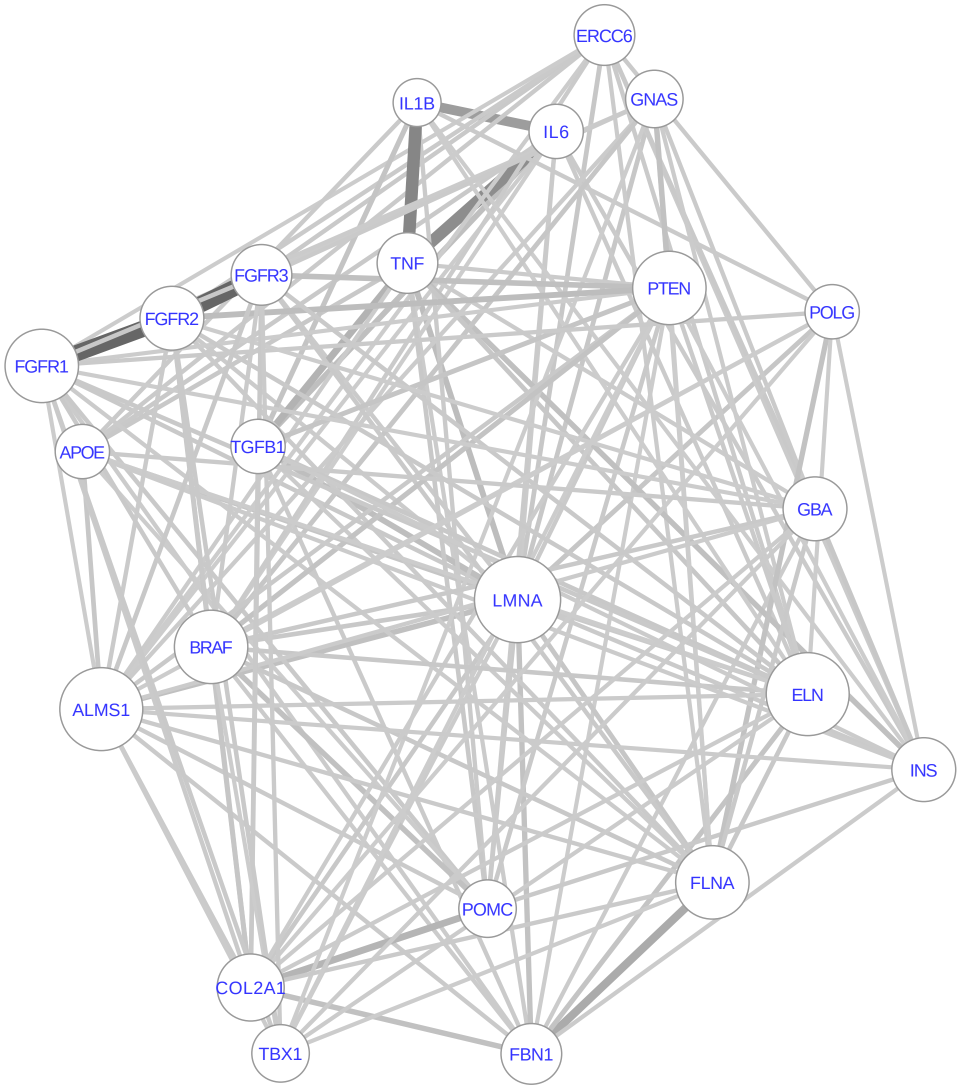


Figure 5.3: Subgraph of the WGN that includes the 23 top genes ranked by EDIL-W when $\alpha = 1.5$. The size of a vertex is proportional to its degree. The width of an edge is proportional to its weight. Thicker edges indicate stronger gene interactions based on our weighting method defined in section 3.4. The gray scale of the edges is also proportional to their weights for visualization purposes. The network is shown using a force-directed layout such that vertices with stronger links appear closer.

