Measuring the Importance of Vertices in the Weighted Human Disease Network

Seyed Mehrzad Almasi[1], Ting Hu[1*]

1 Department of Computer Science, Memorial University, St. John's, NL, Canada

* ting.hu@mun.ca

Abstract

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 provides an important avenue to better understand the common genetic background of diseases and to help develop new drugs that can treat multiple diseases. Meanwhile, network science has seen increasing applications on modeling complex biological systems, and can be a powerful tool to elucidate the correlations of multiple human diseases. In this article, known disease-gene associations were represented using a weighted bipartite network. We extracted a weighted human diseases network from such a bipartite network to show the correlations of diseases. Subsequently, we proposed a new centrality measurement for the weighted human disease network in order to quantify the importance of diseases. Using our centrality measurement to quantify the importance of vertices in the weighted human disease network, we were able to find a set of most central diseases. By investigating the 30 top diseases and their most correlated neighbors in the network, we identified disease linkages including known disease pairs and novel findings. Our research helps better understand the common genetic origin of human diseases and suggests top diseases that likely induce other related diseases.

Author summary

Introduction

During the past decades, significant progress has been made in our understanding of human diseases [1]. However, the genetic architectures of complex diseases are still largely unclear. Many common diseases tend to be related to each other, and it is suspected that they may share common genetic origin. Thus, studying the correlations of human diseases has the potentials of better understanding the genotype to phenotype mapping [2, 3] and better predicting disease association genes [4, 5, 6, 7, 8]. Furthermore, learning which diseases are correlated can help use existing drugs to treat multiple similar diseases [9, 10, 11, 12, 13].

Meanwhile, network science is a rising field where entities and their complex relationships are studied on a global scale [14, 15, 16], and has seen increasing applications to perform advanced analysis on biomedical data [17, 18, 19, 20, 21, 22]. There are various cellular components in the human body that interact with each other within the same cell or across different cells [15]. 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. Those cellular components can be proteins or metabolites, and the network refers to protein-protein interaction (PPI) network [23, 24, 25] or metabolic network [26, 27, 28].

Some studies aimed at identifying the correlations among diseases through network analysis [15, 29, 30]. Goh *et al.* [31] 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 cancer cluster and neurological disease cluster. Zhou *et al.* [32] extracted over twenty million bibliographic records from PubMed [33] in order to obtain 147,978 connections between 322 symptoms and 4,219 diseases. A human symptoms-disease network (HSDN) was then constructed and was able to show the symptom similarity between all pairs of diseases (7,488,851 links) in the network. The weight of links 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.* [34] built a disease metabolism network in order to study disease comorbidity for better disease prediction and prevention. Two diseases are connected with each other if a mutated enzyme catalyzes 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.

Measuring the centrality of vertices helps identify important vertices in the network in terms of connecting to all other vertices. Centrality measures have been used frequently to analyze biological networks over the past decades [35, 36, 37]. 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) [38]. 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 network of interest, in this study, the human disease network.

Köhler *et al.* [39] 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 comparing with conventional distance-based centrality measures. Wu *et al.* [40] 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. Martinez *et al.* [41] 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 diseases.

In the article, we propose a new method for the construction of a weighted human disease network (WHDN) and a new centrality measure to identify the most important diseases. First we use a large database of disease-gene associations to build a weighted bipartite disease-gene network, and then construct a weighted disease network where link weights capture the strengths of the pairwise disease correlations. After the backbone extraction of the WHDN, we design a centrality measure for the context of the WHDN that considers not only the degree of a vertex but also the importance of its

incident edges. Finally, we compare our new centrality measure with degree, closeness and betweenness by evaluating the network efficiency decline rate with the removal of top-ranked vertices by each centrality measurement.

Methods and Results

Given the multiple-step pipeline structure of this study, we show the result of each step after the description of the corresponding method.

