Analysis of a human Disase-Gene-Network reveals the Complexity of Cancer

Christina Morgenstern
christinamorgenstern@lewisu.edu
DATA-51000-002, 20
Data Mining and Analytics
Lewis University

I. Introduction

Biological systems are complex structures that result through the orchestrated interaction of molecules, genes, pathways or organs at different scales within an organism. The recent rise in technologies in the biomedical field have generated a tremendous amount of data at the molecular level. Studying these datasets can reveal insights into the process of development and disease. Network biology is a field addressing the interplay between individual entities and trying to decipher the patterns underlying normal and disease mechanisms [1].

The question we are addressing in this paper is how human diseases and corresponding genes are related to each other. The aim is to find disease patterns such as detecting the intricate structure of genes involved in a certain disease.

This paper describes the process of developing and analyzing a disease-gene network model using visualization techniques and graph metrics. In the data description (section II), an overview of the features and their distribution is given. Section III describes the methodology of developing a graph model using both the network visualization tool Gephi and the Python library NetworkX. The results from the modeling and the graph analysis are presented and discussed in section IV. A summary is given, and conclusions are drawn in section V.

II. Data Description

The dataset used in this work was downloaded from the Stanford Network Analysis Project (SNAP), a general purpose network analysis and graph mining library hosted by Stanford University [2]. This website also offers curated network datasets from the biomedical world. The disease-gene-association network named "DG-AssocMiner" was downloaded as Tab-separated values (.tsv) file and used for subsequent graph analysis using Gephi [3] and Python´s NetworkX [4] library. The data contain information on human diseases and the associated genes [5]. Table 1 lists the attributes with the associated data types and provides a description with an example value. In total, there were 21,357 rows and 3 columns in the dataset.

Table 1. Description of DG-AssocMiner Dataset

Attribute	Type	Example Value	Description
# DISEASE ID	Categorical (string)	C0036095	Identifier for disease name
DISEASE NAME	Categorical (string)	Salivary Gland Neoplasms	Name of the disease
GENE ID	Numeric (integer)	1462	Identifier of gene

For exploratory data analysis (EDA), the data were loaded into the Jupyter notebook [6] environment and analyzed using Python 3 programming software [7]. The following libraries were used: NumPy [8], pandas [9], Matplotlib [10] and seaborn [11]. Initially the dataset was explored in terms of the distribution of the disease names and genes as well as for missing values. However, no missing values were detected. Assessing the frequency of different diseases in the data showed that prostatic neoplasms had the most entries with 485, followed by IgA glomerulonephritis and mammary neoplasms with 450 and 433 entries, respectively (see Fig. 1a. The diseases with the fewest entries are amongst others diarrhea, arthritis and hay fever, with each 10 occurrences. The gene IDs, represented by numbers, with the most nodes are 7124 and 6648 with 155 and 96 entries, respectively (Fig. 1b). Several genes listed have only a single occurrence in the dataset.