Table 4.5 shows the 23 important significant genes extracted by EDIL-W when $\alpha = 1.5$. Here, EDIL-W select the highest number of unique genes after the closeness centrality. Unique genes refer to those genes which cannot be selected by other centrality measures. Since results show that the closeness centrality is the least effective among the four methods, EDIL-W can be considered as the method with the most unique results.

We show a visualization of the subnetwork including the top 23 genes ranked by EDIL-W in Figure 5.3. This figure shows the interconnection among the top 23 genes, excluding other genes connected to them. *ALMS1*, *LMNA*, and *ELN* have the highest degree in the subgraph.

Table 5.4 shows the top 23 genes selected by EDIL-W when $\alpha = 1.5$. This table shows the degree of genes in the WGN and the most interacted gene and the most associated disease with the corresponding gene.

Table 5.4: The genes that have the strongest interaction as well as the diseases that have the strongest association with top 23 ranked genes selected by EDIL-W when $\alpha = 1.5$

Rank	**Gene**	**Degree**	**The most interacted gene**	**The most associated disease**
1	LMNA	449	ZMPSTE24	progeria
2	TNF	429	IL1B	rheumatoid arthritis
3	GBA	369	SCARB2	gaucher disease, type 1
4	FGFR2	338	FGFR3	cutis gyrata syndrome of beare and stevenson
5	PTEN	299	AKT1	macrocephaly/autism syndrome

Table 5.4: The genes that have the strongest interaction as well as the diseases that have the strongest association with top 23 ranked genes selected by EDIL-W when $\alpha = 1.5$

Rank	**Gene**	**Degree**	**The most interacted gene**	**The most associated disease**
6	BRAF	463	PTPN11	noonan syndrome
7	POMC	296	COL2A1	obesity
8	TBX1	204	COMT	digeorge syndrome
9	FGFR1	345	FGFR2	kallmann syndrome
10	APOE	261	LDLR	alzheimer's disease
11	FBN1	241	TGFBR2	marfan syndrome
12	POLG	412	TYMP	alpers syndrome
13	ALMS1	391	DICER1	alstrom syndrome
14	ELN	347	FBLN5	supravalvular aortic stenosis
15	ERCC6	532	ERCC8	cockayne syndrome, type ii
16	TGFB1	277	TNF	camurati-engelmann syndrome
17	FLNA	381	FBN1	oto-palato-digital syndrome type 1
18	GNAS	397	STX16	pseudohypoparathyroidism, type ia
19	COL2A1	199	COL11A2	platyspondylic lethal skeletal dysplasia, torrance type
20	IL6	346	TNF	rheumatoid arthritis, systemic juvenile
21	FGFR3	304	FGFR2	muenke syndrome
22	INS	384	KCNJ11	diabetes mellitus, permanent neonatal

Table 5.4: The genes that have the strongest interaction as well as the diseases that have the strongest association with top 23 ranked genes selected by EDIL-W when $\alpha = 1.5$

Rank	Gene	Degree	The most interacted gene	The most associated disease
23	IL1B	293	KCNJ11	alzheimer's disease

The DAVID application gives us 560 recodes. Because of the large number of records, we filter the records by ignoring some annotation categories including *COG-ONTOLOGY, SP-PIR-KEYWORDS, and UP-SEQ-FEATURE*. In addition, we set arbitrary a minimum number of genes involved in the corresponding term as 7. Then, 48 annotation terms and their related genes are sorted by p-value from top to bottom (Table 5.5). The low p-value confirms the importance of our findings.