Disease-Gene Associations (DGAs)

The data used in this project contains disease-gene associations (DGAs) from multiple curated databases including UNIPROT, CTD (human subset), PsyGeNET, Orphanet, and HPO. The disease-gene association data are conducted by DisGeNet group, available on DisGeNET v4.0 [42]. The current version of the data set contains 130,821 DGAs, between 13,075 diseases and 8,949 genes. Each DGA is assigned with 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 [42]. We first 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 we would like to consider the correlation among all diseases, 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 [42] for this filtering.

Network Construction

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 different sets of vertices. One set includes diseases and another one contains 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. Their link weight is assigned as the score a_i^k, for disease i and gene k, computed in the DGA database described in the previous section. A sample subgraph of the bipartite network is shown in Figure 1.

Figure 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 2 a), the degree distribution of the diseases is right-skewed and approximately follows a heavy-tailed 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 giant component. Figure 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. Figure 3 shows that there is a considerable number of components with two vertices, i.e., 844 isolated disease-gene pairs. Since we are interested in investigating the

Fig 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.

Fig 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.

Fig 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.

large-scale genetic correlations of human diseases, we focus the giant component of the disease-gene bipartite network in the downstream analyses.

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 [14], 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 authors 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, first we consider that the correlation of two diseases through a gene is stronger when they are the sole associated diseases with this gene than when there are many other diseases associated with the same gene. Second, the correlation of two diseases is considered stronger when they share more genes through stronger associations than less genes or weaker associations. Thus, we extend Newman's method to weighted graph and define the weight of edge w_{ij} between two diseases i and j as

$$w_{ij} = \sum_{k \in G} \frac{\delta_i^k \delta_j^k (a_i^k + a_j^k)}{s_k}, \tag{1}$$

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,

$$s_k = \sum_{i \in D} a_i^k. \tag{2}$$

Fig 4. Distribution of edge weights in the WHDN. The weight of an edge quantifies the shared genetic background of two connected diseases. There are 112,324 edges in the graph with weights ranging from 0.0152 to 22.4506.

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

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

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 11,2324 edges and an average vertex degree of 42.56. That is, a disease correlates with on average 42.56 other diseases with varying strengths. Figure 4 depicts the distribution of all the edge weights in the WHDN. As we can see that a large number of edge weights are of small values and may not be particularly interesting for the subsequent analysis. Those weak edges not only add computational overhead to the network analysis, but also render the network difficult to interpret. Therefore, next we perform an edge reduction and only extract the most meaningful structure of the network.

The Multi-Scale Backbone of WHDN

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. Here, to preserve the multi-scale backbone of the weighted human disease network (WHDN) while removing less relevant and

meaningful edges we use a multi-scale filtering method proposed by Serrano *et al.* [43]. Such a multi-scale backbone exaction algorithm has been used to reduce the network size while preserving the meaningful structure of biological networks in multiple studies [32, 44, 45, 46].

First, the weight of edge linking vertex i with its neighbor j can be normalized as

$$p_{ij} = \frac{w_{ij}}{s_i}, \tag{3}$$

where s_i is the vertex strength, i.e., the sum of weights incident to vertex i, defined as

$$s_i = \sum_{j \in \prod(i)} w_{ij}, \tag{4}$$

where $\prod(i)$ 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 the denominator.

Second, a null model is used to assess the expectation if the weights of links connecting to a particular vertex were distributed randomly. That is, the normalized weight p_{ij} that corresponds to the link connecting to a certain vertex of degree k is produced by a random assignment from an uniform distribution. Thus the probability density function for the variable taking a particular value x is

$$p(x)dx = (k-1)(1-x)^{k-2}dx. \tag{5}$$

Then, to identify whether the probability, β_{ij}, of link weight p_{ij} is compatible with the null model with a threshold β is given as

$$\beta_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}dx < \beta. \tag{6}$$

All links with computed β_{ij} lower than a given threshold 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.* [44] empirically found that the AND rule preserve the network features better than using the OR rule in the context of human phenotype networks. In this article, the AND rule is adopted to reduce the size of the network 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. Figure 5 shows the results of network metrics as a function of β cutoffs. 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 $\beta = 0.501$ can be determined, shown as the vertical dashed line in the figure.