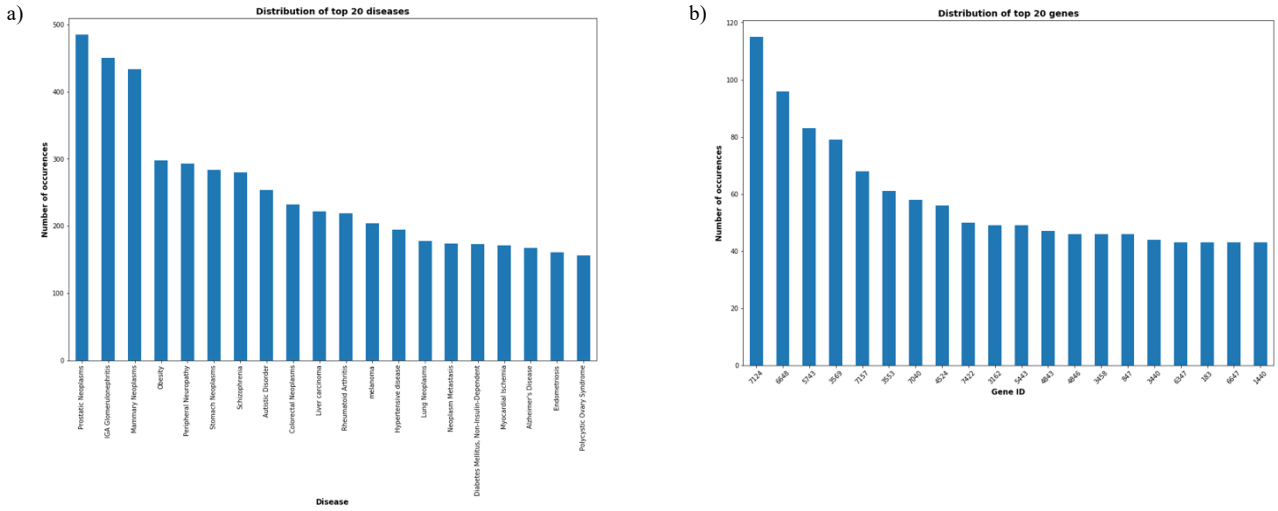

Fig. 1. Distribution of top 20 diseases (a) and top 20 genes (b) in disease-gene-network dataset.

While there is no description what gene ID corresponds to which gene, a list is given whether the gene is essential or not essential in human metabolism where essential means that it is important for normal function. According to the list given, the two most prominent genes in the dataset 7124 and 6648 are non-essential and essential, respectively. Researching for the for gene ID 7124 in the database National Center for Biotechnology Information yields the human Tumor Necrosis Factor (TNF), a proinflammatory cytokine involved in the regulation of a variety of biological processes such as cell proliferation as well as cancer [12]. Gene ID 6648 encodes the human Superoxide Dismutase 2 (SOD2), a protein prevalent in the mitochondria that has been implicated in a number of diseases like cancer [13].

In summary, the rows of the dataset are identifiers of diseases and the involved genes. In the next section the methodology of developing a graph model based on this edge list as well as the analysis of the network is described.

III. METHODOLOGY

In order to create a network graph based on the edge list of the disease-gene data, the NetworkX package was used with Python programming language. Creating a graph based on the edge list using NetworkX is straight-forward, the edge list just needs to be passed to NetworkX graph object. Table 2 describes the graph properties that have been calculated using NetworkX as well as confirmed from the data description listed on the SNAP website. In total, there are 7813 nodes in the graph, that can be of either type disease (519) or type gene (7294). The nodes are connected via 21,357 edges. Since the nodes are divided into two categories, disease and gene, it was investigated whether the graph is of bipartite nature. A bipartite graph is graph, whose nodes can be separated into two independent and disjoint sets. This was confirmed as true using the NetworkX command [14]. Investigating the left and right nodes, showed that the left nodes were of disease type where the right nodes were of gene type.

The size of the network in terms of number of nodes and edges is described in Table 2.

TABLE 2. DESCRIPTION OF DG-ASSOCMINER NETWORK

Attribute	Total number	Description
NODES	7813	Nodes can be of type disease or gene
DISEASE NODES	519	Nodes of type disease
GENE NODES	7294	Nodes of type gene
EDGES	21357	Connections of type disease-gene, gene-gene, disease-disease

Using NetworkX, the disease-gene graph was analyzed in terms of microscale, medium-scale and large-scale structures. Microscale structures are measurements that characterize the network's nodes such as centrality which defines a measure of importance. Degree centrality, betweenness, eigenvector and closeness centrality are metrics that characterize a node in relation to other nodes in the network [15]. Degree centrality processes the number of edges for a certain node. Betweenness centrality assesses the number of shortest paths that pass-through a given node. Closeness centrality determines the average of the shortest paths between a certain node and all other nodes. Furthermore, applying the PageRank algorithm sorts nodes according to their importance [16]. For each metric does apply, the higher the values, the more important a node is.

Large scale structures were determined through finding the distances and shortest paths between nodes. Of special interest are the medium-scale structures of the network like communities and cliques which help to identify subnetworks.

