Application of Link and Attribute Prediction on OMIM Disease-Gene Data Set

Skylar Lysbeth Martin
Undergraduate Student
Computer Science
University of Colorado Boulder
Skylar.Martin@colorado.edu

Kiersten Rose Johnson
Undergraduate Student
Chemical and Biological Engineering
University of Colorado Boulder
Kiersten.Johnson@colorado.edu

1 Introduction

The Online Mendelian Inheritance in Man (OMIM) database contains information about all known Mendelian disorders (over 7,300) as well as over unique 15,000 genes that are connected to these disorders. This is an incredibly valuable set of data because it allows us to further understanding of diseases through their respective genetic components. Since this data has thousands of diseases also allows for inference of gene function across the phenotype. Every year this data set expands as scientists discover more disease-gene correlations, but this data is still very sparse. For any one disease there are a multitude of genes that are influential to its existence. In 2019, there are still many gene-disease pairs that have been yet to be discovered.

This paper aims to utilize networks to further understand the disease-gene data set holistically, while also using these network wide algorithms to predict possible missing gene-disease associations and further classify the diseases provided into their respective disorder classes.

A note on this paper's organization: Section 1: Introduction will cover the methodology and results for general analysis of the network; Section 2: Attribute Prediction will cover the methodology and results for attribute prediction on the subset of data; Section 3: Link Prediction will cover will cover the methodology and results for link prediction on the subset of data. Conclusion will wrap the paper up.

1.1 Methods

As stated above, the data for this is study is the OMIM data set, but we will be primarily referring to a curated subset of the OMIM data produced in "The Human Disease Gene Network" 2007 journal article created by Goh et al. In the OMIM subset the authors have hand classified each disease as a particular disorder class (e.g. metabolic, endocrine, cancer). Ths OMIM subset will be used for each algorithm, since it is a smaller network and easier for computation. The study also did some cleaning to their data to create a more connected network. This same methodology will then applied to the larger OMIM data to see if it holds similar properties to the subset. To determine if we can extrapolate that an algorithm that does well on our subset will have similar results if applied to the whole OMIM data set.

In the OMIM data each disease variation is listed, for our purposes we combined these variations into one disease. For example *Spinocerebellar ataxia 1, 164400 (3)* and *Spinocerebellar ataxia 10, 603516 (3)* are condensed down to Spinocerebellar ataxia. Each variation's associated genes are now associated to the condensed disease. This allows for a more connected graph which is useful when applying different algorithms in sections 2 and 3.

In both sets of data, OMIM and its subset, a bipartite network can be constructed where one set of nodes are genes and one set is diseases, with the edges being the association between two nodes. When projecting the bipartite network the edges become weighted. For the gene projection the edge weight between node i and j represent how many diseases i and j share. Likewise, for the disease projection the edge weight between node i and j represent how many mutated genes i and j share. In the projection each node also takes on an attribute that is its degree in the bipartite network, this will be used for sizing in our visualizations. For the following results section the projections were visualised to look at clustering and other graphical traits.

In order to visualize each graph we used the NetworkX spring layout function to compute node positions, with spring constant relative to number of nodes in the graph of interest, s.t. $k = \frac{1}{(\#Nodes)^3}$ and number of iterations at 100. The spring function uses statics to find the most equilibrium position for all the nodes based on their weights and the spring constant, this process repeats 100 iterations. The color of the nodes in the subset represent that node's disorder class. The "The Human Disease Gene Network" study extrapolates disorder classes onto gene nodes, s.t. for some gene i if all the diseases were associated with a gene i are one

class, then that gene is labeled that class. If the gene has disorders associated with it are from multiple classes then the gene is labeled "Multiple". The different sizes of the nodes in each visualization is the degree it had in its bipartite representation, meaning that a node's size in the gene projection is relative to how many diseases are linked to it.

1.2 General Analysis

Starting with the OMIM subset data, we visualized the Disease-Gene Bipartite Network (Fig 1). In this visualization we noticed a large connected component with small subgraphs around it. Next we projected the subset data from the bipartite network into the disease network. Similarly to the bipartite network, clustering exists in the center. This visualization was interesting but it is difficult to see clustering, so instead we zoomed in and graphed the largest connected component in the network (Fig 2). From this view more clustering is apparent; an edge from i to j with a particular color shows that i and j share the disorder class, i.e. that edge is a in-community edge. Where as the gray edges are between community edges.

