Network Visualization

Grecia Plasencia Acosta, Lionel Dsilva, Joseph Griffin

Week 7/8 Group Project

Dataset 1: Link / Network analysis on the Wikipedia Vote dataset

I. Description of the Dataset

The dataset was retrieved from a text file hosted on the Stanford SNAP website. The data was then loaded into Excel and transformed into a CSV file type, filtering out a few irrelevant rows that describe the data being viewed. This data was then imported into the Gephi software and relevant visualizations and statistics were derived from it after that. From the description of the data, hosted on SNAP – "the dataset is sourced from Wikipedia, an encyclopedia written collaboratively by volunteers around the world. A small part of the contributors are administrators, who are users with access to additional technical features that aid in maintenance. For a user to become an administrator, a request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote that decides who to promote to adminship. Using the latest complete dump of Wikipedia page edit history we extracted all administrator elections and vote history data. This gave us 2,794 elections with 103,663 total votes and 7,066 users participating in the elections (either casting a vote or being voted on). Out of these 1,235 elections resulted in a successful promotion, while 1,559 elections did not result in the promotion. About half of the votes in the dataset are by existing admins, while the other half comes from ordinary Wikipedia users. The network contains all the Wikipedia voting data from the inception of Wikipedia till January 2008. Nodes in the network represent Wikipedia users and a directed edge from node *i* to node *j* represents that user *i* voted on user *j*".

There are 7,115 nodes and 103,689 edges. From the description, there were 7,115 administrators voted on by a total of 103,689 votes. In-links signify the number of votes received by a potential administrator, and each out-link from a node is a voted casted by a user. A sample of this is provided in the table below.

TABLE 1: SAMPLE FROM ADJACENCY TABLE FOR THE WIKI-VOTE DATASET

Source	Target
30	1412
30	3352
3	271
25	28
25	154

In the concept and study of networks, there are three types of graphs that can be analysed. Graphs whose edges have a known direction to them – a specified source and destination, which have an observable order and direction to them are known as 'Directed graphs'. In contrast to directed graphs, an undirected graph is a network that has no discernable direction or order to it. Supplementing the two of these, are mixed graphs – whose edges and nodes either have some direction to them or do not.

A network is defined by its diameter, which is the shortest distance between two of the most distant nodes and its density, which is the ratio of the number of edges to the maximum number of edges possible. The software, 'Gephi', is capable of ranking nodes through two different algorithms and a number of different metrics. These metrics and algorithms can be used to partition the network by modularity, ranking, density, degree and other such metrics to supplement the visualization – which aids in identifying trends and influential nodes. A summary of these algorithms and metrics is given below.

A. PageRank

PageRank assigns numerical values, or ranks, to pages (nodes in this case), based on backlink counts and ranks of pages providing these backlinks. It considers a model where a user starts at a webpage and performs a random walk by following links from the page they are currently based in. The PageRank of a page is the probability of that webpage being visited on a specific random walk. In essence, an in-link from an important page is worth more and a page is considered important if it has more in-links from other important pages. The issue with this algorithm is that some pages are considered 'dead ends. Dead end pages have in-links but have no out-links, which results in a random walk going nowhere after it reaches a page – this causes importance to leak. Another issue is spider traps – where out-links point to pages within a single group, that is – each node (page in this case) points to another node which creates a cycle between these nodes. This spider trap tends to absorb importance of a page. Google's PageRank solves the problem of spider traps by use of 'teleports. At any given time, a web surfer has two options to take: given probability β – to follow a link at random and given a probability 1- β – to jump to a random page. In the event the algorithm encounters a spider trap, the surfer will teleport out of it. This solution also solves the issue of dead ends.

B. HITS

HITS, also known as Hubs and Authorities, is an iterative algorithm that models linked webpages as a directed graph. The algorithm is considered a valid alternative to the PageRank algorithm and is often used as a supplement or validation tool to compare results. Its input is an adjacency matrix which represents a collection of items and value defining the number of k iterations, which in turn produces an output of hub and authority score vectors. The input data is pre-processed first into an adjacency matrix, which is followed by the initialization of the hub and authority vectors. These vectors are then updated using the HITS update rule $a = AT(Aa) = (ATA)a$, and $h = A(ATh) = (AAT)h$; where a represents the authority vector and h represents the hub vector. The updated vectors are then normalized and then sent to output. The update and normalizing processes are repeated until the number of iterations are reached. Although the Hubs and Authorities algorithm is efficient in computation, it can be manipulated through spam and has the potential to perform poorly when the wrong number of iterations 'k' is specified or selected.

C. Metrics

a. Degree: The degree of a node is an indicator of how connected a node is. An in-degree is the number of links coming into a node and an out-degree is the number of nodes exiting the node.
b. Density: The number of existing edges divided by the number of possible edges, with the assumption that there are no loops or duplications. A graph with a higher density suggests that it is strongly connected and robust.
c. Betweenness Centrality: Extended measure of degree which measures distance (shortest paths) between other nodes on a graph. A node is considered to have a high betweenness centrality if the shortest paths of many pairs of nodes in the graph pass through the source node. Nodes with high betweenness also influence the flow of information throughout the graph.
d. Closeness Centrality: Measures the centrality of a node by its closeness to other nodes. The measure decreases if the number of nodes reachable from the source node decreases or distance between node increases.
e. Community modularity: Measures how well a network is divided into communities.
f. Authority: A measure of a node or vertex based on the number of in-links to the node. A large value indicates that a node is considered an authority.
g. Hub: Measure of a node based on out-links from the node. A node is considered a hub if it has more out-links than in-links – meaning that it points to more nodes than other nodes point to it.

II. Results

When visualized, the edges and nodes produced the following directed graph shown in Figure 1. The graph has been partitioned by modularity class, which measures how well a network is divided into smaller communities. Table 2 shows what percentage of the nodes and their links are divided up into communities, which are represented by their numerical value and their colour.

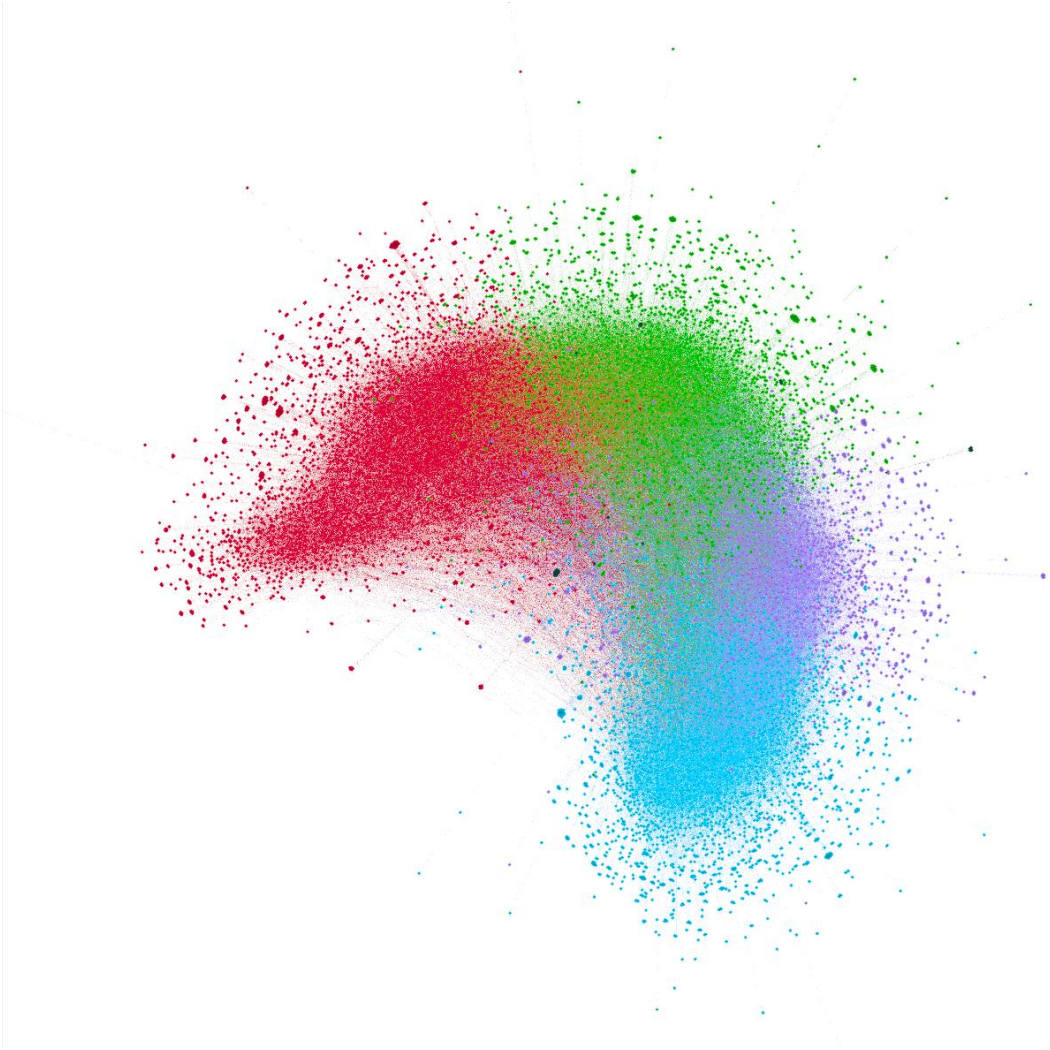


Figure 1: Wikipedia Vote Network by Modularity

TABLE 2: PERCENTAGE OF NODES IN A COMMUNITY

Modularity Class (Colour)	**%age of nodes within community**
5 (Red)	28.51%
3 (Light Blue)	28.47%
2 (Green)	24.65%
4 (Lavender)	16.34%
7 (Dark Green)	1.25%

The modularity for this network is 0.43, and forms 32 different communities – with most of these nodes belonging to communities 5, 3, 2, and 4. The remaining consist of 0.03% of the remaining nodes per community. Figure 2 shows the distribution of the network by modularity class.