Since visualizations of the network did not look appealing using NetworkX, the graph was visualized using Gephi. Before importing into Gephi, the columns had to be renamed to source and target in order to read the data in the form of an edge list. For layouting of the entire network, the Force Atlas 2 [17] layout algorithm was chosen using a scaling factor of 100 to create a more sparse network. This force-directed algorithm is an improved version of the Force Atlas but optimized for large networks.

In Fig. 2, the process of analyzing the disease-gene network is depicted as a flowchart.

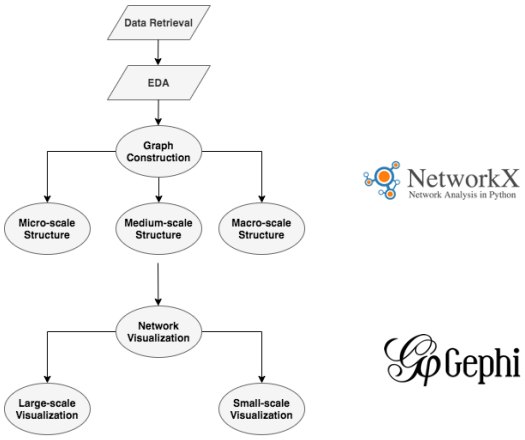

Fig. 2. Flowchart describing the process of network analysis and visualization on the disease-gene dataset.

IV. Results and Discussion

In this section, the results from the network analysis and visualizations are presented and discussed. Results from the analysis of metrics based on micro, macro and medium scale structures are discussed in sections A, B and C, respectively. The visualizations of the disease-gene network at large-scale and small-scale round up the results in sections D and E.

A. Micro-scale Structure Analysis

Different network centrality measures were calculated, and the results ranked from high to low values. The top three nodes of each metric determining degree, betweenness, eigenvector and closeness centrality and their respective normalized values are shown in Table 3 (values were rounded to four decimal places).

TABLE 3. DEGREE CENTRALITY MEASURES

Centrality Measure	Node	Value
DEGREE CENTRALITY	Prostatic Neoplasms	0.0620
	IGA Glomerulonephritis	0.0576
	Mammary Neoplasms	0.0554
BETWEENNESS CENTRALITY	Prostatic Neoplasms	0.0845
	IGA Glomerulonephritis	0.0825
	Mammary Neoplasms	0.0711

Centrality Measure	Node	Value
EIGENVECTOR CENTRALITY	Mammary Neoplasms	0.2379
	Prostatic Neoplasms	0.2369
	Stomach Neoplasms	0.1445
CLOSENESS CENTRALITY	7124	0.3408
	Mammary Neoplasms	0.3394
	Prostatic Neoplasms	0.3372

The centrality analysis using different measures showed two diseases as important nodes or hubs within the graph. Both, prostatic and mammary neoplasms show up in the top three ranking of each centrality measure (see Table 3). Thus, these are well connected nodes with many short paths between other nodes. Listing the neighbor nodes of prostatic and mammary neoplasms showed that the former has with 485 slightly more neighbors than the latter (433).

Table 4 shows the top three listed nodes according to the PageRank algorithm. IgA glomerulonephritis, an autoimmune disease that affects the kidneys ranks top when applying this algorithm [18].

TABLE 4. RESUTLT FOR PAGERANK ALGORITHM

Node	Value
IGA Glomerulonephritis	0.0133
Prostatic Neoplasms	0.0109
Mammary Neoplasms	0.0089

Fig. 3 shows a plot of the fraction of nodes against the degrees. More than 40% of the nodes in the disease-gene network have 0 degrees i.e. no connections to other nodes. Only a few nodes have more than 400 degrees. And these are the top three nodes from the centrality network analysis listed before (Table 3): prostatic neoplasms, mammary neoplasms and IgA glomerulonephritis with 485, 433 and 450 nodes each. Most of the nodes have fewer than 200 connections (see Fig. 3).