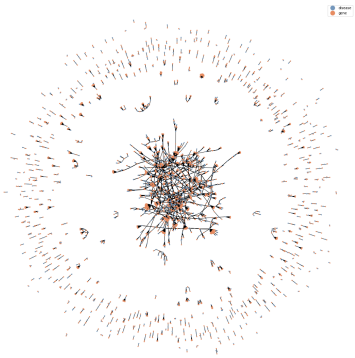


Fig. 1. Disease-Gene Bipartite Network OMIM Subset. Nodes = 5,107; Edges = 6,275

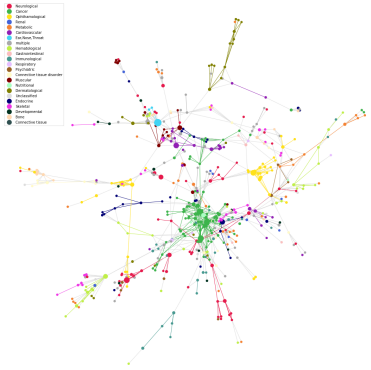


Fig. 2. Disease Proj. OMIM Subset: Largest Connected Component

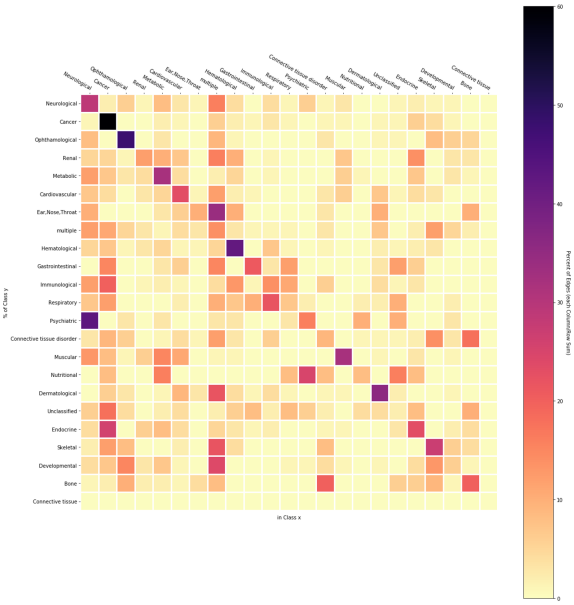


Fig. 3. In v. Between Community Edges for Disease Projection of OMIM Subset

Fig. 4. Disease Proj. OMIM Subset: Largest Connected Component Breadth First Traversal from Largest Degree Node

While clustering is apparent in Figure 2, the in versus out community edge matrix (Fig 3) shows a strong trend down diagonal of the matrix meaning that there is many in-community edges. The box color represents the percent of edges from the classes on the y axis connecting to classes on the x axis. It also shows the correlation between communities, for example a large percentage connected to Psychiatric nodes are linked with Neurological nodes. Cancer is linked primarily with itself, this is because there are many cancer nodes and edges due to a high quantity of research in that field. It is important to remember that the number of genes associated with disease is not because the disease is more complicated in the genome. But instead because of more research put into that disease to find more genes associated

with it. A breath first traversal from the largest degree nodes displays this community structure quite clearly (Fig 4), this is promising because we could use this projection transformation for attribute prediction.

Next we move on to the gene projection of the bipartite graph (Fig 5) starting with a zoom-in view of the central component. This visualization shows many out of community edges. This is largely due to our extrapolation of disorder classes on the genes.

Fig. 5. Gene Proj. OMIM Subset: Largest Connected Component

The same transformation as the disease projection are done. The gene network has more edges and therefore it can be harder to see community structure. The breadth first traversal (Fig 6) takes away a lot of these edges and reveals visible clustering.

Fig. 6. Gene Proj. OMIM Subset: Largest Connected Component Breadth First Traversal from Largest Degree Node

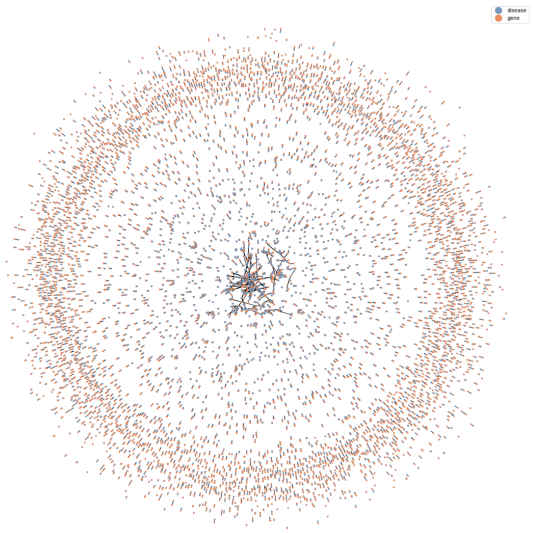


Fig. 7. Disease-Gene Bipartite Network OMIM Subset. Nodes = 21,733; Edges = 24,398