After the backbone extraction, the WHDN has 4,898 vertices and 38,275 edges. Those vertices are no longer connected in a single component. Figure 6 shows the size distribution of its connected components. There is a giant component with 4,810 vertices and its degree distribution is shown in Figure 7. Again the degree distribution is heavy tailed and resembles a power-law relationship. The vertex *epilepsy* has the highest degree of 576. This giant component 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.

Fig 5. Choosing the β value. 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.

Fig 6. The size distribution of connected components in the extracted backbone of the WHDN. The network has a single giant component with 4,810 vertices.

Fig 7. Degree distribution of vertices in the giant component of the extracted backbone of the WHDN. The distribution is shown on a log-log scale.

Fig 8. An example weighted graph.

Measuring Vertex Importance in WHDN

Although various vertex centrality measures have been proposed in the literature, the quantification of the importance of a vertex in a network is often context-specific. For some networks, measuring degree may suffice since a vertex can be considered important when its number of neighbors is the sole criterion. For some networks, e.g., information communication networks, a vertex may be considered more important if its distances to all other vertices are short, then closeness centrality serves this purpose well. For our WHDN, a disease is considered important if it correlates with many other diseases (degree) as well as if the correlations are themselves very important (edge importance).

We propose a vertex importance measure for the weighted human disease network (WHDN) by extending a centrality measure for unweighted networks proposed by Liu *et al.* [47]. 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:

$$I_{e_{ij}} = \frac{U_{e_{ij}}}{\lambda_{e_{ij}}}, \tag{7}$$

where $U_{e_{ij}} = (k_i - p - 1)(k_j - p - 1)$ and $\lambda_{e_{ij}} = \frac{p}{2} + 1$. Following the convention, k_i and k_j are the degrees of vertex v_i and v_j, respectively, and p is the number of triangles with one edge being e_{ij}.

Subsequently, the contribution that vertex v_i makes to the importance of e_{ij} is computed as

$$C_{v_i v_j} = I_{e_{ij}} \times \frac{k_i - 1}{k_i + k_j - 2}, \tag{8}$$

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,

$$\mathrm{DIL}_{v_i} = k_i + \sum_{v_j \in \Gamma_i} C_{v_i v_j}. \tag{9}$$

For weighted networks, we modify the computation of U in Equation (7) as

$$U_{e_{ij}} = (s_i - p_i) \times (s_j - p_j), \tag{10}$$

where s_i is the strength of vertex v_i, calculated by Formula (4), and p_i is the weight sum of links 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 (10).

We define λ for weighted graphs as

$$\lambda_{e_{ij}} = \frac{p_i + p_j}{2} + 1. \tag{11}$$

Finally, the importance of a vertex can be measured by

$$\text{DIL-W}_{v_i} = s_i + \sum_{v_j \in \Gamma_i} C_{v_i v_j}, \tag{12}$$

where $C_{v_i v_j}$ is defined as

$$C_{v_i v_j} = I_{e_{ij}} \times \frac{s_i}{s_i + s_j}. \tag{13}$$

In the weighted graph given in Figure 8, 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,

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

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

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

Fig 9. Distribution of DIL-W centrality in the giant component of the WHDN on a log-log scale.

Fig 10. Correlation of DIL-W scores with a) degree centrality, b) closeness centrality, and c) betweenness centrality in the WHDN.

Similarly, we can compute the DIL-W centrality of vertex b DIL-W$_b = 2.8916$. Therefore, 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 WHDN, the distribution is shown in Figure 9. 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 can have scores that are greater by orders of magnitude.

