Uncovering community structures of disease classes within the human disease network

Mathew Golf
mathewggolf@lewisu.edu
DATA-51000-001, Summer 1
Data Mining and Analytics
Lewis University

I. Introduction

Interestingly, some genes have been found to be connected to multiple different disease phenotypes and they have therefore been suggested as therapeutic targets[1]. This is a logical deduction from the data, and one in which we agree with. However, no matter the interactions and degree of which uncovered, the human system is a dynamic complex system with many moving parts, therefore it is likely that the disease phenotype is caused by a combination of triggers like genetic, epigenetic, environmental, etc. Finding communities within the human disease network will therefore enable the identification of associative disease phenotypes as well as genes involved. Before communities are found it is important that we identify the global qualities of the network and the most important nodes to compare with the communities found within the network.

The human disease network was created by researchers through the use of the Morbid Map accessed through the Online Mendelian Inheritance in Man (OMIM)[1]. The OMIM consists of human disease genes and phenotypes for thousands of disorders and genes[1]. Therefore, the human disease network was created by further classifying the data from the OMIM into 22 disorder classes based on the physiological system affected[1]. Nodes represent disorders and disorders are connected if they share at least one gene where mutations are involved in both disorders[1].

In section II, we provide an overview of the data describing the features, their interactions, and the underlying hierarchy of connections which will be further investigated. Then in section III, the methodology for investigation of the network through the use of network statistics is explained. Section IV, describes the results from the aforementioned analysis and discusses the relevance of said results. Lastly, in section V conclusions from the results and the discussion of the analysis is provided indicating future directions for further investigation.

II. Data Description

For this analysis one dataset was used consisting of IDs for human diseases and genes as well as their types and overall disease classes. Analysis was conducted on 516 diseases and 903 genes for the human disease network (Table I). Directed interactions between nodes are also indicated in the dataset, as well the size of the node indicates the number of connections, and therefore its importance. Degree distribution as visualized in Figure 1 indicates that there is an inherent hierarchy of interactions sufficient to be investigated using modularity, betweenness, PageRank, and HITS to find important nodes and communities within the data. As indicated in Figure 1, by the size and color (greenness) of the nodes, the central nodes are related to diseases and not genes.

TABLE I. Human Disease Network Disease and Gene Values

Attribute	Type	Example Value	Description
ID	Nominal (primary key)	55	Label identifier
LABEL	Nominal (string)	“Deafness”	Name of the disease
TYPE	Nominal (string)	“disease”	Indicates disease or gene
DISCLASS	Nominal (string)	“Ear, Nose, Throat”	Indicates the class of the disease