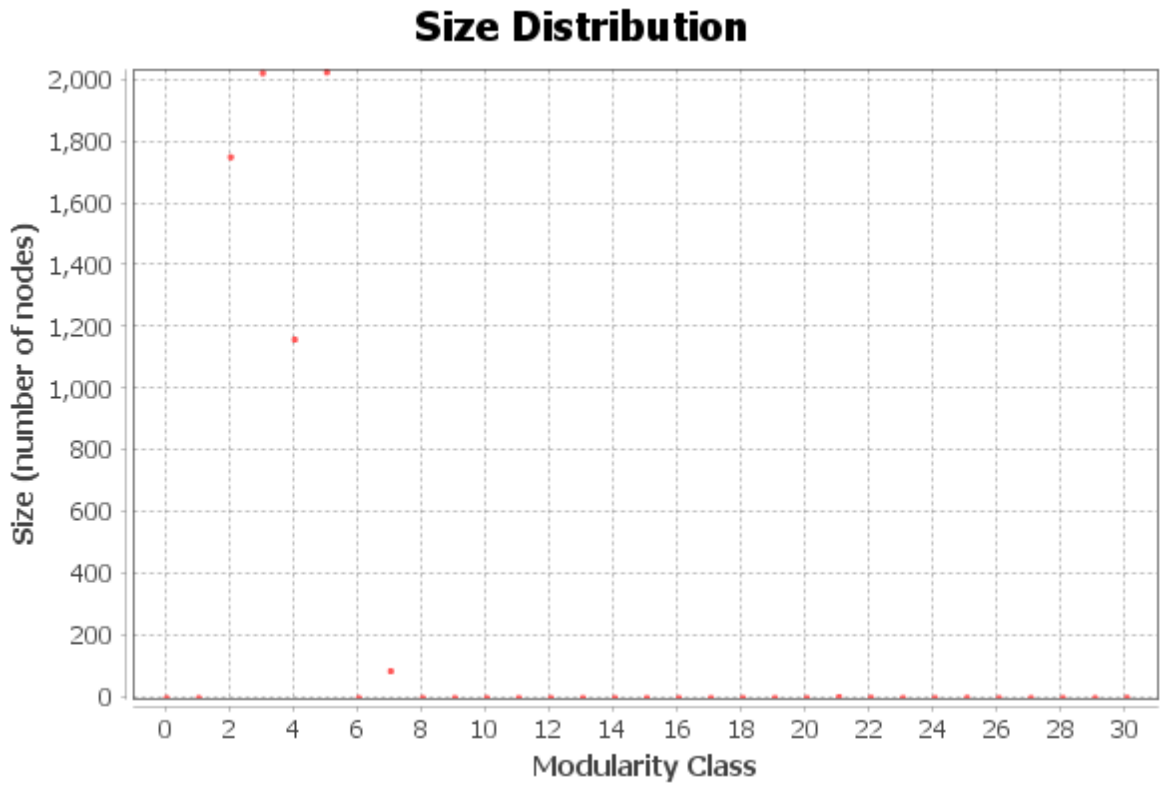


Figure 2: Scatter Plot Showing the Number of Nodes Present within a Modularity Class

The distribution is uneven – as most nodes fall within classes 5,3,2,4, and 7, while the remainder are sparse. Since modularity is a measure that computes the strength of a division of a network into communities or clusters, only networks with high modularity are dense in the number of connected nodes are considered significant. However, it is important to note that modularity suffers from a resolution limit and is unable to detect small clusters, or communities.

If an egocentric network with a depth of 1 is constructed – the network can be filtered by specific nodes. As an example, node 1412 was used to construct the following egocentric network with the ‘OpenOrd’ layout. Figure 3 demonstrates this network.

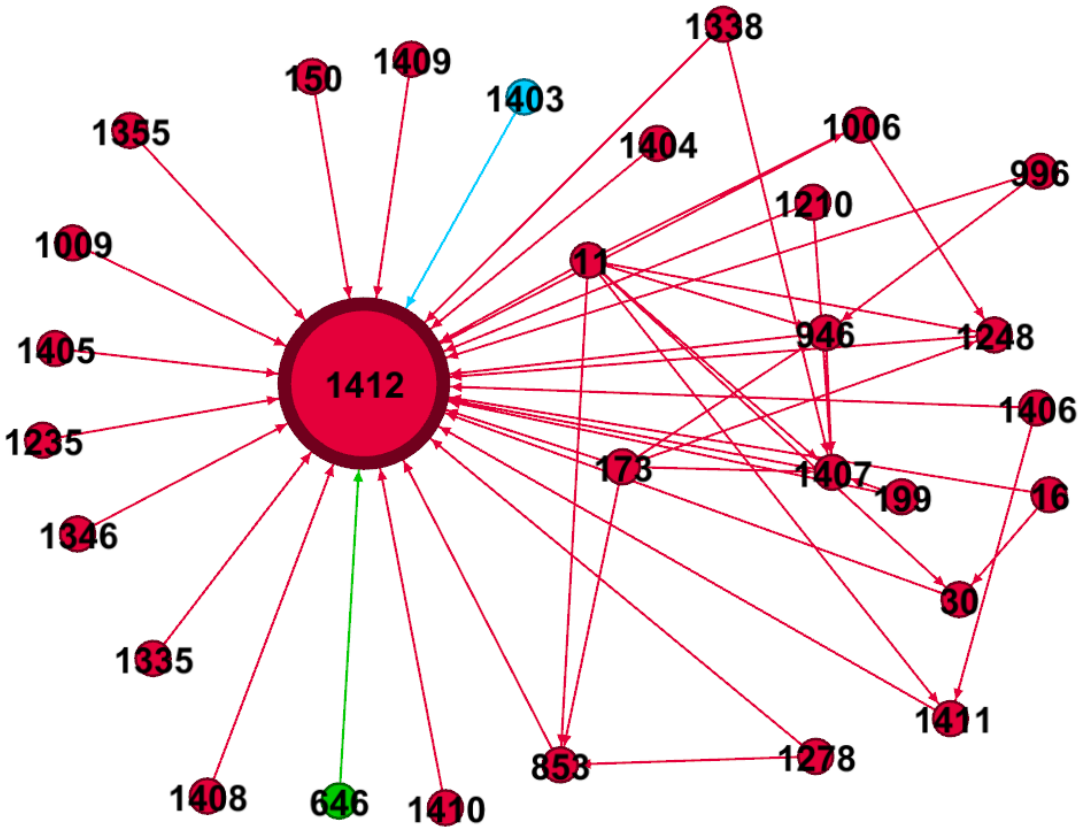


Figure 3: Egocentric Network of Node 1412 with In and Out Links

In the context of node 1412, we can approximate that all in-links from corresponding nodes are users that voted for noted 1412. An interesting observation to note here is that there are different subnetworks within 1412 – such as 1412 votes 1210, who votes for 1407, who then again votes for 1412. These patterns show a series of cyclical events, which can be viewed in greater detail when the depth of the modularity filter is further increased.

In a similar fashion, a PageRank analysis on the Wikipedia vote dataset gives us the following output in Figure 4.