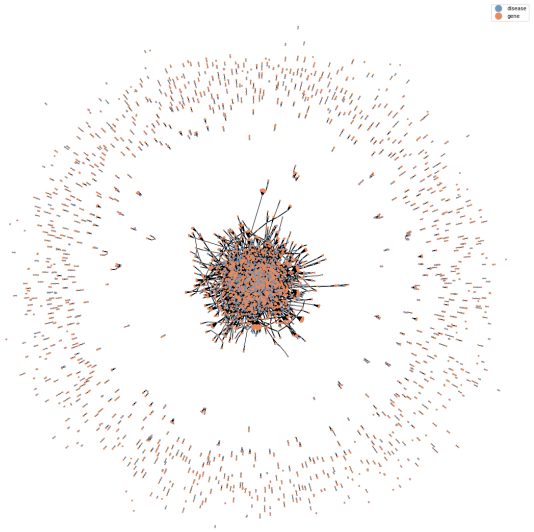


Fig. 8. Disease Projection OMIM Subset. Nodes = 16,844; Edges = 22,804

Looking at the larger scale OMIM network the same trends arise as with the subset. Even though we are primarily using the subset we wanted to see if the full data set showed similar spatial patterns. Since the OMIM data is not disorder classified we only have monotone visuals. In figure 7 is the bipartite gene-disease network visualized before condensing disease variants. Figure 8 shows the drastic increase of the connected component with the condensing of the variants, and looks similar to figure 1 of the subset data.

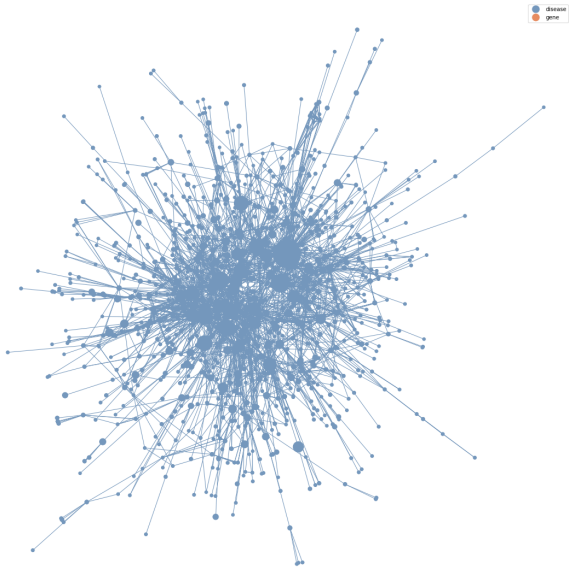


Fig. 9. Disease Projection OMIM Subset

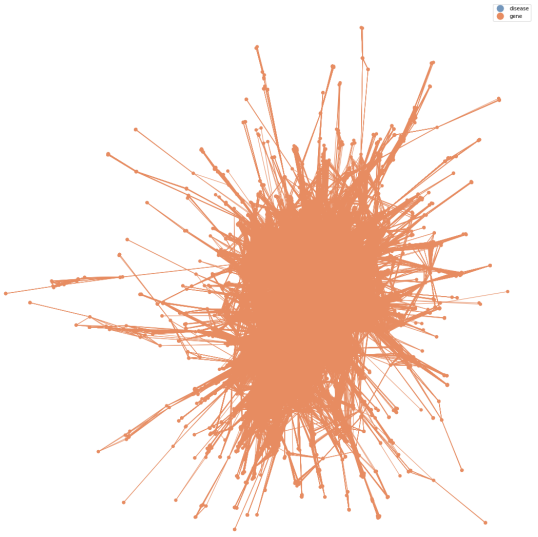


Fig. 11. Gene Projection OMIM Subset

Continuing to do the same analysis as the subset by separating the OMIM data into two projections, gene and disease. In disease projection it is hard to see any structure due to there not being labels on for the data. The gene project (Fig 11) is even worse because of the amount of edges in the projection blocking the spacial patterns from being viewed. When transforming both projections onto the breadth first traversal we see similar structure to our subset data. This is promising sign that attribute prediction could be able to use the manually classified subset classes to predict labels for the whole OMIM data set.