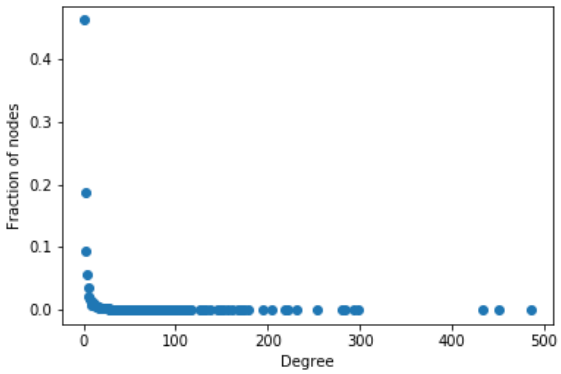

Fig. 3. Degree distribution of nodes in disease-gene network.

B. *Macro-scale Structure Analysis*

Interesting macro-scale structures that characterize the disease-gene network are the path lengths of the graph and the average path length as well as the shortest paths. In general path lengths are a sequence of nodes connected via edges, with the number of edges corresponding to the path length. The diameter of a network is given by the largest path length and was calculated to be 8 for the disease-gene network. The mean shortest path length was found to be 4.23. See Fig. 4 for the distribution of shortest paths in the disease-gene network.

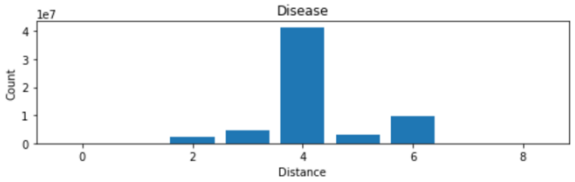

Fig. 4. Histogram of shortest path lengths in disease-gene network.

C. *Medium-scale Structure Analysis*

Many insights of the underlying network can be gained through the study of medium-scale structures such as communities and cliques. Those structures define subareas of the graph that can be analyzed for further insights. Using the Clauset-Newman-Moore method [19], 25 communities were detected in the disease-gene-network. The Girvan-Newman method [20] of community detection based on the betweenness centrality measure only finds 2 communities. Since the latter algorithm works by removing the edges with the highest betweenness values, it is reasonable that only two communities remain. On the other hand, the Clauset-Newman-Moore algorithm finds communities through joining nodes by maximizing modularity. Thus, these complementary approaches determine the minimum and maximum number of communities, for the Girvan-Newman and Clauset-Newman-Moore methods, respectively.

It could also be shown, that the disease-gene network is of bipartite graph type where nodes can be divided into two separated sets. As anticipated, this separation has been shown to be between disease nodes and gene nodes. Among the left nodes the diseases are listed whereas the right nodes are exclusively genes.

D. *Large-scale Visualization of Disease-Gene Network*

Since NetworkX does not produce great visualizations of networks, Gephi was applied to visualize the entire disease-gene graph as well as to find distinct modules, a type of community. The network data was layouted using the Force Atlas 2 algorithm, as described in section III. Subsequently, the modularity algorithm was run to find distinct communities [21]. Gephi´s modularity algorithm produced 20 modules of varying sizes (see Fig. 5). Those classes were further highlighted using different colors in the previously layouted graph (Fig. 7).

Modules 9 (Fig. 7, purple) and 12 (Fig. 7, green) are the biggest with more than 800 nodes each, followed by modules 7 (Fig. 7, blue) and 2 (Fig. 7, dark grey) with approx. 700 nodes and module 4 (Fig. 7, orange) with more than 500 nodes.

See Fig. 7 for visualization of the entire disease-gene network layouted with the Force Atlas 2 algorithm and visually partitioned into the 20 modules found.