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	Genes	p-Value
regulation of protein kinase cascade	ERCC6, FLNA, INS, IL1B, IL6, PTEN, TGFB1, TNF	8.9E-8
MAPKKK cascade	ERCC6, FGFR1, FGFR3, INS, IL1B, TNF, BRAF	3.6E-7

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
skeletal system development	GNAS, TBX1, COL2A1, FBN1, FGFR1, FGFR3, INS, TGFB1	4.8E-7
negative regulation of apoptosis	COL2A1, APOE, INS, IL1B, IL6, PTEN, TNF, BRAF	9.6E-7
negative regulation of programmed cell death	COL2A1, APOE, INS, IL1B, IL6, PTEN, TNF, BRAF	1.1E-6
negative regulation of cell death	COL2A1, APOE, INS, IL1B, IL6, PTEN, TNF, BRAF	1.1E-6
positive regulation of transport	FLNA, APOE, INS, IL1B, IL6, TGFB1, TNF	1.1E-6
transmembrane receptor protein tyrosine kinase signaling pathway	FGFR1, FGFR2, FGFR3, FLNA, INS, PTEN, TGFB1	1.2E-6
regulation of cell proliferation	FGFR1, FGFR2, FGFR3, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF	1.8E-6
regulation of apoptosis	COL2A1, ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF, BRAF	2.2E-6

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
regulation of programmed cell death	COL2A1, ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF, BRAF	2.4E-6
regulation of cell death	COL2A1, ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF, BRAF	2.4E-6
positive regulation of cell proliferation	FGFR1, FGFR2, FGFR3, INS, IL1B, IL6, TGFB1, TNF	2.7E-6
regulation of phosphorylation	ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF	6.0E-6
regulation of phosphorus metabolic process	ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF	7.8E-6
regulation of phosphate metabolic process	ERCC6, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF	7.8E-6
intracellular signaling cascade	ALMS1, GNAS, ERCC6, FGFR1, FGFR3, FLNA, APOE, INS, IL1B, TNF, BRAF	1.0E-5
MAPK signaling pathway	FGFR1, FGFR2, FGFR3, FLNA, IL1B, TGFB1, TNF, BRAF	1.3E-5

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
enzyme linked receptor protein signaling pathway	FGFR1, FGFR2, FGFR3, FLNA, INS, PTEN, TGFB1	1.3E-5
regulation of protein kinase activity	ERCC6, APOE, INS, IL1B, PTEN, TGFB1, TNF	1.4E-5
extracellular region	GNAS, COL2A1, ELN, FBN1, FGFR2, FLNA, APOE, INS, IL1B, IL6, POMC, TGFB1, TNF	1.6E-5
regulation of kinase activity	ERCC6, APOE, INS, IL1B, PTEN, TGFB1, TNF	1.7E-5
response to hormone stimulus	GNAS, INS, IL1B, IL6, PTEN, TGFB1, TNF	2.0E-5
protein kinase cascade	ERCC6, FGFR1, FGFR3, INS, IL1B, TNF, BRAF	2.1E-5
regulation of transferase activity	ERCC6, APOE, INS, IL1B, PTEN, TGFB1, TNF	2.1E-5
positive regulation of molecular function	GNAS, ERCC6, APOE, INS, IL1B, IL6, TGFB1, TNF	2.7E-5
response to endogenous stimulus	GNAS, INS, IL1B, IL6, PTEN, TGFB1, TNF	3.4E-5

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
identical protein binding	GNAS, TBX1, COL2A1, FGFR3, FLNA, APOE, TGFB1, TNF	4.3E-5
positive regulation of nitrogen compound metabolic process	TBX1, ERCC6, APOE, INS, IL1B, IL6, TGFB1, TNF	4.9E-5
protein amino acid phosphorylation	ERCC6, FGFR1, FGFR2, FGFR3, IL1B, TGFB1, TNF, BRAF	6.1E-5
extracellular space	COL2A1, FBN1, APOE, INS, IL1B, IL6, TGFB1, TNF	7.4E-5
extracellular region part	COL2A1, ELN, FBN1, APOE, INS, IL1B, IL6, TGFB1, TNF	8.3E-5
regulation of cellular protein metabolic process	ERCC6, APOE, INS, IL1B, IL6, TGFB1, TNF	8.3E-5
phosphate metabolic process	ERCC6, FGFR1, FGFR2, FGFR3, IL1B, PTEN, TGFB1, TNF, BRAF	9.0E-5
response to organic substance	GNAS, APOE, INS, IL1B, IL6, PTEN, TGFB1, TNF	1.0E-4
positive regulation of catalytic activity	GNAS, ERCC6, APOE, INS, IL1B, TGFB1, TNF	1.4E-4