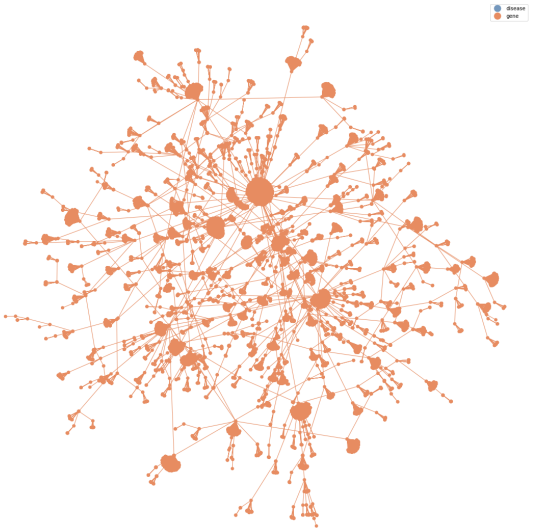


Fig. 12. Gene Projection OMIM Subset

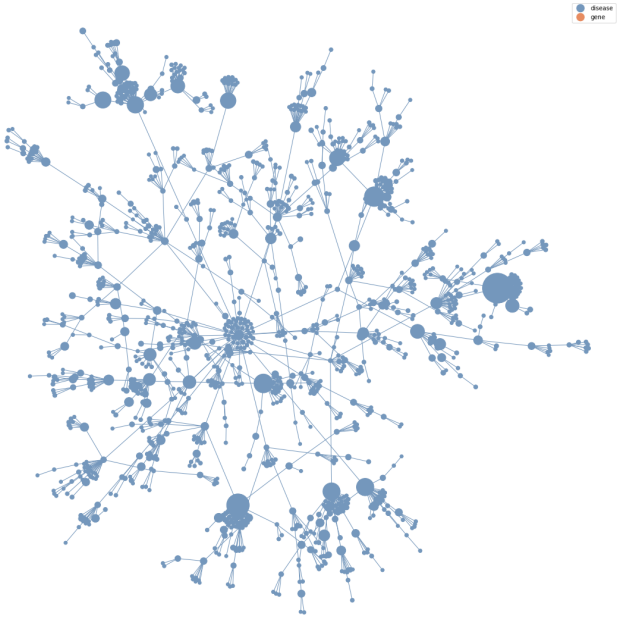


Fig. 10. Disease Projection OMIM Subset

2 Attribute Prediction

The attribute of disorder class was predicted for the disease projection. This is shown visually in the subset graphs above as the color of the nodes.

2.1 Methods

The data set used had manually assigned predictions for each of the disorders; these assignments were used to check the efficiency of the prediction algorithms created. Four different algorithms were used to predict: first neighbor, second neighbor, weighted, and a baseline (random).

The first neighbor algorithm looked at all of the first neighboring disorders. That is, all of the disorders that had

at least one gene mutation in common with the central disorder were compiled. From this list, the function assigned the most frequent first neighbor disorder class to the central disorder. If there was a tie for the max number of neighboring disorders, the central disease was labeled as "Multiple".

The second neighbor had the same basis as the previous function. Instead of looking at first neighbors, a list was compiled of all disorder classes for second neighbors and the function found the most frequent class in this list to assign to the central node.

The weighted function takes into account how many mutated genes the first neighbor disorders have in common with the central disorder. The disease projection is formatted so that edges between the two disorder nodes are weighted according this number of shared disorders. For this function, the weight is taken into account and the class of a neighboring disorder has an impact in the prediction of the central disorder's class in accordance to how many mutated genes they have in common.

The baseline function randomly predicts one of the 16 disorder classes to assign to each node.

2.2 Results

Excluding the baseline prediction, each method produced has a differing accuracy as more of the model is observed (Fig 13).

The baseline has a constant 6.25 percent chance of assigning the correct class to disorders. The weighted accuracy appears to have a constantly increasing accuracy as more of the graph is available for the function to work with. The first neighbor prediction has the best accuracy on average as more of the graph is observed. This shows that disorders with at least one mutated gene in common with another disease are most likely to hold the same class identification. The weighted prediction accuracy has the best accuracy only at the end when the largest percent of the graph is observed. This shows that it does make a difference in taking into account how many mutated genes two disorders have in common to determine if they are classified the same. All of the prediction algorithms do about 2x to 3x better than the baseline random assignment.