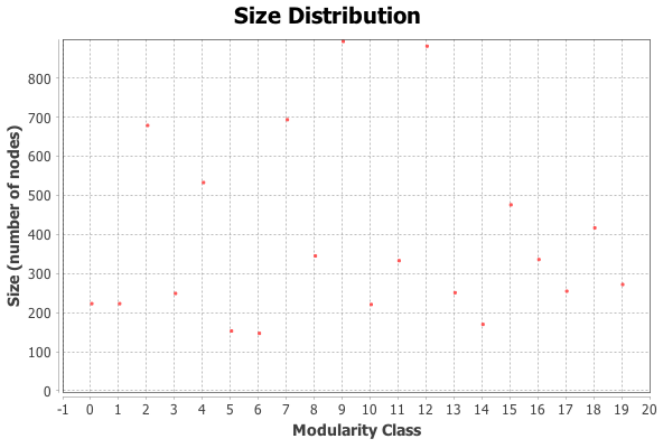

Fig. 5. Result of modularity analysis on disease-gene network using Gephi.

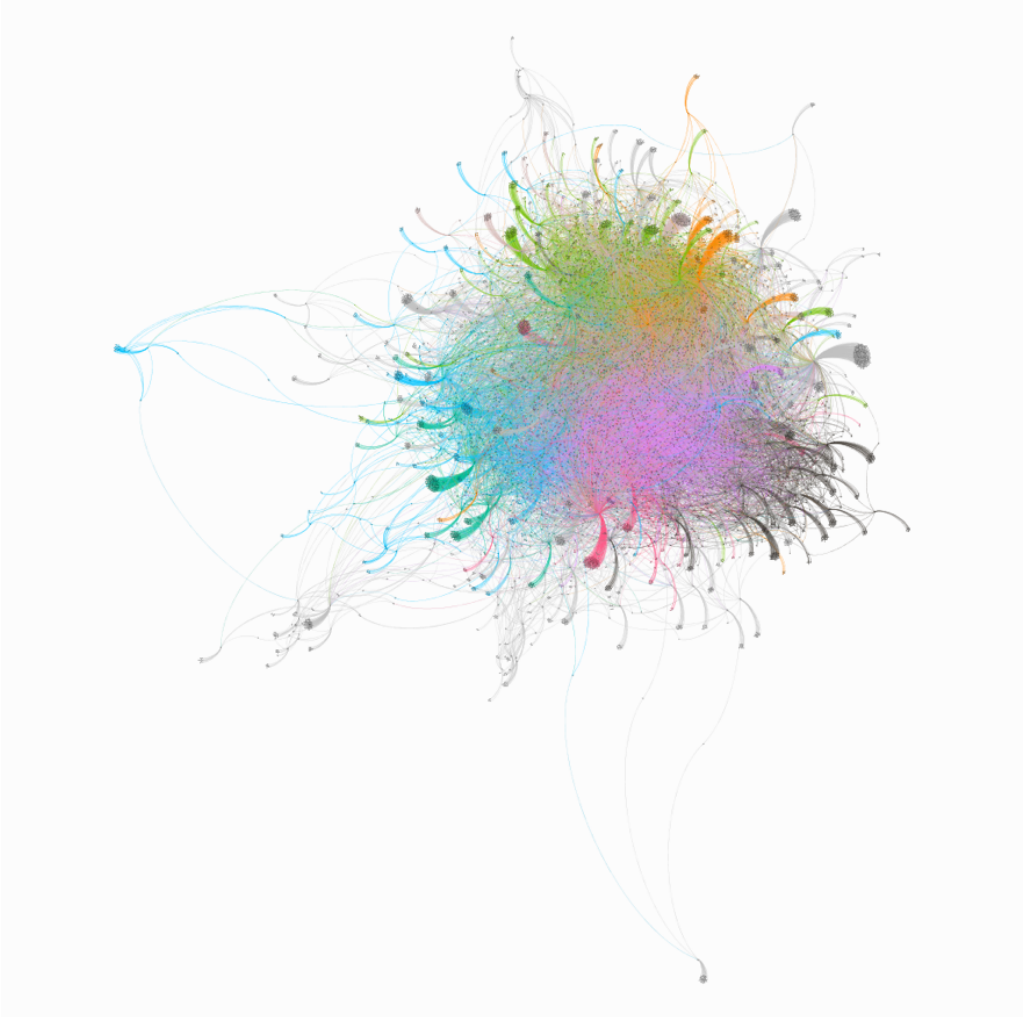

Fig. 6. Visualization of full scale disease-gene network and labeled according to the 20 modularity classes.