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
phosphorylation	ERCC6, FGFR1, FGFR2, FGFR3, IL1B, TGFB1, TNF, BRAF	1.9E-4
positive regulation of macromolecule metabolic process	TBX1, ERCC6, APOE, INS, IL1B, IL6, TGFB1, TNF	2.9E-4
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	TBX1, ERCC6, APOE, INS, IL6, TGFB1, TNF	3.7E-4
neurological system process	ALMS1, GNAS, TBX1, COL2A1, ERCC6, APOE, IL1B, PTEN, TGFB1	4.1E-4
pathways in cancer	FGFR1, FGFR2, FGFR3, IL6, PTEN, TGFB1, BRAF	4.6E-4
positive regulation of cellular biosynthetic process	TBX1, APOE, INS, IL1B, IL6, TGFB1, TNF	6.1E-4
positive regulation of biosynthetic process	TBX1, APOE, INS, IL1B, IL6, TGFB1, TNF	6.6E-4
cell death	ALMS1, INS, IL1B, IL6, PTEN, TGFB1, TNF	7.9E-4

Table 5.5: Functional annotation chart of top 23 ranked genes selected by EDIL-W with $\alpha = 1.5$

Term	**Genes**	**p-Value**
death	ALMS1, INS, IL1B, IL6, PTEN, TGFB1, TNF	8.2E-4
homeostatic process	ALMS1, COL2A1, APOE, INS, IL1B, IL6, TGFB1	9.9E-4
cell surface receptor linked signal transduction	GNAS, FGFR1, FGFR2, FGFR3, FLNA, APOE, INS, PTEN, POMC, TGFB1	1.6E-3
cognition	ALMS1, GNAS, TBX1, COL2A1, ERCC6, IL1B, PTEN	2.7E-3

Conclusion

To conclude, we construct a weighted gene-disease bipartite network to represent the associations among genes and diseases. Then, we construct two new networks, called the weighted human disease network (WHDN) and the weighted gene network (WGN). In addition, we propose a new centrality measure, called DIL-W, to find the most important diseases in the WHDN and extend the DIL-W method, called EDIL-W, to identify the most important genes in the WGN. Our network-based analysis methods are shown to be able to identify more important diseases and genes in terms of network efficiency compared to degree, closeness, and betweenness centralities. The identified disease-disease correlations include previous knowledge supported by published literature as well as less known and novel correlations that can be valuable for future studies. Meanwhile, the identified gene-gene interactions are supported by DAVID through functional annotation and enrichment analysis.

Our understanding of complex human diseases is still largely unclear, and the disease-gene associations are far from being complete. Future studies could explore the utilization of multiple types of data and more powerful computational tools to better cluster and categorize human diseases and to predict new genes and other factors that can explain diseases.

Bibliography

[1] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: generalizing degree and shortest paths. *Social Networks*, 32(3):245–251, 2010.

[2] Peilin Jia and Zhongming Zhao. Network-assisted analysis to prioritize gwas results: principles, methods and perspectives. *Human genetics*, 133(2):125–138, 2014.

[3] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[4] Gang Zheng, Mark Meyer, Wentian Li, and Yaning Yang. Comparison of two-phase analyses for case–control genetic association studies. *Statistics in medicine*, 27(24):5054–5075, 2008.

[5] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

[6] Jason H Moore and Scott M Williams. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*, 85(3):309–320, 2009.

[7] Anna L Tyler, Folkert W Asselbergs, Scott M Williams, and Jason H Moore. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays*, 31(2):220–227, 2009.

[8] Jin Li, Dongli Huang, Maozu Guo, Xiaoyan Liu, Chunyu Wang, Zhixia Teng, Ruijie Zhang, Yongshuai Jiang, Hongchao Lv, and Limei Wang. A gene-based information gain method for detecting gene–gene interactions in case–control studies. *European Journal of Human Genetics*, 23(11):1566–1572, 2015.