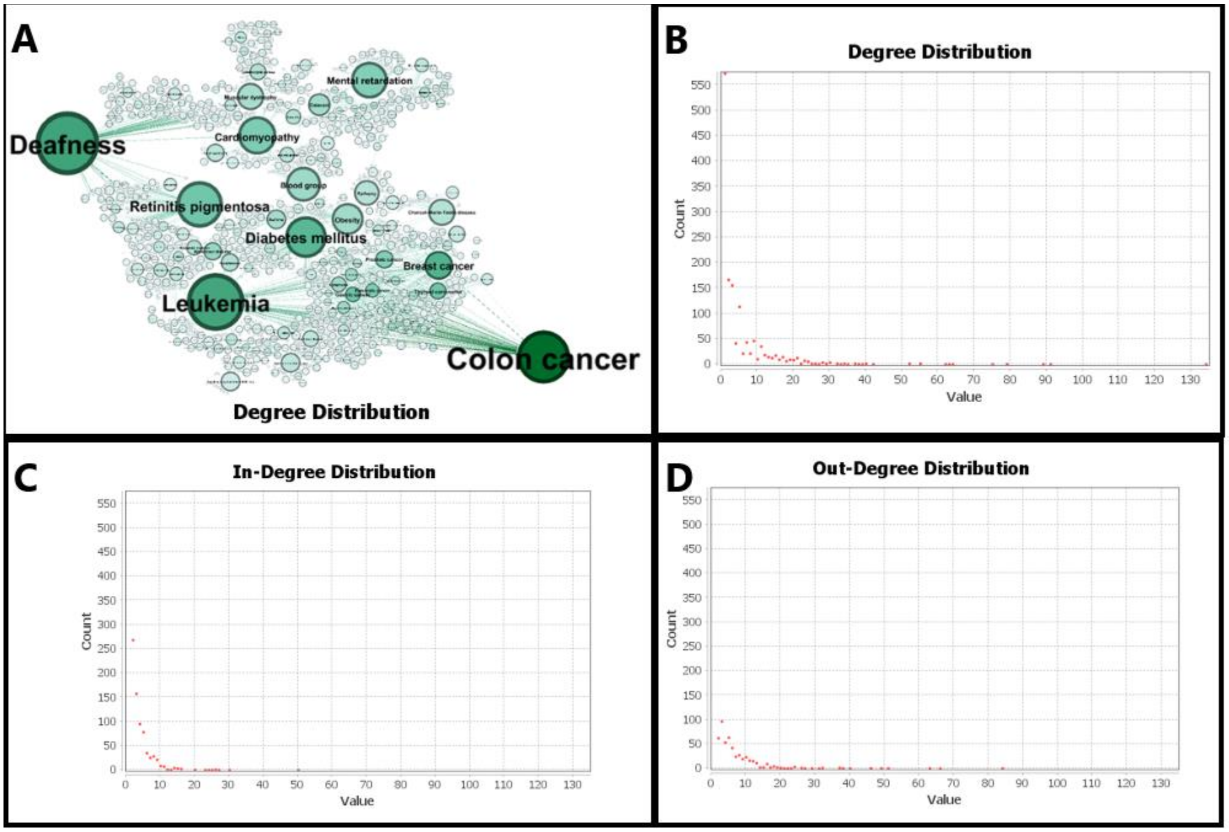

Fig. 1. Human disease network visualized through the distribution of degrees, average degree: 2.767. Green is the greatest, white is the least.

III. Methodology

Data for the human disease network was obtained through the Gephi wiki and analyzed using Gephi[2]. Initially, the distribution of degrees within the network was visualized to ensure that there is an underlying network structure sufficient for further statistical investigation (Fig. 1). Likewise, the size of the nodes were scaled relative to the importance of the nodes, the distribution of degrees confirms this as the larger nodes have more connections, indicated by the greenness (Fig. 1). We then used the betweenness centrality algorithm selecting for a directed network and normalizing the centralities [0, 1][3]. We then investigated the relative hub and authority scores as generated by the HITS algorithm using an epsilon value of 1.0E-4[4]. Next, we used the PageRank algorithm for a directed network with a probability value of 0.85 and an epsilon value of 0.001[5]. Lastly, we investigated the modularity of the human disease network data to detect communities setting the parameters as randomized, using the edge weights, and setting the resolution at 1.0[6, 7]. After statistical analysis we then selected the largest nodes thus highlighting the communities connected to these most important nodes.

IV. Results and Discussion

Since we are interested in identifying the underlying communities of disease phenotypes, we must first show statistical evidence of such communities. Figure 1 suggests there is an underlying hierarchy to the network as some nodes are much more connected than other nodes. Thus, we next investigated the betweenness centrality of the human disease network to find the average graph-distance between all pairs of nodes (Fig. 2). In doing so, we found the average path length to be around 6.65 and the diameter of the network to be 15 (Fig. 2). Betweenness centrality indicates the frequency a node appears on the shortest paths between nodes in the network, clearly most nodes appear on a few and a few nodes appear on most (Fig. 2B). Eccentricity describes the distance from a given starting node to the farthest node in the network, thus having a value around 900 for a network of 1419 nodes indicates a well-connected network (Fig. 2C). Likewise, the closeness centrality measure indicates the average distance from a given starting node to all other nodes in the network and here it indicates that on average the path length is small with some very well-connected path lengths (Fig. 2D).