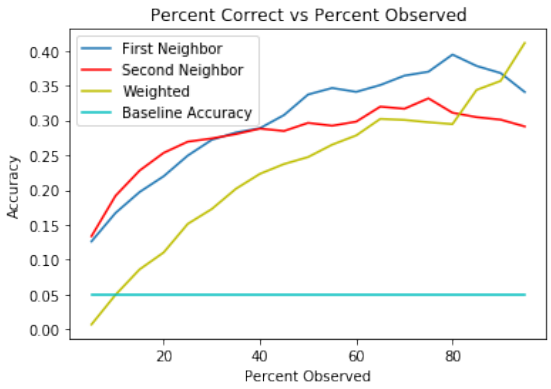


Fig. 13. Accuracy of each disorder class prediction method as percent of graph observed increases

The breadth first projection of the model has slightly different results compared to the disease projection (Fig 14). This model has a much higher accuracy for second neighbor prediction and lower accuracy for first neighbor prediction. This is likely due to the tree-like structure of the graph (Fig 4). Many disorders of the same class are not connected with first neighbors, even though they are "leaves" of the same "tree". By traversing two neighbors, these leaves are able to influence the class prediction for each other. Leaves are very likely to hold the same disorder class so this raises the accuracy a large amount.

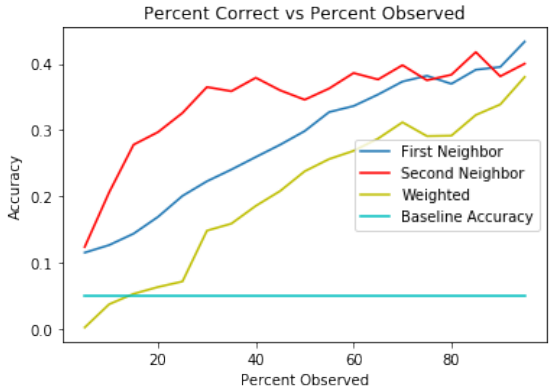


Fig. 14. Accuracy of each disorder class prediction method on the breadth first projection as percent of graph observed increases

3 Link Prediction

Link prediction on the OMIM data predicts missing associations between nodes.

3.1 Methods

In order to predict missing links you must iterate through all possible, and non-existent, links and index them based on their likelihood of having a edge. This score is called a similarity measure. The higher the measure the greater the possibility that there is a missing edge (by the structural assumptions of a certain measure). Five similarity measures are used on the OMIM data subset to predict missing links: Common Neighbors, Jaccard, Preferential Attatchment, Leicht-Holme-Newman, and Katz$_\beta$ Index. In order to predict edges, the possible missing edges are sorted descending by their similarity index and top N are picked to be predicted missing edges (N being a hyper parameter).

Normally, to find the accuracy of these different measures on a data set a certain number of observed edges are hidden, then the algorithm runs and predicts N edges exist. Many statistics describe the goodness of a certain measure like true precision, TPR, FPR, recall, and accuracy.

In our application we are predicting edges on the largely connected component in the OMIM data subset. The problem with the standard methodology in the case of our data is that the network is incredibly sparse and there are almost 2000^2 possible edges. Deleting edges at random could lead

to a disconnected graph. Because of this, many of the goodness statistics will not give valuable information. For example, if it is predicted that no edges are missing, the accuracy will still be very high because 2000^2 non-existent edges were not falsely predicted. Another problem with our data is that it is not complete, there could be missing links that we predict that haven't been found in research yet. These attributes of our data makes it more complicated to test and evaluate link prediction.

The summary statistics that are valuable to this large class difference, since there are more unobserved edges than observed, are precision and recall. Recall reveals the proportion of the hidden edges that are captured by our prediction method. Precision records the proportion of predicted edges that are actual edges; it is the probability that a predicted edge is a hidden edge.

Edge prediction would benefit geneticists researching a particular disease because it could to give them suggestions on what genes that might be associated. We decided to run our experiment from this premise.

Our process was to select 30 nodes at random, independently hide a node's edges with probability alpha then predict N edges, and then average all 30 nodes' prediction goodness statistics. We repeated this process for 10 iterations for each value of N.

Another problem arose: our network is bipartite. Many measures of goodness of fit are based around common neighbors, but a disease and a gene will never share neighbors. Therefore, we decided to do predictions on the projection graphs of the OMIM data subset.