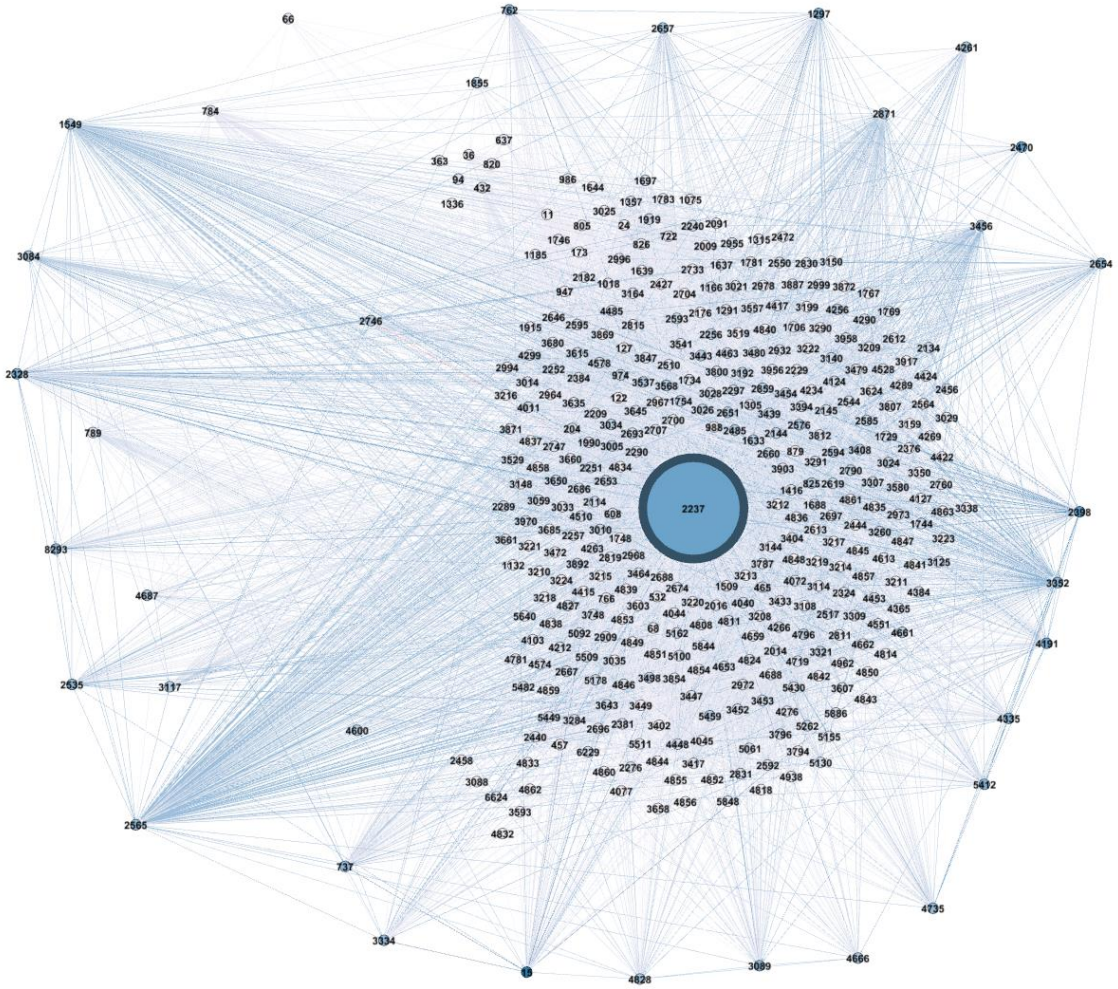


Figure 4: PageRank of Node 2237

In Figure 4, node 2237 is selected at random as an example of an influential node. The darker the node is coloured blue, the more connected it is to others. Node 2237, with a PageRank value of 0.002496, shows a high degree of connectedness – meaning that there are a lot of nodes that it points to, and there are an equal or more number of nodes that point to node 2237. Similarly, if we take the example of Node 4037, with the highest PageRank value in the dataset of 0.004606 returns the following result in Figure 5 when partitioned by modularity.

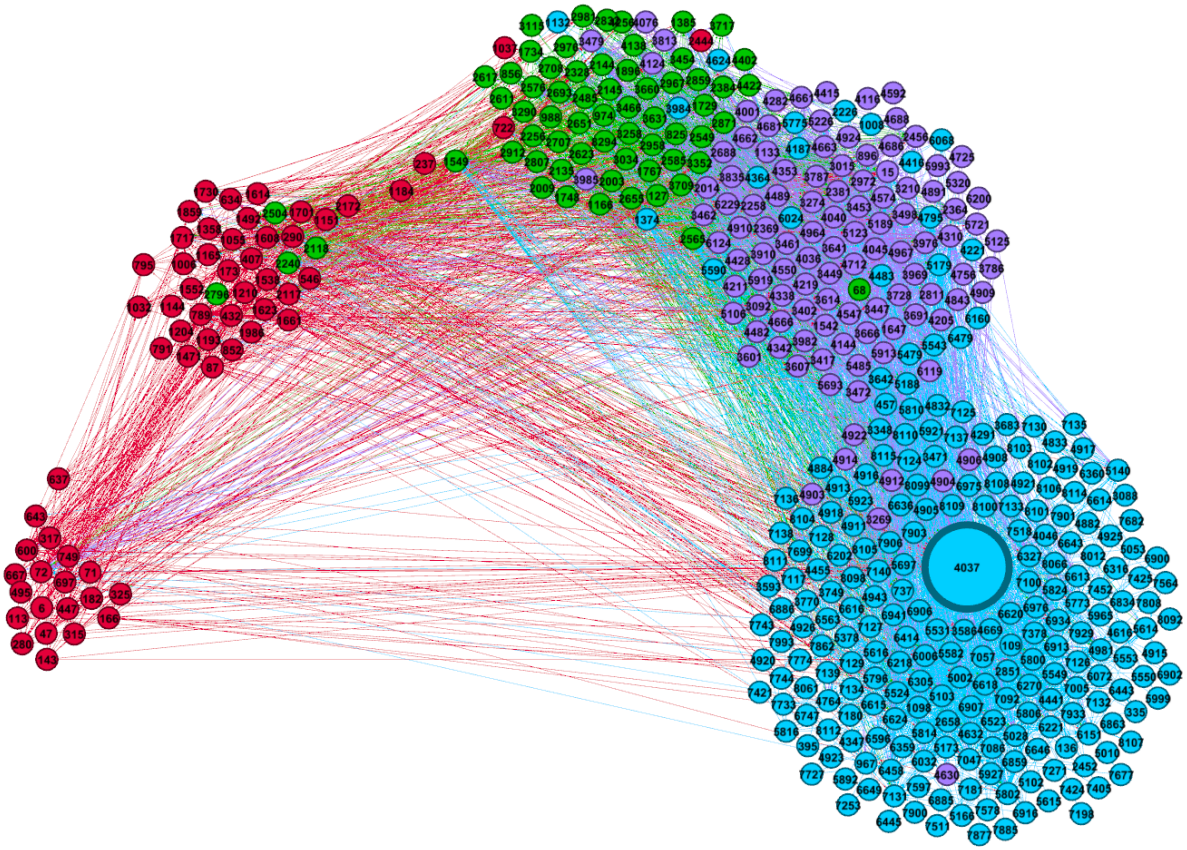


Figure 5: PageRank of Node 4037, Partitioned by Modularity Class

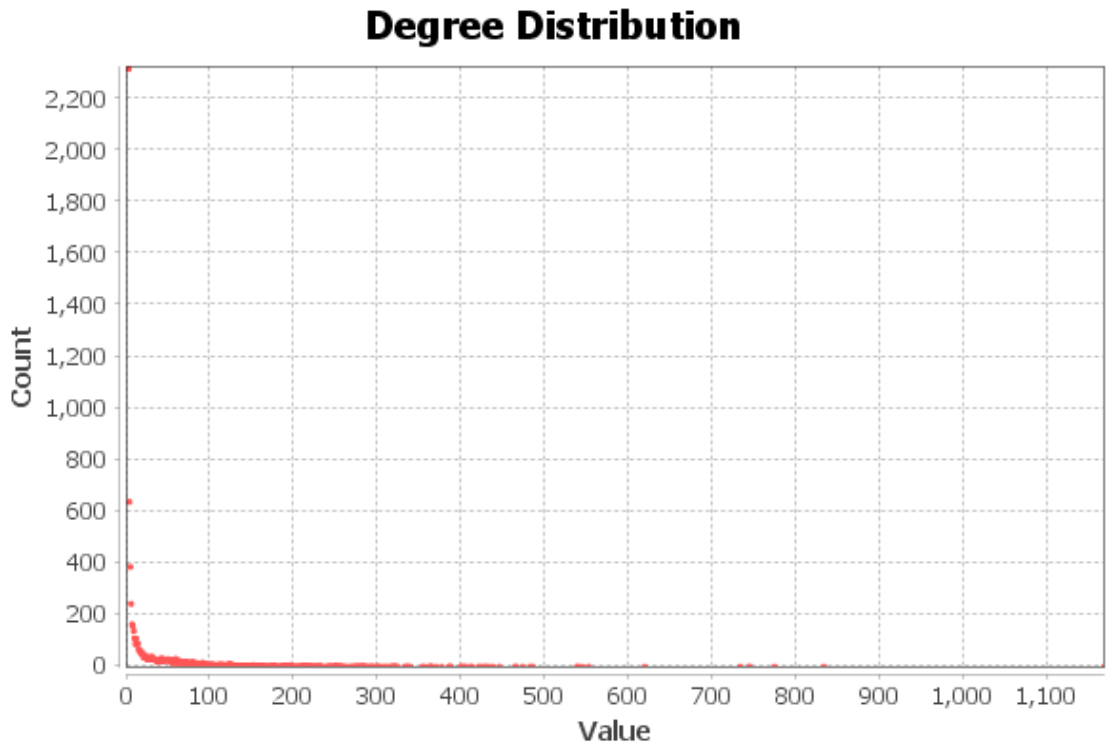
The egocentric network of Node 4037, when partitioned by modularity, shows how different communities of networks are able to interact with each other solely based on their in-links and out-links.