[9] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[10] Heather J Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[11] Peter M Visscher, Gibran Hemani, Anna AE Vinkhuyzen, Guo-Bo Chen, Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Jian Yang. Statistical power to detect genetic (co) variance of complex traits using snp data in unrelated samples. *PLoS genetics*, 10(4):e1004269, 2014.

[12] Patrick C Phillips. Epistasisthe essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[13] C Clark Cockerham. An extension of the concept of partitioning hereditary

variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39(6):859, 1954.

[14] Oscar Kempthorne. The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London B: Biological Sciences*, 143(910):103–113, 1954.

[15] Holger Schwender and Katja Ickstadt. Identification of snp interactions using logic regression. *Biostatistics*, 9(1):187–198, 2007.

[16] Changzheng Dong, Xun Chu, Ying Wang, Yi Wang, Li Jin, Tieliu Shi, Wei Huang, and Yixue Li. Exploration of gene–gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16(2):229–235, 2008.

[17] Guolian Kang, Weihua Yue, Jifeng Zhang, Yuehua Cui, Yijun Zuo, and Dai Zhang. An entropy-based approach for testing genetic epistasis underlying complex diseases. *Journal of theoretical biology*, 250(2):362–374, 2008.

[18] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[19] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature genetics*, 39(9):1167–1173, 2007.

[20] Xia Jiang, M Michael Barmada, and Shyam Visweswaran. Identifying genetic

interactions in genome-wide data using bayesian networks. *Genetic epidemiology*, 34(6):575–581, 2010.

[21] Xiang Chen, Ching-Ti Liu, Meizhuo Zhang, and Heping Zhang. A forest-based approach to identifying gene and gene–gene interactions. *Proceedings of the National Academy of Sciences*, 104(49):19199–19203, 2007.

[22] Daniel F Schwarz, Inke R König, and Andreas Ziegler. On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26(14):1752–1758, 2010.

[23] Ching Lee Koo, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed research international*, 2013, 2013.

[24] Rosanna Upstill-Goddard, Diana Eccles, Joerg Fliege, and Andrew Collins. Machine learning approaches for the discovery of gene–gene interactions in disease data. *Briefings in bioinformatics*, 14(2):251–260, 2012.

[25] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3s):228, 2003.

[26] Kasper Lage, E Olof Karlberg, Zenia M Størling, Páll I Olason, Anders G Pedersen, Olga Rigina, Anders M Hinsby, Zeynep Tümer, Flemming Pociot, Niels Tommerup, et al. A human phenome-interactome network of protein

complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316, 2007.

[27] Xuebing Wu, Qifang Liu, and Rui Jiang. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics*, 25(1):98–104, 2008.

[28] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(1):189, 2008.

[29] Fredrik Barrenas, Sreenivas Chavali, Petter Holme, Reza Mobini, and Mikael Benson. Network properties of complex human disease genes identified through genome-wide association studies. *PloS One*, 4(11):e8090, 2009.

[30] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641, 2010.

[31] Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, 10(5):280–293, 2011.

[32] Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536, 2012.

[33] Silpa Suthram, Joel T Dudley, Annie P Chiang, Rong Chen, Trevor J Hastie, and Atul J Butte. Network-based elucidation of human disease similarities

reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6(2):e1000662, 2010.

[34] Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Xiaoqing Peng, Fang-Xiang Wu, and Yi Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.

[35] Annie P Chiang and Atul J Butte. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, 86(5):507–510, 2009.

[36] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1):496, 2011.

[37] Hailin Chen, Heng Zhang, Zuping Zhang, Yiqin Cao, and Wenliang Tang. Network-based inference methods for drug repositioning. *Computational and Mathematical Methods in Medicine*, 2015, 2015.

[38] Mark EJ Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, 2001.

[39] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[40] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[41] Ting Hu, Nicholas A. Sinnott-Armstrong, Jeff W. Kiralis, Angeline S. Andrew, Margaret R. Karagas, and Jason H. Moore. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 12:364, 2011.