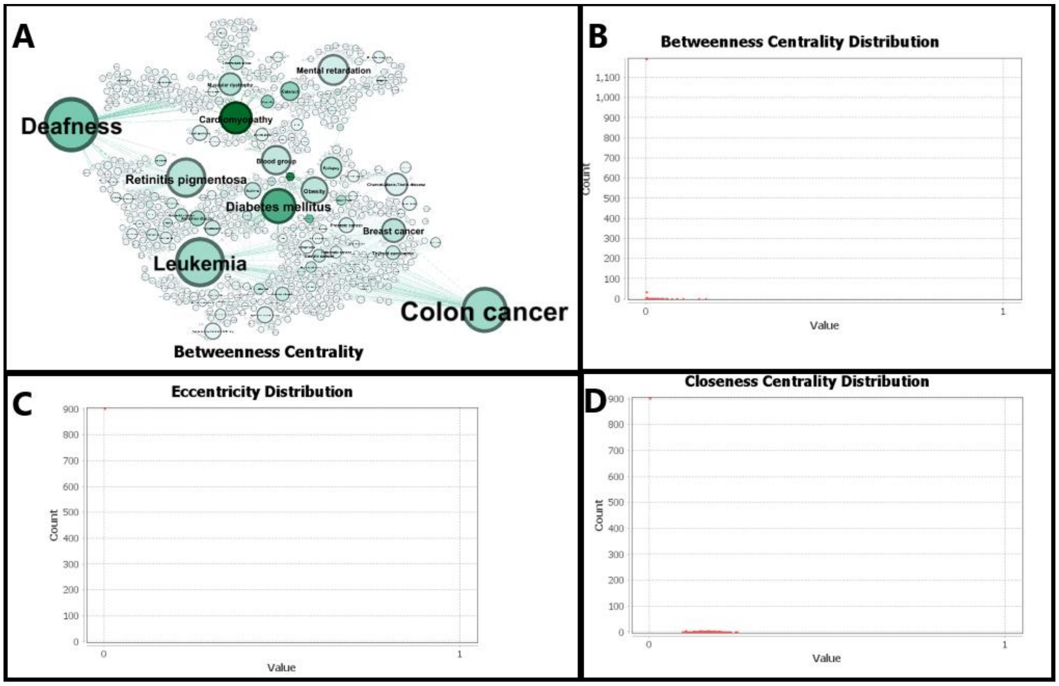

Fig. 2. Betweenness centrality for the directed human disease network, where the distance is the average graph-distance between all pairs of nodes (= 6.65). Connected nodes have a graph distance of 1. The diameter is the longest graph distance between any two nodes in the network and was 15. Centralities were normalized [0, 1]. Green is the greatest, white is the least.

Next, we wanted to use the HITS algorithm to find the relative hubs and authority scores of the nodes (Fig. 3). As our network is associated with disease classes we wanted to observe if the important nodes were more likely to be authorities rather than hubs (Fig. 3A). As predicted the distribution of authorities was increased relative to the distribution of hubs (Fig. 3B & C). What is interesting in this data is that not all large nodes are authorities or hubs, and there are only a few nodes that seem to have both the most authority and the most hubness, i.e. colon cancer (Fig. 3A). This makes sense as biological data is often interconnected and cyclic, so having a true authority or a true hub does not make sense in the messy reality of the biological system.

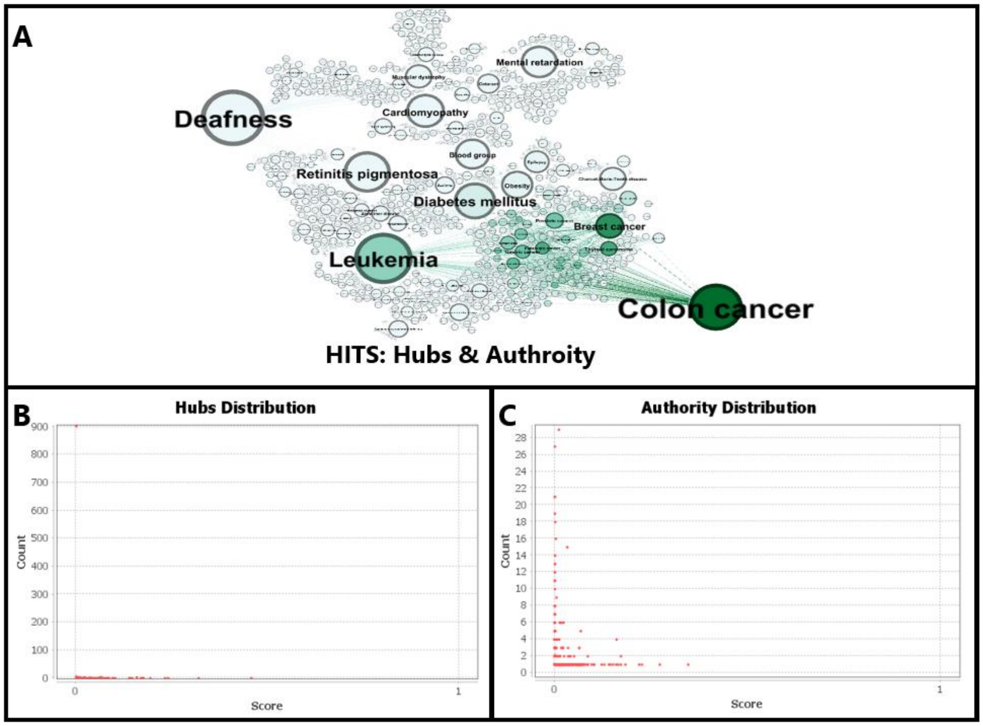

Fig. 3. Authority and hub scores generated from the HITS algorithm (with epsilon of E=1.0E-4). Green is the greatest, white is the least.