Table 3 shows the top five nodes that have the highest PageRank values – which determines how often they are pointed to by other nodes. We can infer from this table that according to the PageRank algorithm, the following nodes were voted on the most.

TABLE 3: PAGERANK RESULT SAMPLE

NodeId	Modularity Class	PageRank	Degree
4037	3	0.004606	457
15	4	0.003679	361
6634	3	0.003585	203
2625	2	0.003283	331
2398	2	0.002608	340

To supplement the results of the PageRank algorithm, the statistics for the degree distribution was generated, which is found in Figure 6.



As specified in the previous section – the Degree measure is an indicator of how connected the node is. When compared to Table 3, there is a noticeable relationship between PageRank value and the value of the degree measure. High degrees of connectedness indicate that specific nodes have a high number of in and out degrees, while high PageRank values indicate how often a specific node is pointed to, so it would not be inaccurate to hypothesize that a node with a high degree would be pointed to by other nodes more often than others.

In a similar fashion, the Hubs and Authorities algorithm (HITS) gives better insight into which nodes are most likely to be considered potential supervisors and which nodes are uses through its Hubs and Authority measure. Succinctly, a node is considered an authority if it has more in-links than out-links; which means that a node with a high authority value would more likely be a potential administrator in the context of this dataset. A node with a high number of out-links would be considered a hub, and as a result – a user. Table 4 illustrates five nodes with high authority values, alongside their hub values.

TABLE 4: HITS RESULTS (TOP 5)

NodeId	Authority	Hub
2398	0.092119	0.022428
3352	0.083132	0.093338
1549	0.08225	0.157912
1297	0.080337	0.023572
2565	0.079388	0.219184

Note that there are nodes that have a high authority value as well as high hub values – case in point, Node 3352. Figure 6 shows as egocentric network around 3352, filtered by in-degrees in the range of 1 to 457, and ranked by authority; the higher the authority value, the greener a node is coloured. This shows how many nodes provide in-links to 3352.

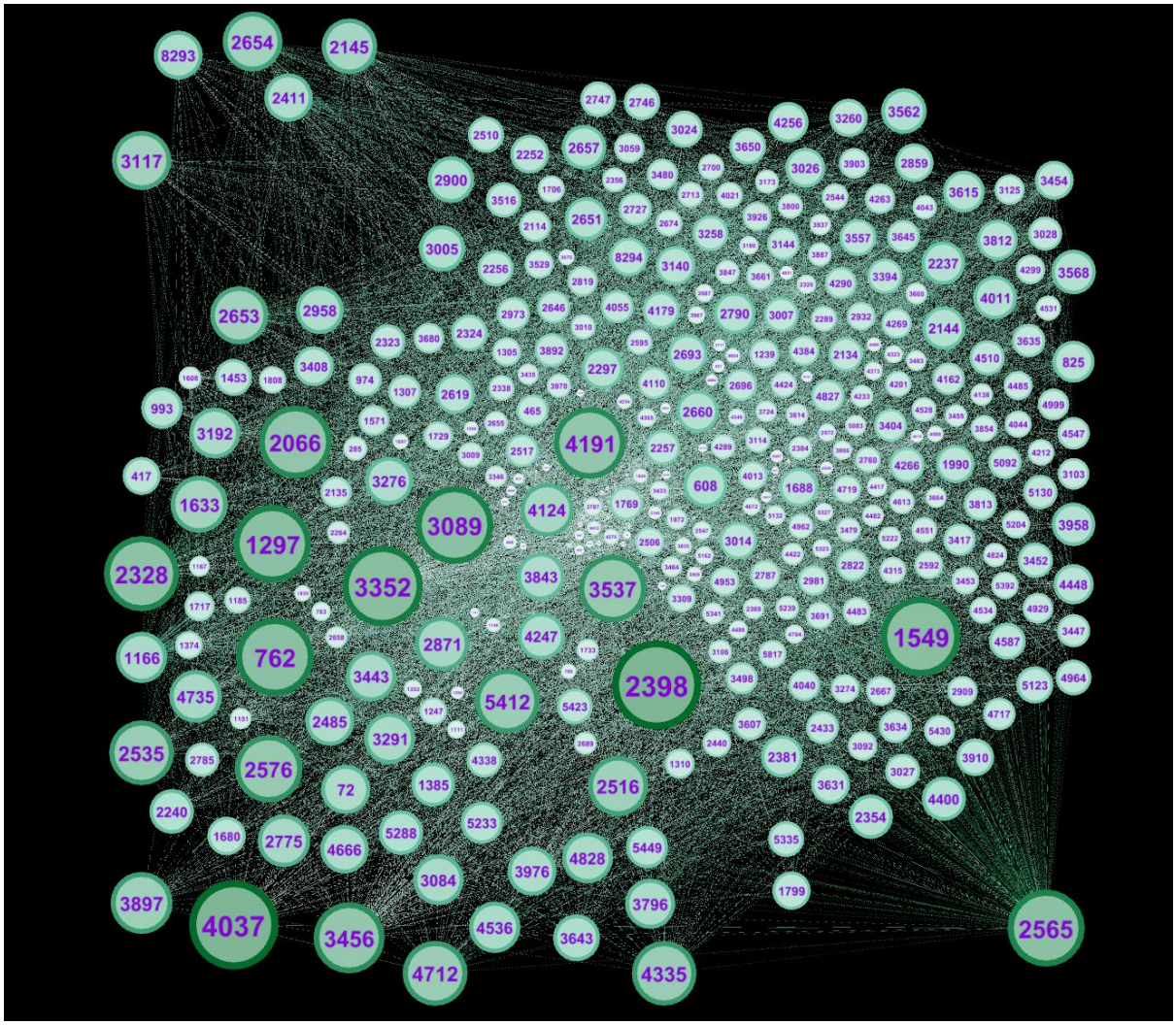


Figure 6: Egocentric Network of Node 3352 Filtered by In-links to 3352, and Ranked by Authority Values of Connected Nodes

If the filter is changed to an out-link filter, the result is Figure 7.