[42] Ting Hu, Yuanzhu Chen, Jeff W. Kiralis, and Jason H. Moore. ViSEN: Methodology and software for visualization of statistical epistasis networks. *Genetic Epidemiology*, 37:283–285, 2013.

[43] Tianshu Yin, Shu Chen, Xiaohui Wu, and Weidong Tian. GenePANDA a novel network-based gene prioritizing tool for complex diseases. *Scientific Reports*, 7, 2017.

[44] Björn H Junker, Dirk Koschützki, and Falk Schreiber. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, 7(1):219, 2006.

[45] Tim Kacprowski, Nadezhda T Doncheva, and Mario Albrecht. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–1473, 2013.

[46] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.

[47] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of

the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[48] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[49] Martin Oti, Berend Snel, Martijn A Huynen, and Han G Brunner. Predicting disease genes using protein–protein interactions. *Journal of Medical Genetics*, 43(8):691–698, 2006.

[50] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[51] David A Fell and Andreas Wagner. The small world of metabolism. *Nature Biotechnology*, 18(11):1121–1122, 2000.

[52] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.

[53] Pea Carninci, T Kasukawa, S Katayama, J Gough, MC Frith, N Maeda, R Oyama, T Ravasi, B Lenhard, C Wells, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, 2005.

[54] Rune Linding, Lars Juhl Jensen, Adrian Pasculescu, Marina Olhovsky, Karen Colwill, Peer Bork, Michael B Yaffe, and Tony Pawson. Networkin: a resource for exploring cellular phosphorylation networks. *Nucleic acids research*, 36(suppl_1):D695–D699, 2007.

[55] Angela Reynolds, Devin Leake, Queta Boese, Stephen Scaringe, William S Marshall, and Anastasia Khvorova. Rational sirna design for rna interference. *Nature biotechnology*, 22(3):326, 2004.

[56] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *cell*, 120(1):15–20, 2005.

[57] Mark Newman. *Networks: an Introduction*. Oxford university press, 2010.

[58] Chao Wu, Jun Zhu, and Xuegong Zhang. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13(1):182, 2012.

[59] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, 21(6):697, 2003.

[60] Ting Hu, Nicholas A Sinnott-Armstrong, Jeff W Kiralis, Angeline S Andrew, Margaret R Karagas, and Jason H Moore. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC bioinformatics*, 12(1):364, 2011.

[61] Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285, 2008.

[62] Young-Rae Cho and Aidong Zhang. Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. In *BMC bioinformatics*, volume 11, page S3. BioMed Central, 2010.

[63] Andrey Rzhetsky, David Wajngurt, Naeun Park, and Tian Zheng. Probing genetic overlap among complex human phenotypes. *Proceedings of the National Academy of Sciences*, 104(28):11694–11699, 2007.

[64] César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353, 2009.

[65] D-S Lee, J Park, KA Kay, NA Christakis, ZN Oltvai, and A-L Barabási. The implications of human metabolic network topology for disease comorbidity. *Proceedings of the National Academy of Sciences*, 105(29):9880–9885, 2008.

[66] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[67] Charles R Scriver. Allelic and locus heterogeneity. *ELS*, 2005.

[68] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online Mendelian Inheritance in Man (OMIM), a knowl-

edge base of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[69] XueZhong Zhou, Jörg Menche, Albert-Lászlό Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5:4212, 2014.

[70] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(suppl_1):D13–D21, 2007.

[71] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[72] Víctor Martínez, Carlos Cano, and Armando Blanco. ProphNet: A generic prioritization method through propagation of information. *BMC Bioinformatics*, 15(1):S5, 2014.

[73] Jun Liu, Qingyu Xiong, Weiren Shi, Xin Shi, and Kai Wang. Evaluating the importance of nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, 452:209–219, 2016.