Since the HITS algorithm did not generate any useful information regarding the underlying community structure of the human disease network besides highlighting a few central nodes like colon cancer, we next investigated the data using the PageRank algorithm (Fig. 4). As evident from figure 4B, there is a clear distribution in the likelihood of ending at specific nodes, and from figure 4A it is clear that the largest nodes have the highest likelihood. Thus, it is probable that the greenest nodes (largest) represent their own communities (Fig. 4A). For example, the colon cancer node and all its immediate connections would be the colon cancer community.

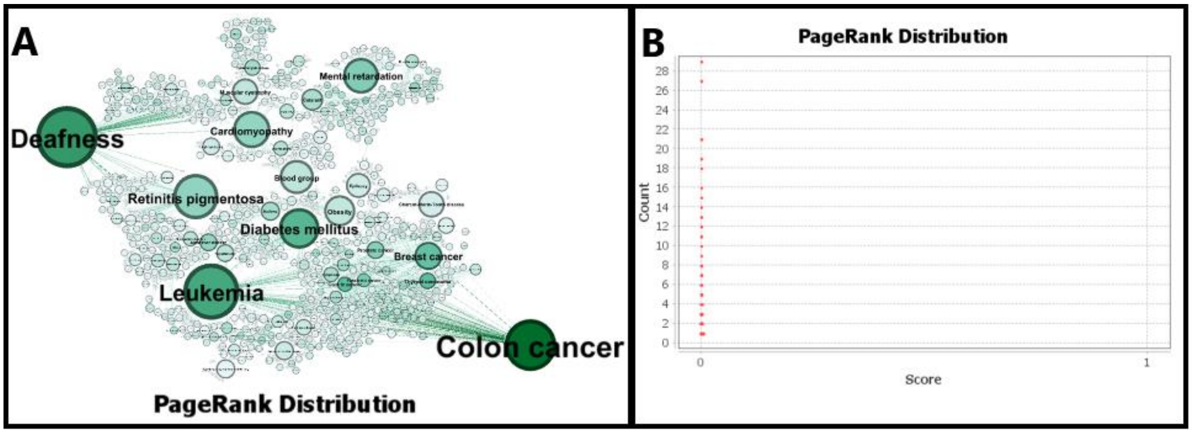

Fig. 4. PageRank distribution of the directed human disease network, nodes are ranked according to how often a user following links will non-randomly reach the node. Probability value of p=0.85 was used to stimulate the user randomly restarting the web-surfing, and an epsilon value of 0.001 was selected as the stopping criterion. Green is the greatest, white is the least.

Therefore, to further investigate the community structure of the human disease network we found the modularity of the data which represents the community structure of the data (Fig. 5). As apparent in figure 5B the overall size of each community varied from small (10) to large (150) for a total of 29 communities when using a resolution value of 1.0. Thus, we found a modularity value of 0.87 indicating significant community structure and trending toward a complete network. Selecting a different resolution value would change our results as below 1.0 would generate smaller communities, and above 1.0 would generate larger communities, however a resolution value of 1.0 opts for the sweet spot in size of communities.