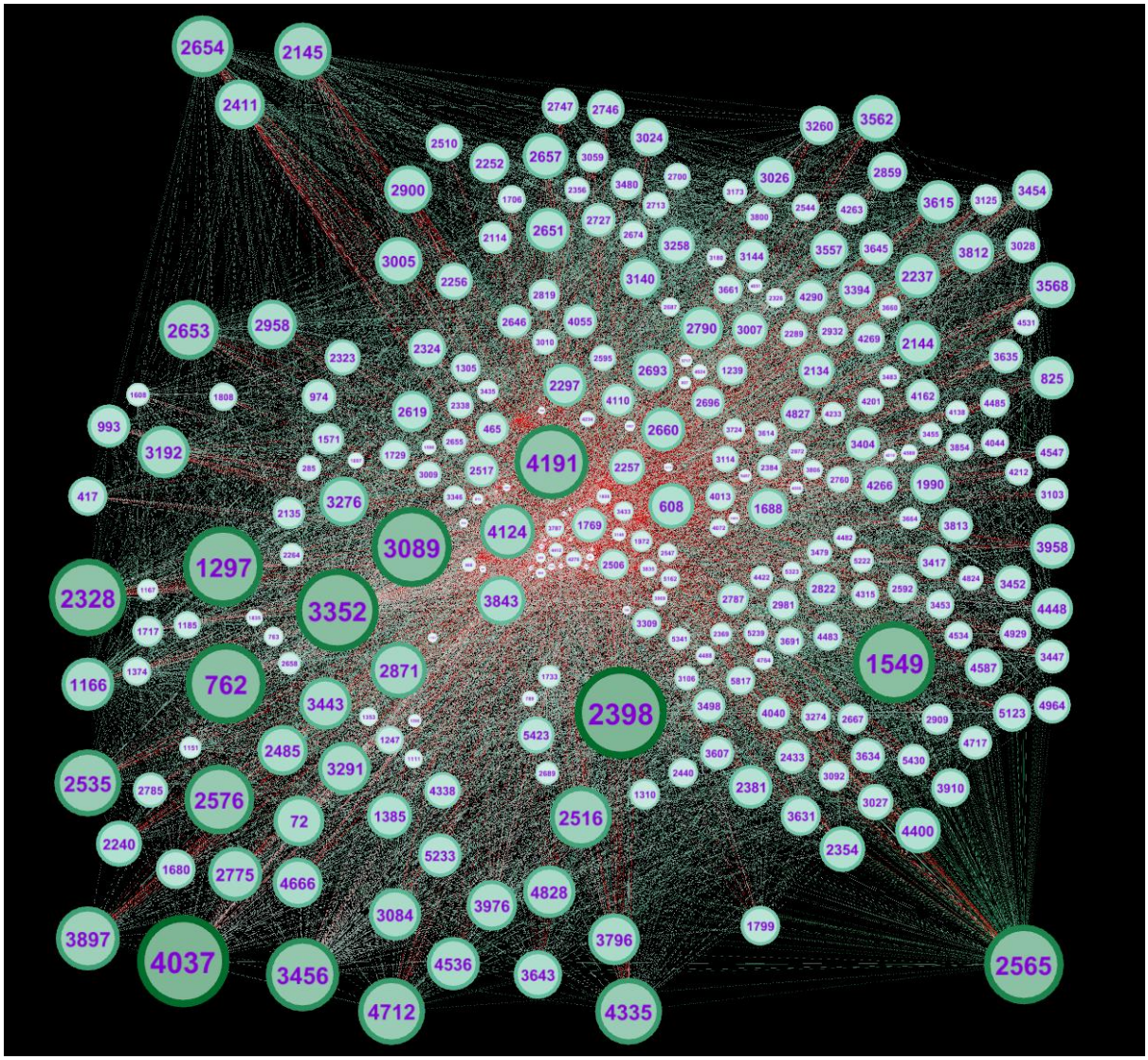


Figure 7: Egocentric Network of Node 3352 Filtered by Out-links from 3352, and Ranked by Authority Values of Connected Nodes

While not visible immediately, there are more out-links than in-links to 3352 (497 to 893). When cross referenced and ranked with the hub values, we see a similar network, however a new node is revealed to have a higher hub value than 3352 – namely node 2565 (Hub value 0.2917), as seen in Figure 8.

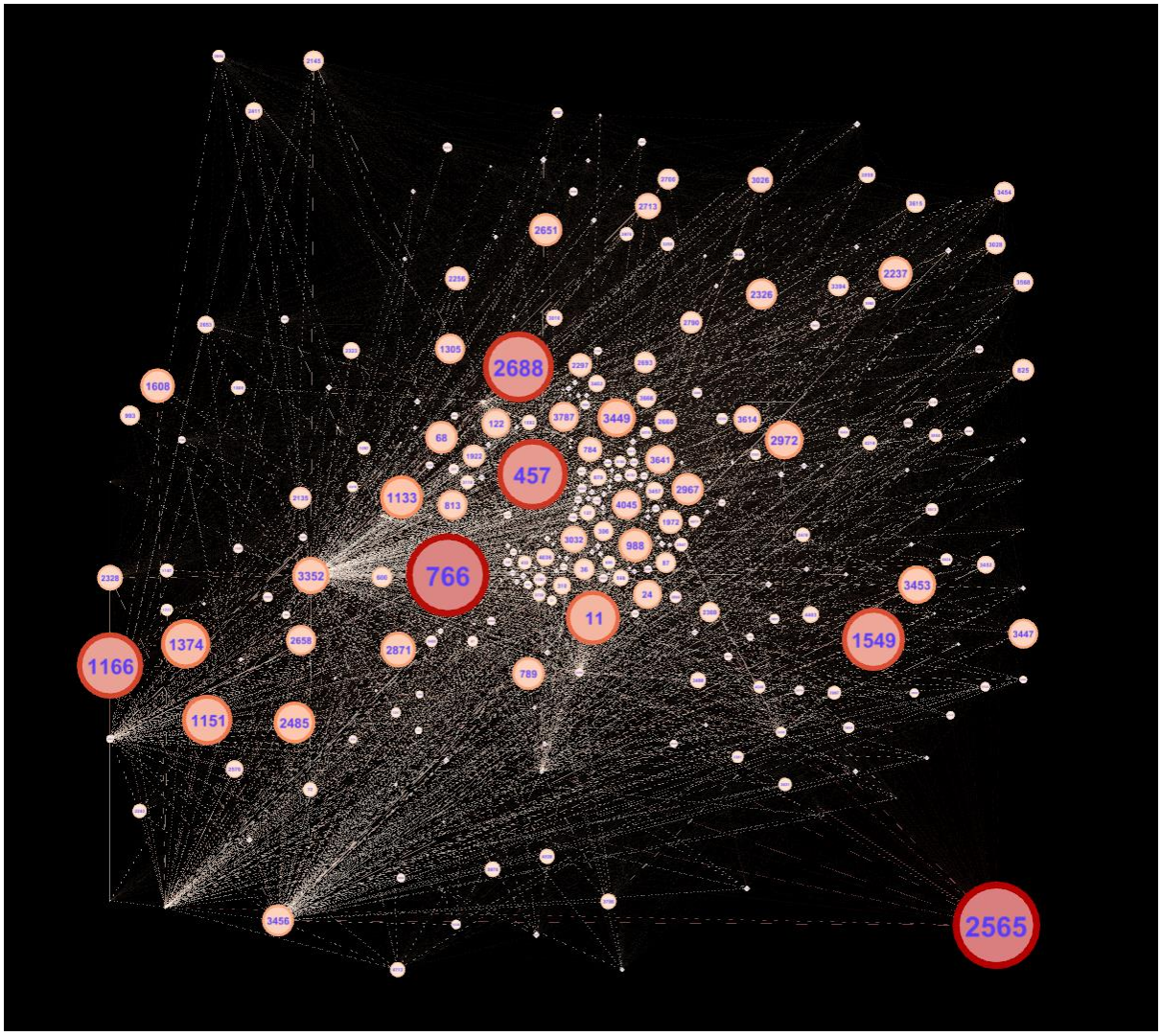


Figure 8: Egocentric Network of 3352 Ranked by Hub Values, Revealing Nodes with Higher Hub Values

III. Conclusions

Both the PageRank and HITS algorithms give us some insight into the nodes that received the most votes, or in-links. By PageRank results – node 4037 is the likeliest winner of the election this dataset is based on, which is corroborated by the In-degree and Out-degree values that are available in the data laboratory. In contrast, the HITS algorithm says that node 2398 is the likely winner, given its high authority score, however its degree value is lower than that of node 4037. Table 5 compares relevant values of the two competing nodes.

TABLE 5: LIKELY WINNER RESULTS

NodeId	In-degree	Out-degree	Degree	Authority	Hub	PageRank
2398	340	62	402	0.0921	0.0224	0.00260
4037	457	15	472	0.0918	0.0050	0.00460

There are multiple nodes that have a higher degree value than the two nodes in Table 5, but their in-degree, out-degree, authority, hub, and PageRank values are significantly lower than nodes 2398 and 4037, which in context, means that other nodes acted either as competition or acted as out-links, providing votes to nodes 2398 and 4037.

Dataset 2: Network Analysis of the LastFM Users Dataset

I. Description of the Dataset

LastFM is a music tracking and streaming service that creates personalized recommendations based on listening habits. It also allows for follower relationships amongst friends and listeners with similar musical tastes [4]. The dataset selected was obtained from the Stanford Snap website. It is a social network composed of LastFM users based in Asia in CSV format [1]. The data was collected from the public API in March 2020. The dataset's nodes are users and the edges are mutual follower relationships between them. There are a total of 7,626 nodes and 27,807 edges in this dataset and it is undirected. There are no clear labels for this data so it was not possible to obtain information apart from the relationships presented.

The Gephi software allowed for analysis and exploration of the selected data. Once the dataset was uploaded into Gephi the ForceAtlas 2 layout was selected to better visualize the relationships within the dataset. Clustering quickly became apparent with this layout and clear divisions were displayed. Once the data was better visualized the average degree statistic was run. The result was an average degree of 7.293. The degrees for this data were the number of mutual follower relationships so for each user, or node, there are approximately 7 follower connections as shown in Figure 9.