Comparison and Evaluation

We compare our DIL-W measurement with three most commonly used centralities, i.e., degree, closeness, and betweenness, when applied to the giant component of the backbone of WHDN. For weighted graphs, degree centrality is calculated as vertex strength given by Equation (4). Closeness and betweenness are shortest-path-based centralities. Shortest path computation can be extended for weighted graph as follows,

$$d_{ij}^{w} = min(\frac{1}{w_{ih}} + ... + \frac{1}{w_{hj}}). \tag{14}$$

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 suggests 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 10 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 DIL-W measure and all other 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. In the context of the WHDN, the network efficiency indicates the extend to which the original connectivity of the network is maintained. We calculate the decline rate of network efficiency after removing m top-rank vertices. The network efficiency [48] is computed based on the connectivity of a network. A higher connectivity suggests a higher network efficiency. The network efficiency is defined by

$$\eta = \frac{1}{n(n-1)} \sum_{v_i \neq v_j \in V} \frac{1}{d_{ij}}, \tag{15}$$

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

$$\mu = 1 - \frac{\eta}{\eta_0}, \tag{16}$$

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

Fig 11. 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.

Fig 12. 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.

When a more importance 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 removing a vertex in the network. Figure 11 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. As shown in the figure, we do not observe a monotonic relationship across all four centrality methods. However, the correlation analysis shows that our method, DIL-W, has a slighter stronger negative correlation between the decline rate and the rank of the removed vertex than the other three. The Spearman's rank correlation coefficient, ρ, for degree, closeness, and betweenness is -0.1801, -0.0017, and -0.0679, respectively. In comparison, DIL-W has a negative correlation coefficient -0.2698.

We also consider removing all m top-rank vertices at once and see how this accumulative removal affects the efficiency of the network. Figure 12 shows the decline rate of the network efficiency after removing all 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 12, the four methods are very comparable 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 around point 30 and becomes very comparable afterwards.

Table 1 shows the top 30 diseases ranked by our DIL-W method, their degrees, and their neighbors that have the strongest correlations (i.e., edge weights). References that support the known comorbidity of the disease pairs are also given.

Table 1. The 30 top-ranked diseases by DIL-W and their most correlated dieases

Rank	Disease	Degree	The most correlated disease	Ref.
1	Epilepsy	576	Pediatric failure to thrive	–
2	Pediatric failure to thrive	462	Epilepsy	–
3	Sensorineural hearing loss (disorder)	313	Retinitis pigmentosa	[54]
4	Anemia	327	Pediatric failure to thrive	[50]
5	Obesity	268	Retinitis Pigmentosa	[49]
6	Osteoporosis	326	Osteopenia	[55]
7	Nystagmus	276	Epilepsy	[56]
8	Liver cirrhosis	278	Chemical and drug induced liver injury	[51]
9	Low vision	270	Nystagmus	[57]
10	Heart failure	311	Obesity	[58]
11	Muscle degeneration	277	Amyotrophic lateral sclerosis	[59]

Table 1. The 30 top-ranked diseases by DIL-W and their most correlated dieases

Rank	Disease	Degree	The most correlated disease	Ref.
12	Diabetes mellitus, non-insulin-dependent	245	Obesity	[60]
13	Strabismus	293	Epilepsy	[61]
14	Exophthalmos	302	Strabismus	[62]
15	Myopia	266	Sensorineural hearing loss (disorder)	[63]
16	Degenerative polyarthritis	239	Rheumatoid arthritis	[64]
17	Cerebral atrophy	267	Epilepsy	[65]
18	Optic atrophy	236	Nystagmus	–
19	Rheumatoid arthritis	188	Lupus erythematosus, systemic	[66]
20	Hydrocephalus	250	Epilepsy	[67]
21	Alopecia	241	Dystrophia unguium	–
22	Myocardial ischemia	166	Obesity	–
23	Myocardial infarction	228	Coronary artery disease	[68]
24	Chemical and drug induced liver injury	174	Cholestasis	[69]
25	Asthma	198	Dermatitis, atopic	[70]
26	Endometriosis	135	Obesity	[71]
27	Hypertrophic cardiomyopathy	187	Pediatric failure to thrive	[72]
28	Conductive hearing loss	163	Sensorineural hearing loss (disorder)	[73]
29	Brain ischemia	191	Diabetes mellitus, non-insulin-dependent	–
30	Gastroesophageal reflux disease	190	Epilepsy	[74]