3.2 Results

For disease projection link prediction the two best methods were Common Neighbors and Katz. Looking at the recall, precision, accuracy versus top N picked for Common Neighbors (Fig 15) we can see that the intersection of recall and precision lies around 0.30. This shows that links are most likely to be found between two nodes with common neighbors. The Katz B has a similar intersection at 0.3, but the recall increases to 0.9 as N increases. This is important because with 30 predictions almost 90% of the hidden edges are found. The performance of Katz B shows that viewing all paths, beyond just common neighbors, between each set of nodes is valuable for the disease projection.

For the gene projection the best two methods were Common Neighbors and Jaccard (Fig 15 & 16). Both of these methods rely on Common Neighbors. Therefore on this data set Common Neighbors is an important metric to predict missing edges. The intersection for both methods is around 25 predicted and 0.7 recall and precision.

While other methods were tested, shown in Appendix C, we decided to just analyze the top two prediction methods for both projections.

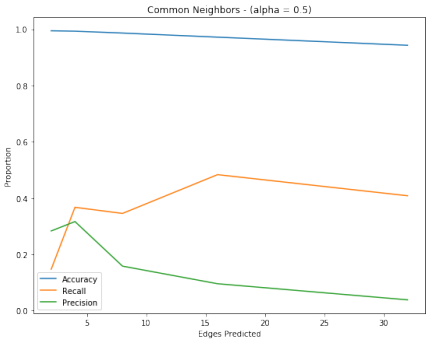


Fig. 15. Link Prediction on Disease Projection: Common Neighbors

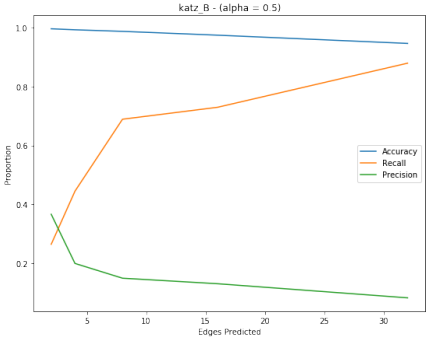


Fig. 16. Link Prediction on Disease Projection: Katz

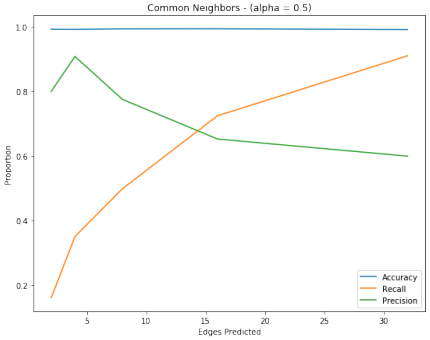


Fig. 17. Link Prediction on Gene Projection: Common Neighbors

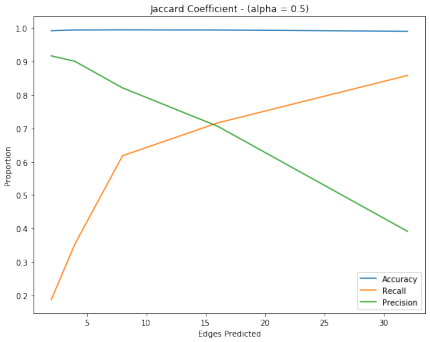


Fig. 18. Link Prediction on Genee Projection: Jaccard

4 Discussions

The implications that can be drawn from these analyses are vast. By imaging the networks, this alone has revealed patterns in the graphs both overall, and for the subset

and breadth first traversals. The clustered component makes it visually apparent that there are certain types of disorders that have many common mutated genes. The breadth first traversal shows how the disorder classes are often grouped. Furthermore, the visualizations within this paper are particularly helpful since they are subsets of the original data which is very convoluted (compare Fig 5 to Fig 9).

Attribute prediction also has potential uses in further research. By knowing which diseases belong to which disorder class, connected mutated genes are more likely to have impacts throughout all disorders of that class. Different methods were tested and revealed that overall the weighted function and first neighbors algorithms usually completed the best predictions.

Link prediction has a similar potential advantage to the research field, as noted before. Researchers would have suggestions from this as to which genes may be mutated and causing a disorder. This would save lots of time and money invested in research techniques if these connections to mutated genes may be made instead of guessing which ones may have an impact.

5 Conclusions

Networks created from a subset of data from the OMIM data set have shown visually and through analyzation how the relationships between diseases and mutated genes associated with them are predictable. In particular, we have shown these patterns visually overall, by predicting the disorder class attribute, and by conducting analysis to predict possible gene-disease associations that are currently missing

Acknowledgements

Aaron Clauset, Ph. D., for teaching the Biological Networks course at University of Colorado Boulder

References

[1] Goh, Kwang-Il, et al. "The Human Disease Network." PNAS: Proceedings of the National Academy of Sciences of the United States of America, vol. 104, no. 22, 22 May 2007 doi:https://doi:org/10.1073/pnas.0701361104.