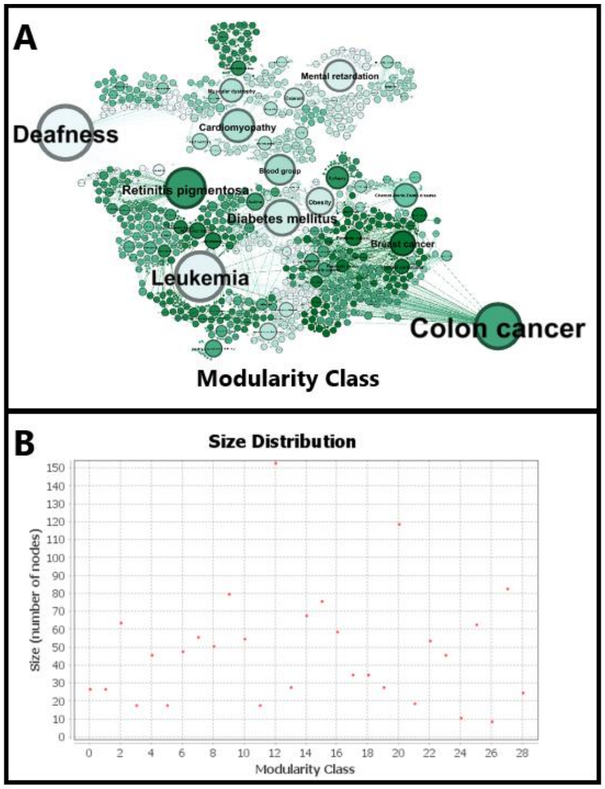

Fig. 5. Community detection for the human disease network using the modularity class method with weights and a resolution value of 1.0. A modularity value of 0.87 and 29 total communities were found. Green is the greatest, white is the least.

Lastly, we then selected some of the largest nodes – which indicate the most connections and therefore importance, in order to visualize the communities associated with these nodes (Fig. 6). Since we found 29 communities using the modularity class method with a resolution value of 1.0, we wanted to select some of the more important ones to view their connections (Fig. 5 & 6). Some of these communities are the: (B) blood group, (C) mental retardation, (D) cardiomyopathy, (E) diabetes mellitus, (F) retinitis pigmentosa, (G) colon cancer, (H) leukemia, and (I) deafness (Fig. 6). Interestingly, connections which seem logical were not observed, for example diabetes is not connected with colon cancer, blood group is not connected with leukemia, and obesity is not connected with more diseases (Fig. 6). However, there are clear connections between colon cancer and leukemia and vice versa (Fig. 6G & H), as well as the interesting connections between deafness and cardiomyopathy (Fig. 6D & I). Furthermore, the graph density of Figure 6A was found to be 0.002 and the average clustering coefficient value was found to be 0.414.