II. Results

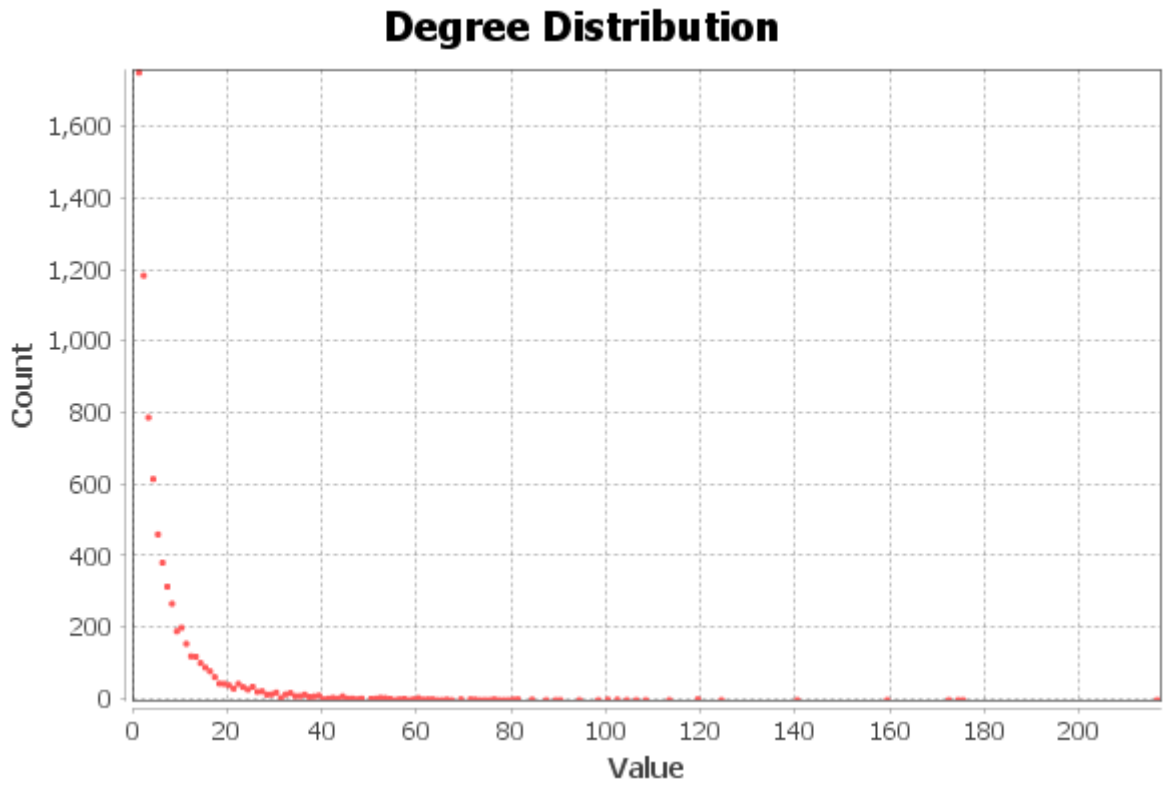


Figure 9: Average Degree Graph for the LastFM Users Dataset

Once the layout was run the Modularity algorithm was used. Modularity measures the solidity of relationships between nodes and clusters. The modularity for this dataset was calculated at 0.814 with 23 communities. Figure 10 displays the size distribution for the dataset.

Figure 10: Modularity Size Distribution [2] for the LastFM Users Dataset

The appearance was modified such that the nodes were partitioned by Modularity Class with the default color palette as shown in Figure 11. The major node clusters became much clearer and Table 6 shows the percentages of the main clusters. Class 2 was the largest with 14.35% of the total closely followed by class 7 which had 13.11%. Given the modularity score of 0.814 it follows that the clusters would be dense with possible overlap between some of them. There are also smaller clusters within the dataset that are not as significant but worth mentioning. Over 20% of the total, or 15 of the 23 classes, fall within this category. These clusters are also visualized in Figure 11 and given a gray color.

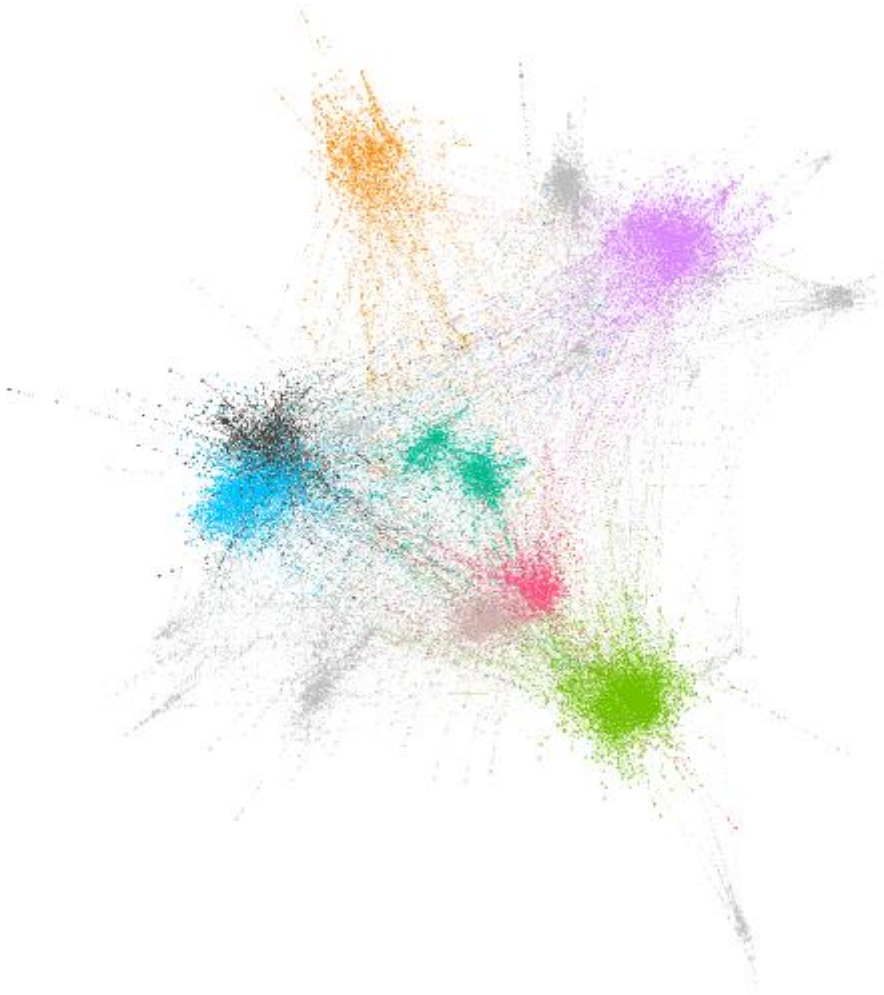


Figure 11: Mutual Follower Relationships amongst Asian LastFM users by Modularity

TABLE 6: MODULARITY CLASS PERCENTAGES FOR LASTFM USERS

Modularity Class (Colour)	%age of nodes within community
2 (Pink-Purple)	14.35%
7 (Green)	13.11%
12 (Light Blue)	10.71%
11 (Black)	8.79%
1 (Orange)	8.75%
8 (Bright Pink)	6.5%
6 (Turquoise)	6.45%
14 (Light Brown)	6.33%

After the modularity was calculated the data table was examined, in particular the degree column. The Degree column was sorted from largest to smallest to identify the node with the most connections, this had the node ID 7237 with a degree of 216. As an aside, the node with the second highest degree had node ID 3530 with a total of 175 degrees, thus there was a difference of over 40 degrees between the two 'top' nodes. With the previous information the graph was filtered under Topology with an Ego Network. The node ID input was 7237 to visualize that particular node with a depth of 1. Node 7237 was part of class 7, the second largest. Since the interest was in the node, and not the class, more alterations were made. The

size of the nodes were increased to a maximum of the degree of the largest node, in this case the maximum degree was 216 as previously mentioned. The Noverlap layout was run to better visualize the connections between the main node and the rest. Following that the node ID was labeled to clearly identify the outlier and it's degree network, the labels were sized by the degree of the individual node. Figure 12 below visualizes this. It also visualizes the connections node 7237 has with some of the nodes outside of class 7.