[2] Online Mendelian Inheritance in Man, OMIM® McKusick-Nathans Institute of Genetic Medicine Johns Hopkins University (Baltimore, MD), Oct, 14 2019. World Wide Web URL: https:/omim.org/

Appendix A: OMIM Analysis Project Code Repository

github.com/sky123martinOmim_Network_Analysis_Project

Appendix B: The human disease network Study Visualization

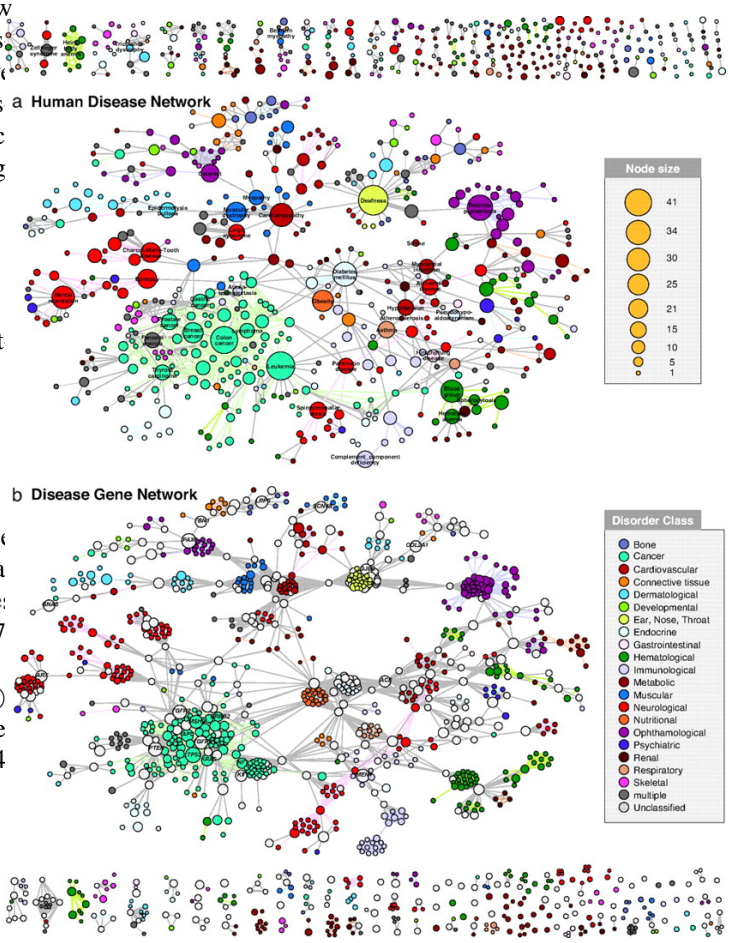


Fig. 19.