Discussion

In this article, we use a network-based analysis to identify important human diseases that share genetic background with many other diseases through strong associations. We collect a large number of known disease-gene associations (DGAs) using DisGeNET 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 and the edge weights reflect the number of genes they share as well as the strength of the DGAs. Then 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 three commonly used centrality measures, degree, closeness and betweenness, and has identified top diseases including *epilepsy*, *anemia*, and *obesity*.

Table 1 shows the degree in the WHDN and the most correlated disease of those 30 top-rank diseases. We are also able to find previous publications that verify almost all the correlations of those pairs of diseases, shown as references in the table. Besides some very well-known correlations such as *heart failure* - *obesity* and *diabetes* - *obesity*, the table also reports some less known but interesting correlations. For instance, Savin [49] 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.* [50] suggested *anemia* as one of the novel causes of *failure to thrive* in children. Zimmerman [51] 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*.

The disease-gene associations come from DisGeNet [42] only. While this is a valuable resource, it is merely one of the many databases that have disease gene information (including Jensen Lab's DISEASES [52] and DiseaseConnect [53] databases), all of which have their own disease association scoring convention. The alternative databases will be explored in our future study.

Conclusion

Apart from many existing related work, in this article, we construct a *weighted* human disease network (WHDN) and propose a new centrality measure DIL-W designed specifically for the WHDN. Our network-based analysis methods are shown to be able to identify more important diseases comparing 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. Our understanding of 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.

Acknowledgments

This research was supported by the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) RGPIN-2016-04699 to TH. The computation was feasible with the help from the IBM HPC cluster at the Center for Health Informatics & Analytics (CHIA), Faculty of Medicine, Memorial University.

References

1. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nature Genetics. 2003;33(3s):228.

2. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nature Biotechnology. 2007;25(3):309–316.

3. Wu X, Liu Q, Jiang R. Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics. 2008;25(1):98–104.

4. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Molecular Systems Biology. 2008;4(1):189.

5. Barrenas F, Chavali S, Holme P, Mobini R, Benson M. Network properties of complex human disease genes identified through genome-wide association studies. PloS One. 2009;4(11):e8090.

6. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Computational Biology. 2010;6(1):e1000641.

7. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Briefings in Functional Genomics. 2011;10(5):280–293.

8. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics. 2012;13(8):523–536.

9. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Computational Biology. 2010;6(2):e1000662.

10. Luo H, Wang J, Li M, Luo J, Peng X, Wu FX, et al. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics. 2016;32(17):2664–2671.

11. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. Clinical Pharmacology & Therapeutics. 2009;86(5):507–510.

12. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular Systems Biology. 2011;7(1):496.

13. Chen H, Zhang H, Zhang Z, Cao Y, Tang W. Network-based inference methods for drug repositioning. Computational and Mathematical Methods in Medicine. 2015;2015.

14. Newman ME. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E. 2001;64(1):016132.

15. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011;12(1):56–68.

16. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. Cell. 2011;144(6):986–998.

17. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. BMC Bioinformatics. 2011;12:364.

18. Hu T, Chen Y, Kiralis JW, Moore JH. ViSEN: Methodology and software for visualization of statistical epistasis networks. Genetic Epidemiology. 2013;37:283–285.

19. Yin T, Chen S, Wu X, Tian W. GenePANDA — a novel network-based gene prioritizing tool for complex diseases. Scientific Reports. 2017;7.

20. Junker BH, Koschützki D, Schreiber F. Exploration of biological network centralities with CentiBiN. BMC Bioinformatics. 2006;7(1):219.