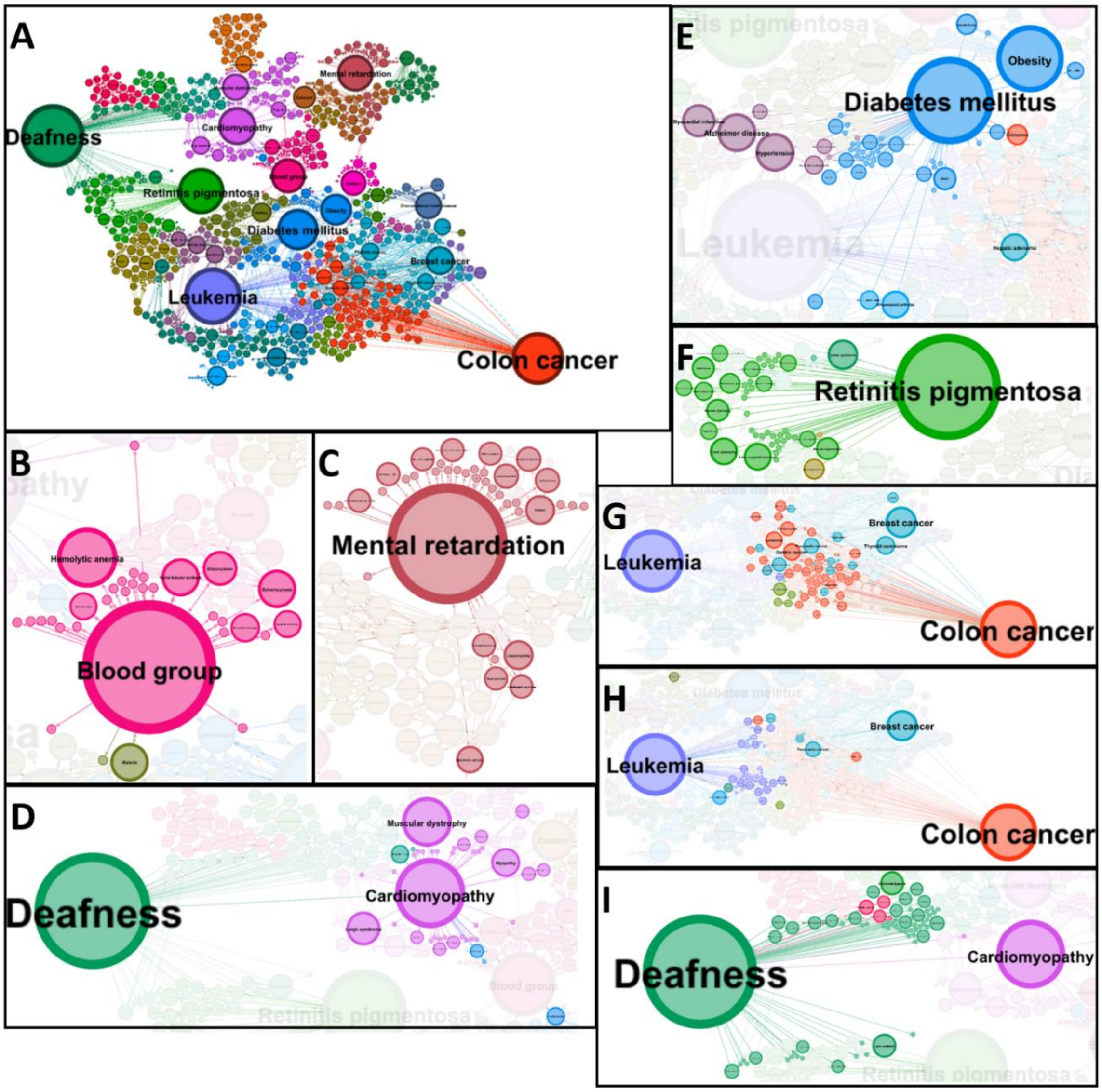

Fig. 6. (A) The human disease network, colors indicate shared disease class [label adjust: speed 1.0, include node size]; (B) Blood group community; (C) Mental retardation community; (D) Cardiomyopathy community; (E) Diabetes mellitus community; (F) Retinitis pigmentosa community; (G) Colon cancer community; (H) Leukemia community; (I) Deafness community.

V. Conclusions

Finding the underlying community structure of diseases is important in order to understand which genes give rise to these communities, why are some diseases connected throughout many communities, and why are other diseases not connected. Thus, to find these communities, we first wanted to ensure that an underlying network structure was present, therefore degree distribution was visualized (Fig. 1). This suggested that there was an underlying hierarchy as some nodes had significantly more connections than other nodes. To further investigate this structure, we found the betweenness centrality, the hub and authority scores, and the PageRank distribution of the human disease network (Fig. 2, 3 & 4). Considering the networks and values generated from these methods it was clear that the underlying structure of the network consists of communities, therefore we next found the modularity score of the network (Fig. 5). Lastly, we then visualized the most important communities and their connections in order to understand the connections between diseases and the lack of connections between otherwise similar diseases (Fig. 6).

Future directions would be to take the communities identified and query the genes within finding the most important (connected) nodes and using them as potential therapeutic targets for intervention on the related disease. Personally, I believe the absence of connection is almost as important as the presence of a connection. For example, diabetes is not related to colon cancer in the human disease network despite there being clear connections in the real-world. Therefore, this absence of connection in the disease-gene network indicates that the relationship must be more environmental and therefore an environmental or nutritional intervention might yield more results than a genetic intervention. Thus, future directions should focus on the missed connections as well as the present connections, in order to provide a more comprehensive intervention plan outside of just genetic intervention.

References

[1] K.-I. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Szló Barabá, “The human disease network,.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1885563/pdf/zpq8685.pdf.

[2] “gephi/gephi,” GitHub. https://github.com/gephi/gephi/wiki/Datasets.

[3] U. Brandes, “A faster algorithm for betweenness centrality*,” The Journal of Mathematical Sociology, vol. 25, no. 2, pp. 163–177, Jun. 2001, doi: 10.1080/0022250x.2001.9990249.

[4] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM, vol. 46, no. 5, pp. 604–632, Sep. 1999, doi: 10.1145/324133.324140.

[5] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Computer Networks and ISDN Systems, vol. 30, no. 1–7, pp. 107–117, Apr. 1998, doi: 10.1016/s0169-7552(98)00110-x.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/p10008.

[7] R. Lambiotte, J.-C. . Delvenne, and M. Barahona, “Laplacian Dynamics and Multiscale Modular Structure in Networks,” IEEE Transactions on Network Science and Engineering, vol. 1, no. 2, pp. 76–90, Jul. 2014, doi: 10.1109/TNSE.2015.2391998.