Appendix C: Link Prediction Full Results

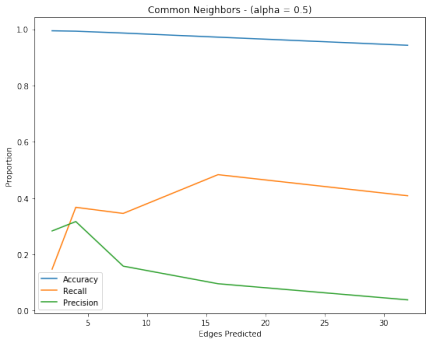


Fig. 20. Link Prediction on Disease Projection: Common Neighbors

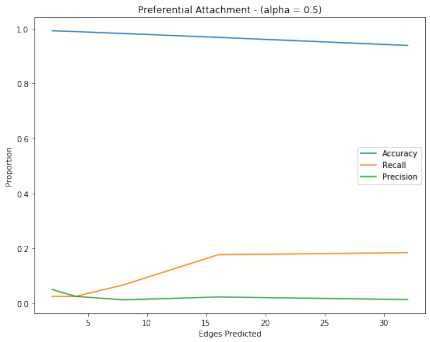


Fig. 21. Link Prediction on Disease Projection: Pref. Attachment

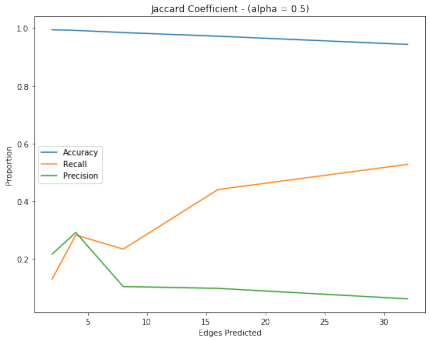


Fig. 22. Link Prediction on Disease Projection: Jaccard

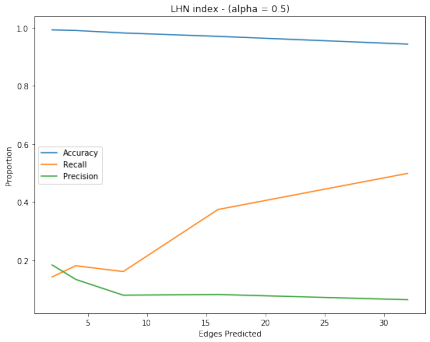
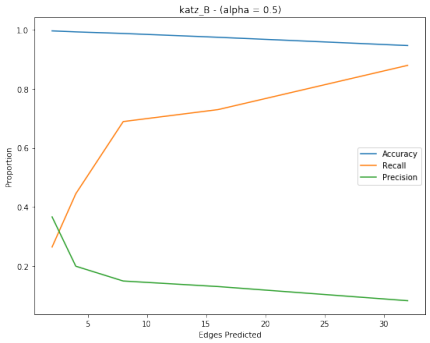


Fig. 23. Link Prediction on Disease Projection: LHN Index



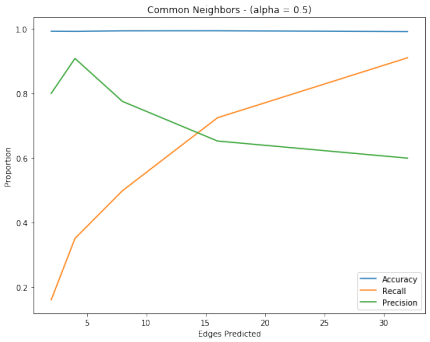


Fig. 25. Link Prediction on Gene Projection: Common Neighbors

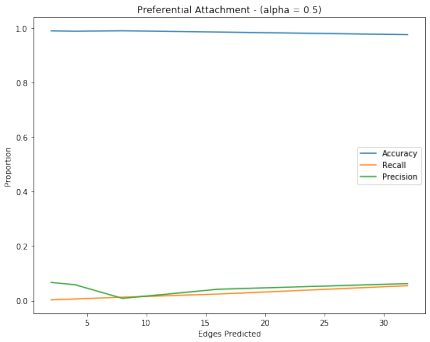


Fig. 26. Link Prediction on Gene Projection: Pref. Attachment

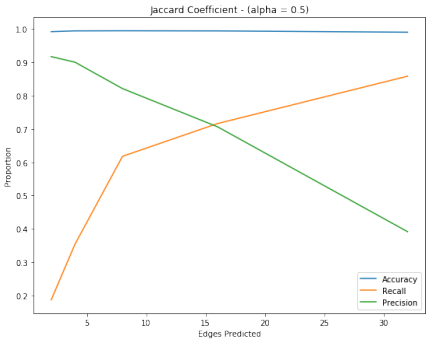


Fig. 27. Link Prediction on Genee Projection: Jaccard

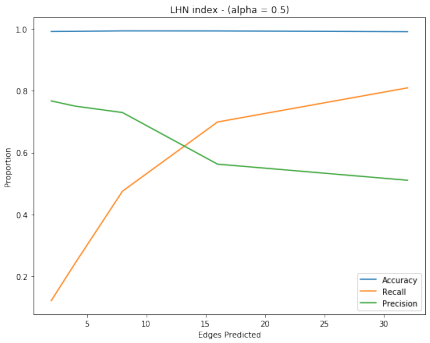


Fig. 28. Link Prediction on Gene Projection: LHN Index

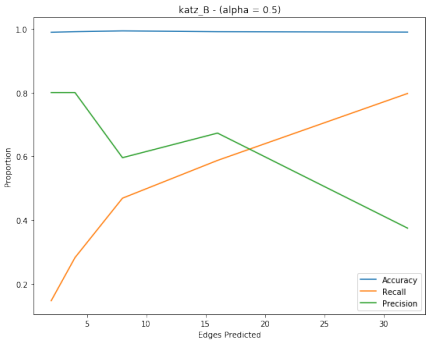


Fig. 29. Link Prediction on Gene Projection: Katz