[74] Daniela Nitsch, Leon-Charles Tranchevent, Joana P Goncalves, Josef Korbinian Vogt, Sara C Madeira, and Yves Moreau. Pinta: a web server for network-based gene prioritization from expression data. *Nucleic acids research*, 39(suppl_2):W334–W338, 2011.

[75] Laszlo Lovasz. Random walks on graphs: A survey. *Combinatorics, Paul Erdos in Eighty*, 2, 1993.

[76] Ping Hu, Wenli Fan, and Shengwei Mei. Identifying node importance in complex networks. *Physica A: Statistical Mechanics and its Applications*, 429:169–176, 2015.

[77] Zhao YihuanWang ZulinZheng JingGuo Xujing. Finding most vital node by node importance contribution matrix in communication netwoks [j]. *Journal of Beijing University of Aeronautics and Astronautics*, 9:009, 2009.

[78] Tingyuan Nie, Zheng Guo, Kun Zhao, and Zhe-Ming Lu. Using mapping entropy to identify node centrality in complex networks. *Physica A: Statistical Mechanics and its Applications*, 453:290–297, 2016.

[79] Xiangbin Yan, Li Zhai, and Weiguo Fan. C-index: A weighted network node centrality measure for collaboration competence. *Journal of Informetrics*, 7(1):223–239, 2013.

[80] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569, 2005.

[81] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, 2017.

[82] DisGeNET. http://www.disgenet.org/web/disgenet/menu/home, 2017. [Online; accessed 8-December-2017].

[83] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

[84] Christian Darabos, Marquitta J White, Britney E Graham, Derek N Leung, Scott M Williams, and Jason H Moore. The multiscale backbone of the human phenotype network based on biological pathways. *BioData Mining*, 7(1):1, 2014.

[85] Zhuo-Ming Ren, Feng Shao, Jian-Guo Liu, Qiang Guo, and Bing-Hong Wang. Node importance measurement based on the degree and clustering coefficient information. *Acta Phys. Sin.*, 6:128901, 2013.

[86] LH Savin. Atypical retinitis pigmentosa associated with obesity, polydactyly, hypogenitalism, and mental retardation (the Laurence-Moon-Biedl Syndrome)(clinical and genealogical notes on a case). *The British Journal of Ophthalmology*, 19(11):597, 1935.

[87] David Dimmock, Keiko Kobayashi, Mikio Iijima, Ayako Tabata, Lee-Jun Wong, Takeyori Saheki, Brendan Lee, and Fernando Scaglia. Citrin deficiency: a novel cause of failure to thrive that responds to a high-protein, low-carbohydrate diet. *Pediatrics*, 119(3):e773–e777, 2007.

[88] Hyman J Zimmerman. Drug-induced liver disease. *Clinics in Liver Disease*, 4(1):73–96, 2000.

[89] Fiona C Mansergh, Sophia Millington-Ward, Avril Kennan, Anna-Sophia Kiang, Marian Humphries, G Jane Farrar, Peter Humphries, and Paul F Kenna. Retinitis pigmentosa and progressive sensorineural hearing loss caused by a C12258A mutation in the mitochondrial MTTS2 gene. *The American Journal of Human Genetics*, 64(4):971–985, 1999.

[90] Denise Rossato Silva, Ana Cláudia Coelho, Anelise Dumke, Jorge Diego Valentini, Juliana Nunes de Nunes, Clarisse Luisa Stefani, Lívia Fontes da Silva Mendes, and Marli Maria Knorst. Osteoporosis prevalence and associated factors in patients with COPD: a cross-sectional study. *Respiratory Care*, 56(7):961–968, 2011.

[91] Sarah E Stolz, Gian-Emilio Chatrian, and Alexander M Spence. Epileptic nystagmus. *Epilepsia*, 32(6):910–918, 1991.

[92] American Optometric Association. https://www.aoa.org/, 2017. [Online; accessed 30-December-2017].

[93] Satish Kenchaiah, Jane C Evans, Daniel Levy, Peter WF Wilson, Emelia J Benjamin, Martin G Larson, William B Kannel, and Ramachandran S Vasan. Obesity and the risk of heart failure. *New England Journal of Medicine*, 347(5):305–313, 2002.