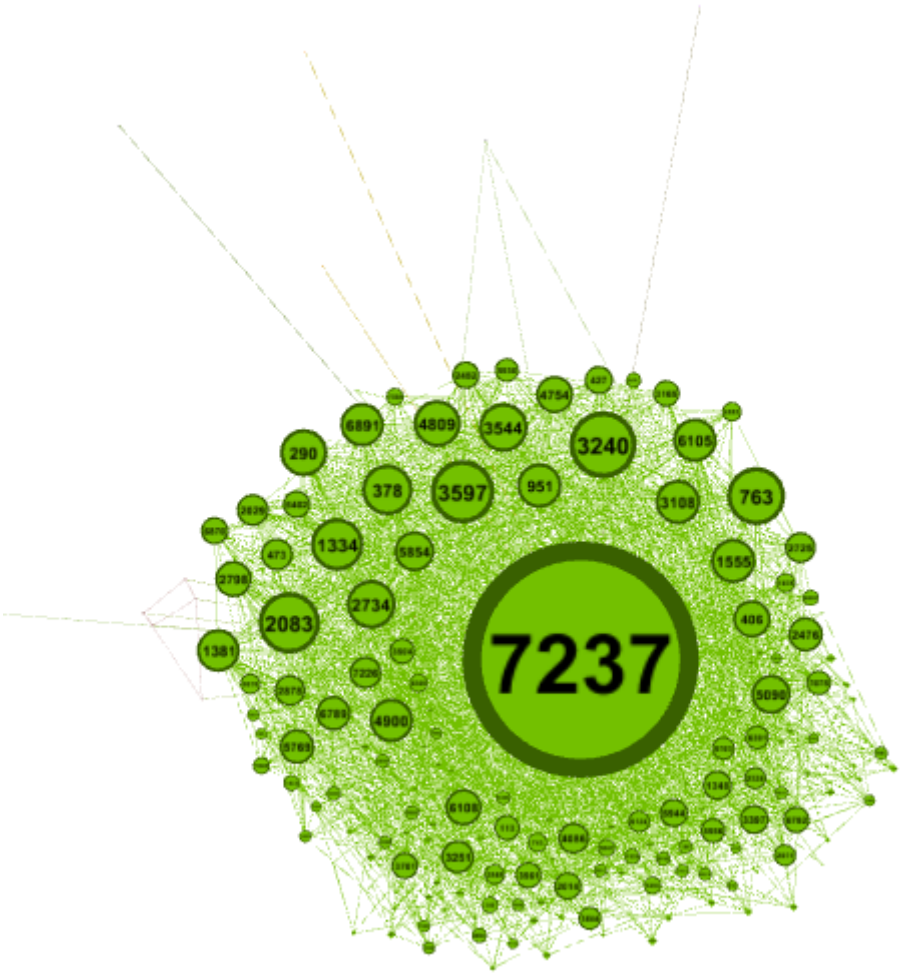


Figure 12: Egocentric Network of the Node with the Highest Degree (Node 7237)

After the Degree column was sorted and the highest node visualized the same was done with the PageRank column. PageRank is used to measure significance, in this case of a particular node. When the PageRank algorithm was run it had an epsilon of 0.001 and probability 0.85. The highest node was ID 4811 with a PageRank score of 0.003271. Again the graph was filtered by that node using Ego Network but the labels were modified for clarity. This node had a degree of 113 and a modularity class of 11 as is displayed in Figure 13.

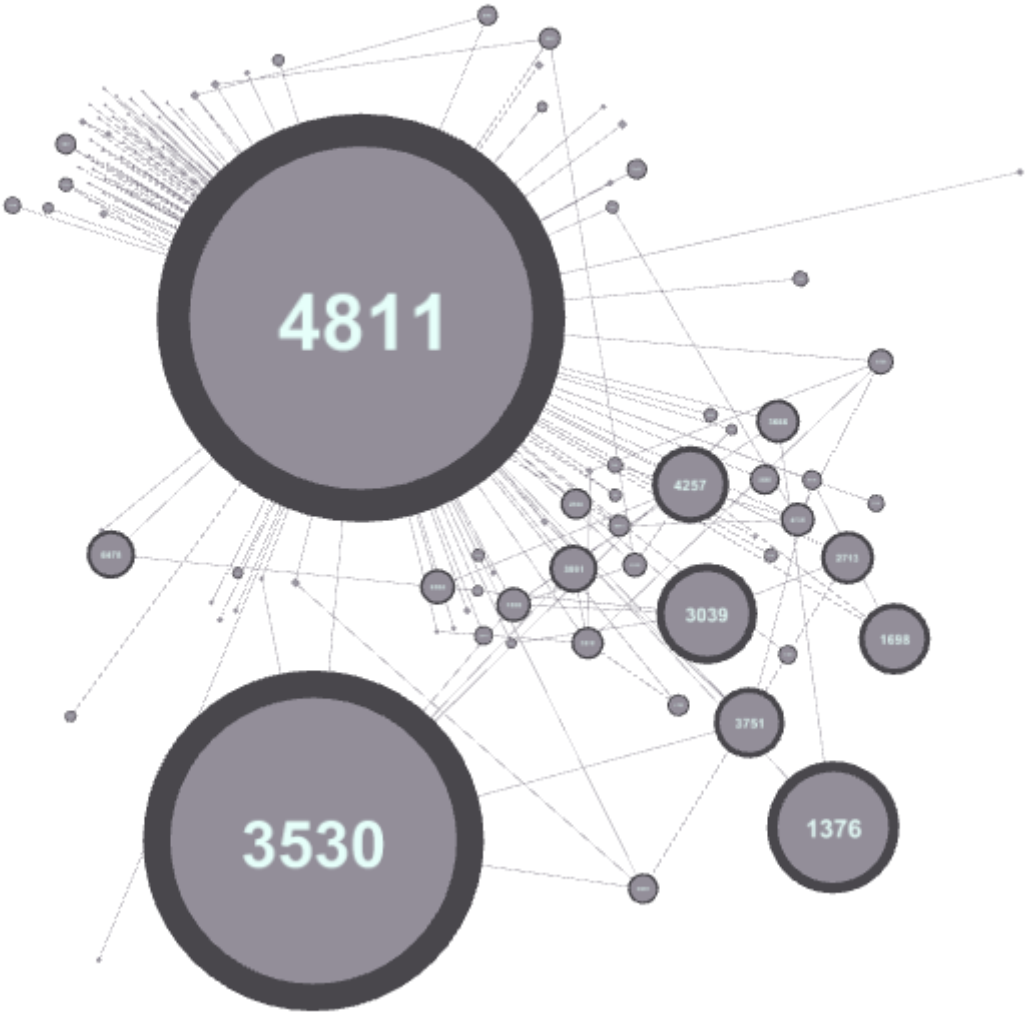


Figure 13: PageRank Partitioning of Node 4811

Also displayed in Figure 13 are nodes 3530, 1376, 3039 and 4257 which rounded out the top five nodes with the highest PageRank scores. These nodes are not in the same cluster, however they are an example of the overlap between the classes, in this case due to the fact that these nodes display the highest connectedness. To further visualize this Figure 11 above shows the overlap between classes 11 and 12 while Figure 14 below contains the same PageRank graph but partitioned by Modularity to display the links between the nodes and the classes. Table 7 below describes the statistics of these nodes to further cement the strength of their relationships with each other.

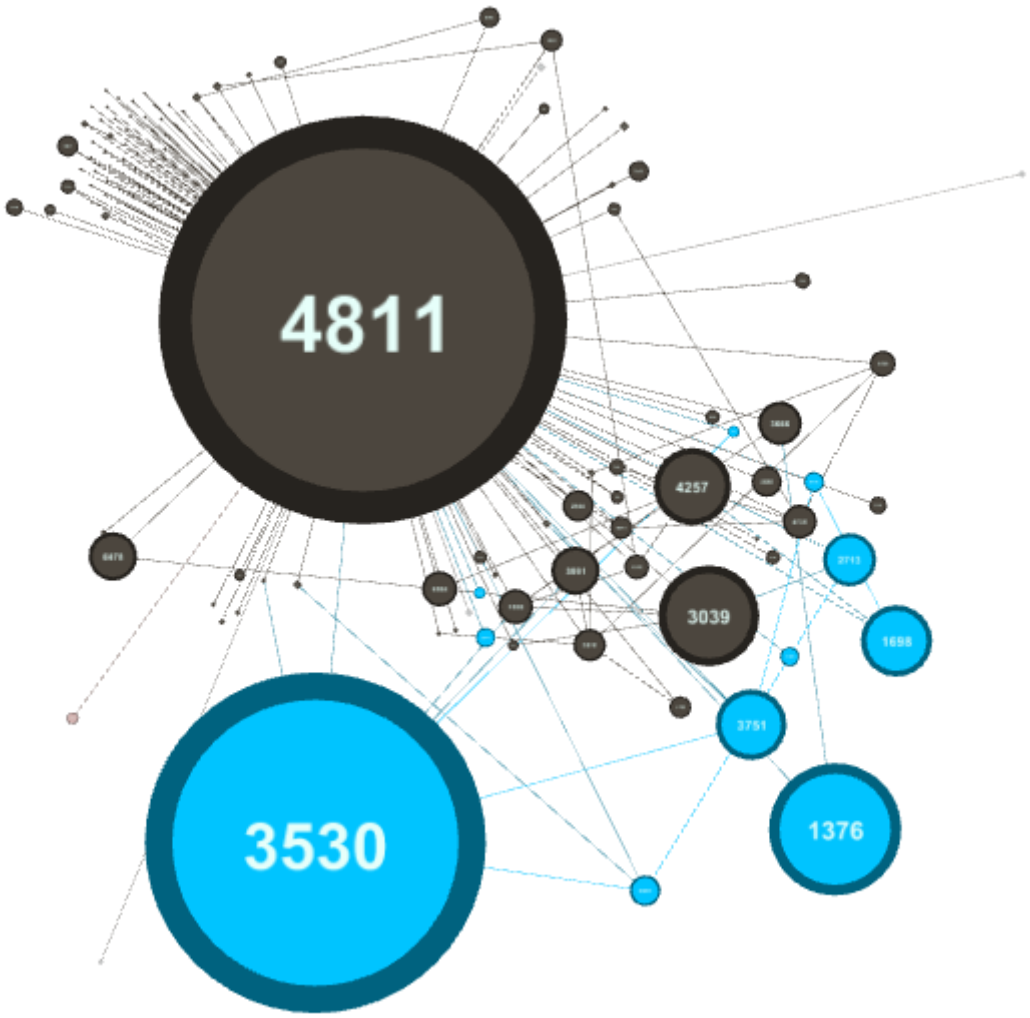


Figure 14: PageRank of Node 4811, Partitioned by Modularity

TABLE 7: TOP FIVE PAGERANK RESULTS FOR LASTFM USERS

NodeId	Modularity Class	PageRank	Degree
4811	11	0.003271	113
3530	12	0.002731	175
1376	12	0.001067	57
3039	11	0.00082	44
4257	12	0.000631	27

Following the PageRank algorithm the Betweenness Centrality was computed. Betweenness centrality measures centrality based on the shortest path. Firstly, the nodes with the highest Betweenness Centrality scores were visualized. For this the color range was changed from pink to green with green being the nodes with the highest scores. Those same nodes were highlighted by changing the size of them. Figure 15 below displays this image.