21. Kacprowski T, Doncheva NT, Albrecht M. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. Bioinformatics. 2013;29(11):1471–1473.

22. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4(1):2.

23. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature. 2005;437(7062):1173–1178.

24. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005;122(6):957–968.

25. Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein–protein interactions. Journal of Medical Genetics. 2006;43(8):691–698.

26. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. 2000;407(6804):651–654.

27. Fell DA, Wagner A. The small world of metabolism. Nature Biotechnology. 2000;18(11):1121–1122.

28. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences. 2007;104(6):1777–1782.

29. Rzhetsky A, Wajngurt D, Park N, Zheng T. Probing genetic overlap among complex human phenotypes. Proceedings of the National Academy of Sciences. 2007;104(28):11694–11699.

30. Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Computational Biology. 2009;5(4):e1000353.

31. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proceedings of the National Academy of Sciences. 2007;104(21):8685–8690.

32. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms–disease network. Nature Communications. 2014;5:4212.

33. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2007;36(suppl_1):D13–D21.

34. Lee DS, Park J, Kay K, Christakis N, Oltvai Z, Barabási AL. The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences. 2008;105(29):9880–9885.

35. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene regulation and systems biology. 2008;2:GRSB–S702.

36. Özgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24(13):i277–i285.

37. Chavali S, Barrenas F, Kanduri K, Benson M. Network properties of human disease genes with pleiotropic effects. BMC systems biology. 2010;4(1):78.

38. Newman M. Networks: an Introduction. Oxford university press; 2010.

39. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics. 2008;82(4):949–958.

40. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinformatics. 2012;13(1):182.

41. Martínez V, Cano C, Blanco A. ProphNet: A generic prioritization method through propagation of information. BMC Bioinformatics. 2014;15(1):S5.

42. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2017;45(D1):D833–D839.

43. Serrano MÁ, Boguná M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences. 2009;106(16):6483–6488.

44. Darabos C, White MJ, Graham BE, Leung DN, Williams SM, Moore JH. The multiscale backbone of the human phenotype network based on biological pathways. BioData Mining. 2014;7(1):1.

45. Serrano MÁ, Boguná M, Sagués F. Uncovering the hidden geometry behind metabolic networks. Molecular biosystems. 2012;8(3):843–850.

46. Cantini L, Medico E, Fortunato S, Caselle M. Detection of gene communities in multi-networks reveals cancer drivers. Scientific reports. 2015;5:17386.

47. Liu J, Xiong Q, Shi W, Shi X, Wang K. Evaluating the importance of nodes in complex networks. Physica A: Statistical Mechanics and its Applications. 2016;452:209–219.

48. Ren ZM, Shao F, Liu JG, Guo Q, Wang BH. Node importance measurement based on the degree and clustering coefficient information. Acta Phys Sin. 2013;6:128901.

49. Savin L. 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. 1935;19(11):597.

50. Dimmock D, Kobayashi K, Iijima M, Tabata A, Wong LJ, Saheki T, et al. Citrin deficiency: a novel cause of failure to thrive that responds to a high-protein, low-carbohydrate diet. Pediatrics. 2007;119(3):e773–e777.

51. Zimmerman HJ. Drug-induced liver disease. Clinics in Liver Disease. 2000;4(1):73–96.

52. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease–gene associations. Methods. 2015;74:83–89.

53. Liu CC, Tseng YT, Li W, Wu CY, Mayzus I, Rzhetsky A, et al. DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. Nucleic Acids Research. 2014;42(W1):W137–W146.

54. Mansergh FC, Millington-Ward S, Kennan A, Kiang AS, Humphries M, Farrar GJ, et al. Retinitis pigmentosa and progressive sensorineural hearing loss caused by a C12258A mutation in the mitochondrial MTTS2 gene. The American Journal of Human Genetics. 1999;64(4):971–985.