[94] Lewis P Rowland. Diagnosis of amyotrophic lateral sclerosis. *Journal of the Neurological Sciences*, 160:S6–S24, 1998.

[95] W Rodger. Non-insulin-dependent (type II) diabetes mellitus. *CMAJ: Canadian Medical Association Journal*, 145(12):1571, 1991.

[96] JHD Millar. Epilepsy and strabismus. *Epilepsia*, 6(1):43–46, 1965.

[97] Sarah L Czerwinski, Caryn E Plummer, Shari M Greenberg, William F Craft, Julia A Conway, Mayrim L Perez, Kirsten L Cooke, and Matthew D Winter. Dynamic exophthalmos and lateral strabismus in a dog caused by masticatory muscle myositis. *Veterinary Ophthalmology*, 18(6):515–520, 2015.

[98] Patrick E Brookhouser. Sensorineural hearing loss in children. *Pediatric Clinics of North America*, 43(6):1195–1216, 1996.

[99] Flemming Nørgaard. Earliest roentgenological changes in polyarthritis of the rheumatoid type: rheumatoid arthritis. *Radiology*, 85(2):325–329, 1965.

[100] MI Botez, Ezzedine Attig, and Jean Lorrain Vézina. Cerebellar atrophy in epileptic patients. *CanadianJournal of Neurological Sciences*, 15(3):299–303, 1988.

[101] Gerald Weissmann. Rheumatoid arthritis and systemic lupus erythematosus as immune complex diseases. *Bulletin of the NYU Hospital for Joint Diseases*, 67(3):251, 2009.

[102] Osamu Sato, Tsuyoshi Yamguchi, Mamoru Kittaka, and Hiroyuki Toyama. Hydrocephalus and epilepsy. *Child's Nervous System*, 17(1):76–86, 2001.

[103] Elizabeth G Nabel and Eugene Braunwald. A tale of coronary artery disease and myocardial infarction. *New England Journal of Medicine*, 366(1):54–63, 2012.

[104] Neil Kaplowitz. Drug-induced liver injury. *Clinical Infectious Diseases*, 38(Supplement_2):S44–S48, 2004.

[105] Elena Galli, Simona Gianni, Giovanni Auricchio, Ercole Brunetti, Giorgio Mancino, and Paolo Rossi. Atopic dermatitis and asthma. In *Allergy and Asthma Proceedings*, volume 28, pages 540–543. OceanSide Publications, Inc, 2007.

[106] K Arumugam. Endometriosis and obesity. *Journal of Obstetrics and Gynaecology*, 12(4):266–268, 1992.

[107] Robert Gajarski, David C Naftel, Elfriede Pahl, Juan Alejos, F Bennett Pearce, James K Kirklin, Mary Zamberlan, Anne I Dipchand, Pediatric Heart Transplant Study Investigators, et al. Outcomes of pediatric patients with hypertrophic cardiomyopathy listed for transplant. *The Journal of Heart and Lung Transplantation*, 28(12):1329–1334, 2009.

[108] Debara L Tucci, Donald E Born, and Edwin W Rubel. Changes in spontaneous activity and CNS morphology associated with conductive and sensorineural hearing loss in chickens. *Annals of Otology, Rhinology & Laryngology*, 96(3):343–350, 1987.

[109] Eugenio Fiorentino, Gianni Pantuso, Alessia Cusimano, Stefania Latteri, Achille Mastrosimone, and Calogero Cipolla. Gastro-oesophageal reflux and epileptic attacks: casually associated or related efficiency of antireflux surgery. *Chirurgia Italiana*, 58(6):689–696, 2008.

[110] Anja Kurt, Femke Nijboer, Tamara Matuz, and Andrea Kübler. Depression and

anxiety in individuals with amyotrophic lateral sclerosis. *CNS Drugs*, 21(4):279–291, 2007.

[111] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2008.

[112] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.