E. Small-scale Visualization of selected Disease-Gene Subnetwork

Using the modularity class algorithm of Gephi the graph was partitioned into 20 modules with modules 9 and 12 being the biggest. Examining the individual modules showed that the previously highly ranked node mammary neoplasms belongs to node 12 whereas the node prostatic neoplasms was found in module 8. The focus of subsequent investigations were modules 12 and 8 because these are the biggest modules in the network and the disease classes mammary neoplasms and prostatic neoplasms which have been deemed important by previous analyses are part of these two modules.

Examining module 12 showed that many other cancers are also part of module 12, such as nasopharyngeal carcinoma, lymphoma, non-small cell lung carcinoma, colon carcinoma and melanoma (Fig. 7). These results show that similar disease types such as cancers are clustered together because they have genes in common that are dysfunctional and thus lead to the development of cancer. Also, as cancer spreads in the human body, one type of cancer can lead to spread of the cancer to different tissues, suggesting links between individual cancer types or disease-disease interactions.

Interestingly, prostatic neoplasms cluster in the same module with cancers of the female and male urogenital tract such as ovarian neoplasm and breast cancer suggesting that in this case the location of the cancer as well as the genes expressed within these tissues are the characterizing features (Fig. 8).

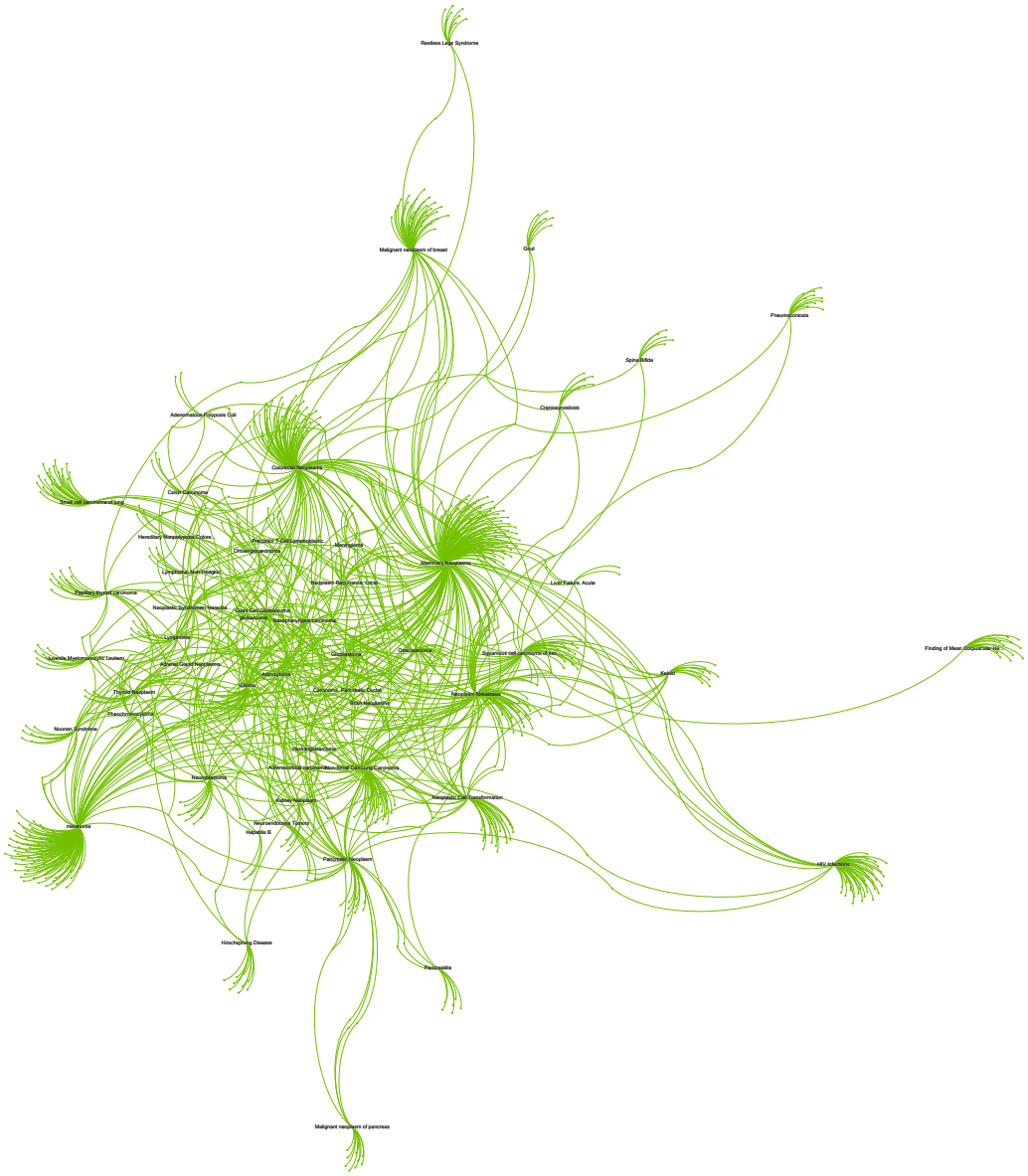

Fig. 7. Visualization of modularity class 12 of disease-gene subnetwork.

V. CONCLUSIONS