55. Silva DR, Coelho AC, Dumke A, Valentini JD, de Nunes JN, Stefani CL, et al. Osteoporosis prevalence and associated factors in patients with COPD: a cross-sectional study. Respiratory Care. 2011;56(7):961–968.

56. Stolz SE, Chatrian GE, Spence AM. Epileptic nystagmus. Epilepsia. 1991;32(6):910–918.

57. American Optometric Association. https://www.aoa.org/; 2017. Available from: `https://www.aoa.org/patients-and-public/eye-and-vision-problems/glossary-of-eye-and-vision-conditions/nystagmus`.

58. Kenchaiah S, Evans JC, Levy D, Wilson PW, Benjamin EJ, Larson MG, et al. Obesity and the risk of heart failure. New England Journal of Medicine. 2002;347(5):305–313.

59. Rowland LP. Diagnosis of amyotrophic lateral sclerosis. Journal of the Neurological Sciences. 1998;160:S6–S24.

60. Rodger W. Non-insulin-dependent (type II) diabetes mellitus. CMAJ: Canadian Medical Association Journal. 1991;145(12):1571.

61. Millar J. Epilepsy and strabismus. Epilepsia. 1965;6(1):43–46.

62. Czerwinski SL, Plummer CE, Greenberg SM, Craft WF, Conway JA, Perez ML, et al. Dynamic exophthalmos and lateral strabismus in a dog caused by masticatory muscle myositis. Veterinary Ophthalmology. 2015;18(6):515–520.

63. Brookhouser PE. Sensorineural hearing loss in children. Pediatric Clinics of North America. 1996;43(6):1195–1216.

64. Nørgaard F. Earliest roentgenological changes in polyarthritis of the rheumatoid type: rheumatoid arthritis. Radiology. 1965;85(2):325–329.

65. Botez M, Attig E, Vézina JL. Cerebellar atrophy in epileptic patients. CanadianJournal of Neurological Sciences. 1988;15(3):299–303.

66. Weissmann G. Rheumatoid arthritis and systemic lupus erythematosus as immune complex diseases. Bulletin of the NYU Hospital for Joint Diseases. 2009;67(3):251.

67. Sato O, Yamguchi T, Kittaka M, Toyama H. Hydrocephalus and epilepsy. Child's Nervous System. 2001;17(1):76–86.

68. Nabel EG, Braunwald E. A tale of coronary artery disease and myocardial infarction. New England Journal of Medicine. 2012;366(1):54–63.

69. Kaplowitz N. Drug-induced liver injury. Clinical Infectious Diseases. 2004;38(Supplement_2):S44–S48.

70. Galli E, Gianni S, Auricchio G, Brunetti E, Mancino G, Rossi P. Atopic dermatitis and asthma. In: Allergy and Asthma Proceedings. vol. 28. OceanSide Publications, Inc; 2007. p. 540–543.

71. Arumugam K. Endometriosis and obesity. Journal of Obstetrics and Gynaecology. 1992;12(4):266–268.

72. Gajarski R, Naftel DC, Pahl E, Alejos J, Pearce FB, Kirklin JK, et al. Outcomes of pediatric patients with hypertrophic cardiomyopathy listed for transplant. The Journal of Heart and Lung Transplantation. 2009;28(12):1329–1334.

73. Tucci DL, Born DE, Rubel EW. Changes in spontaneous activity and CNS morphology associated with conductive and sensorineural hearing loss in chickens. Annals of Otology, Rhinology & Laryngology. 1987;96(3):343–350.

74. Fiorentino E, Pantuso G, Cusimano A, Latteri S, Mastrosimone A, Cipolla C. Gastro-oesophageal reflux and "epileptic" attacks: casually associated or related efficiency of antireflux surgery. Chirurgia Italiana. 2008;58(6):689–696.

75. Kurt A, Nijboer F, Matuz T, Kübler A. Depression and anxiety in individuals with amyotrophic lateral sclerosis. CNS Drugs. 2007;21(4):279–291.