Biological systems are complex interwoven networks of individual entities such as genes. Gene-gene communication networks play an essential role in guiding the development of an organism or when malfunctioning driving disease. In this work, the interactions between human diseases and the involved genes were dissected in order to find underlying patterns in the mechanisms of diseases. To achieve this, graph analysis using two different software packages, NetworkX and Gephi, was applied to study micro, medium and macro scale structures of the network as well as to visualize the entire network and selected subnetworks, respectively.

The 7813 nodes in the disease-gene network are divided into 7294 gene and 519 disease nodes with the disease nodes serving as hubs with many connections to disease-relevant genes. Especially cancers have a high connectivity and rank in the top ranges of importance within this network when assessing centrality measures or PageRank. The two neoplastic transformations that stand out as being major nodes are the prostatic neoplasms and the mammary neoplasms nodes with associated gene connections. The biggest communities were found among different types of cancers. While each cancer has its own set of genes that are connected, they also share certain genes. Those genes seem to be implicated in many different cancer types. What can also be seen from this network analysis is that cancer are complex diseases with many genes implicated in the formation of tumors. Other diseases have fewer disease implications suggesting that fewer genes are involved.

While this analysis did not investigate in the underlying genes and their functions, future work should focus on identifying the genes in the cancer-gene subnetworks in order to get a bigger picture on the involved genes.

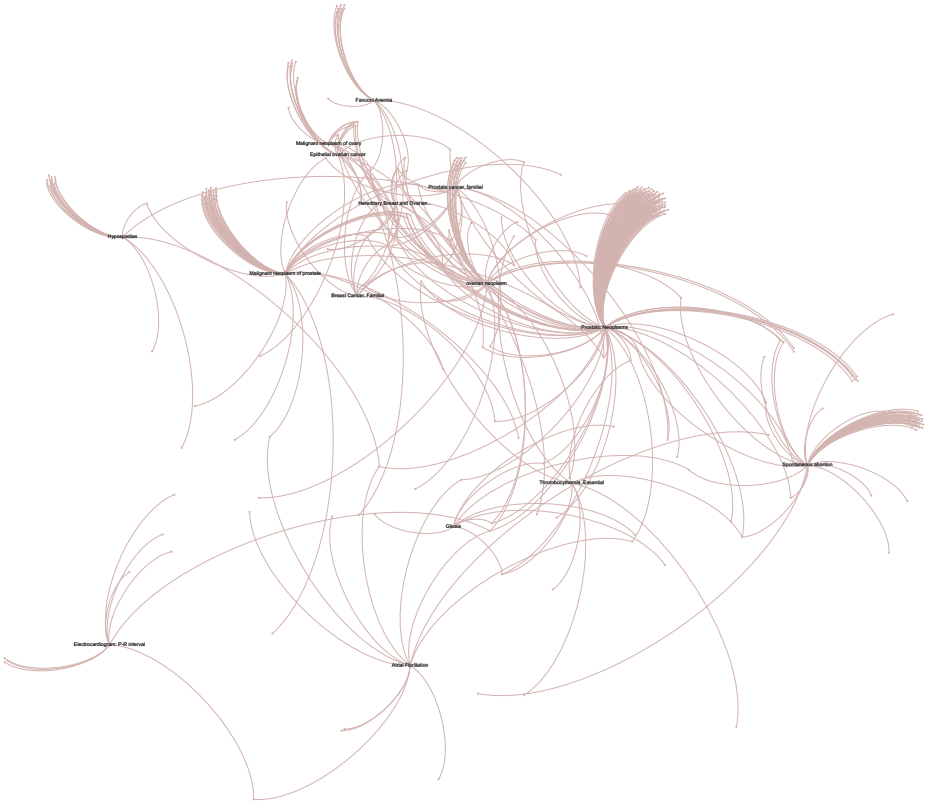

Fig. 8. Visualization of modularity class 8 of the disease-gene subnetwork.

REFERENCES

[1] G. A. Pavlopoulos *et al.*, "Using graph theory to analyze biological networks," *BioData Min.*, vol. 4, p. 10, Apr. 2011, doi: 10.1186/1756-0381-4-10.

[2] "SNAP: Stanford Network Analysis Project." https://snap.stanford.edu/index.html (accessed Jun. 25, 2020).

[3] M. Bastian, S. Heymann, and M. Jacomy, "Gephi : An Open Source Software for Exploring and Manipulating Networks," p. 2.

[4] A. Hagberg, P. Swart, and D. Chult, "Exploring Network Structure, Dynamics, and Function Using NetworkX," presented at the Proceedings of the 7th Python in Science Conference, Jan. 2008.

[5] "BioSNAP: Network datasets: Disease-gene association network." https://snap.stanford.edu/biodata/datasets/10012/10012-DG-AssocMiner.html (accessed Jun. 25, 2020).

[6] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 21–29, 2007, doi: 10.1109/MCSE.2007.53.

[7] Pilgrim, M., & Willison, S. (, "Dive Into Python 3," vol. Springer, .

[8] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi: 10.1109/MCSE.2011.37.

[9] W. McKinney, "Data Structures for Statistical Computing in Python," Austin, Texas, 2010, pp. 56–61, doi: 10.25080/Majora-92bf1922-00a.

[10] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, May 2007, doi: 10.1109/MCSE.2007.55.

[11] Michael Waskom *et al.*, *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo, 2017.

[12] "TNF tumor necrosis factor [Homo sapiens (human)] - Gene - NCBI." https://www.ncbi.nlm.nih.gov/gene/7124 (accessed Jun. 26, 2020).

[13] "SOD2 superoxide dismutase 2 [Homo sapiens (human)] - Gene - NCBI." https://www.ncbi.nlm.nih.gov/gene/6648 (accessed Jun. 26, 2020).

[14] "Bipartite — NetworkX 2.5rc1.dev20200626122636 documentation." https://networkx.github.io/documentation/latest/reference/algorithms/bipartite.html?highlight=bipartite (accessed Jun. 28, 2020).

[15] "Centrality — NetworkX 2.5rc1.dev20200626122636 documentation." https://networkx.github.io/documentation/latest/reference/algorithms/centrality.html?highlight=degree%20centrality (accessed Jun. 28, 2020).

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. 1998.

[17] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLOS ONE*, vol. 9, no. 6, p. e98679, Jun. 2014, doi: 10.1371/journal.pone.0098679.

[18] "IgA nephropathy - Wikipedia." https://en.wikipedia.org/wiki/IgA_nephropathy (accessed Jun. 27, 2020).

[19] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, Dec. 2004, doi: 10.1103/PhysRevE.70.066111.

[20] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Natl. Acad. Sci.*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002, doi: 10.1073/pnas.122653799.

[21] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.