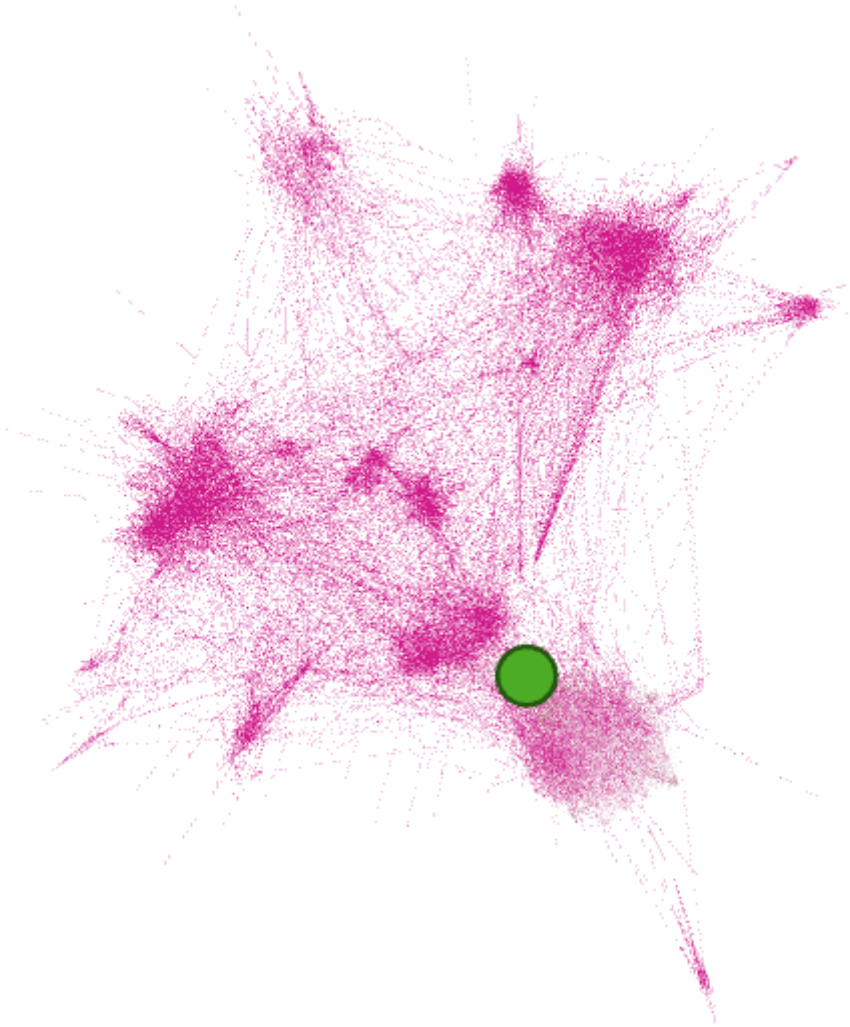


Figure 15: Betweenness Centrality Scores for the LastFM Users Dataset; Green Points Have Higher Scores

As with the previous algorithms the Data Table was analyzed to verify the nodes with the most significant scores. In this case the top node was 7237 which, as stated above, belongs to class 7. Node 7237 had been previously analyzed in regards to the degree and with this new measure it is clear that this node is an outlier, the betweenness centrality for it was 15368.46 whereas the next largest had a score of 329.82. Whether it is an error in the data or this truly is an exceptional node is uncertain however given the influence it has had throughout this work a second node was visualized alongside it. The second most significant node in this category was node 3597 which, as previously stated, had a score of 329.82. Due to the abnormality of node 7237 its influence is apparent in the previous image. Having such an anomaly detracts from graphical analysis so Figure 16 below provides a representation of this outlier alongside a more typical node. The color scheme was maintained to better exhibit the predominant nodes and Table 8 provides straightforward statistics for these nodes.

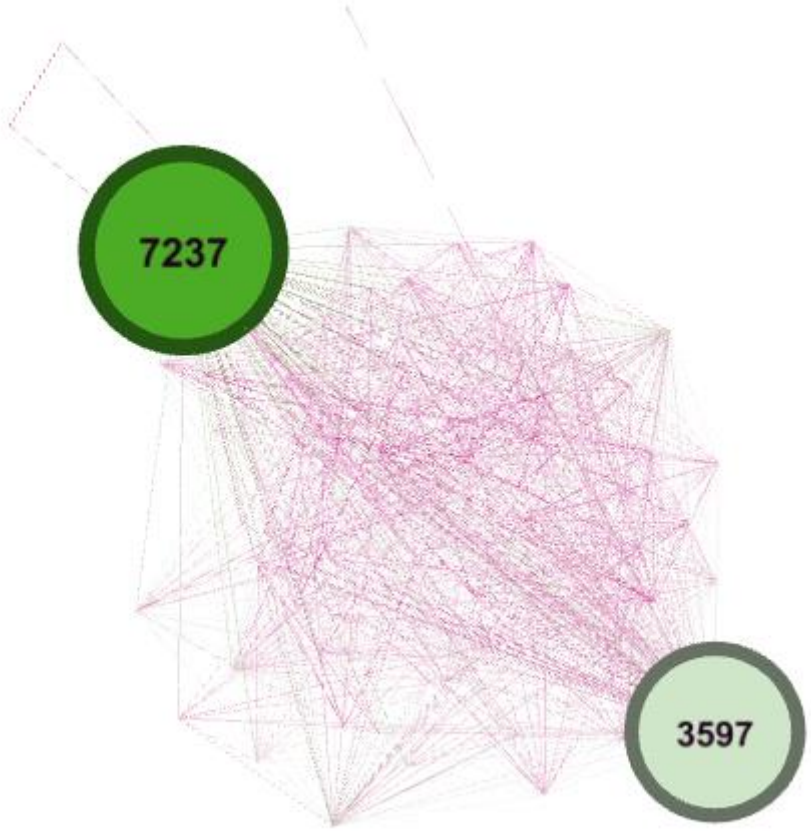


Figure 16: Betweenness Centrality of Nodes 7237 and 3597

TABLE 8: TOP FIVE BETWEENNESS CENTRALITY RESULTS FOR LASTFM USERS

NodeId	Modularity Class	Betweenness Centrality	Degree
7237	7	15368.46	216
3597	7	329.82	124
2083	7	319.10	90
3240	7	305.20	102
6891	7	218.68	74

III. Discussion

Given that there was not much information available for this dataset it is difficult to create significant conclusions. This dataset is a social network thus it is assumed that the nodes with higher degrees have more relationships. In terms of social networks it is also assumed that these nodes, or users, are more popular and influential. With more data it might be possible to provide a better interpretation but from the information that was analyzed one can assume that the user that corresponds to node 7237 is very highly connected.

Another interesting conclusion that may be better supported with more data would be analyzing location details. This dataset is explicitly derived from Asian users therefore understanding the country of origin for these users may be worth examining. As of now it is unclear whether the classes described in this work are location based, such as countries or metropolitan areas, genre based or possibly relating to specific artists. With further knowledge

of this and the relationships presented in this work it would be easy to create more appropriate recommendations to users whether it be musically or socially.

Dataset 3: Diseasome Biological Network Analysis

I. Description of the Dataset

The diseasome dataset was taken from the Gephi website. It contains information about a biological network of a wide variety of human diseases and their associated genes. The description with the dataset includes the following: “Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules.

There are 1419 nodes and 2738 edges. In addition to the labels of the diseases and the genes, there are also labels for two different sets of categories. The first identifies the nodes as either diseases or genes, and the second classifies the nodes by their biological function, system, or location. The labels in this category include cardiovascular, neurological, metabolic, immunological, cancer, muscular, skeletal, respiratory, etc.

II. Results

Upon loading the dataset into Gephi, the dataset was first organized by using the ForceAtlas 2 algorithm. There are a number of groups that immediately become apparent. The distribution of the data is shown below in Figure 17.

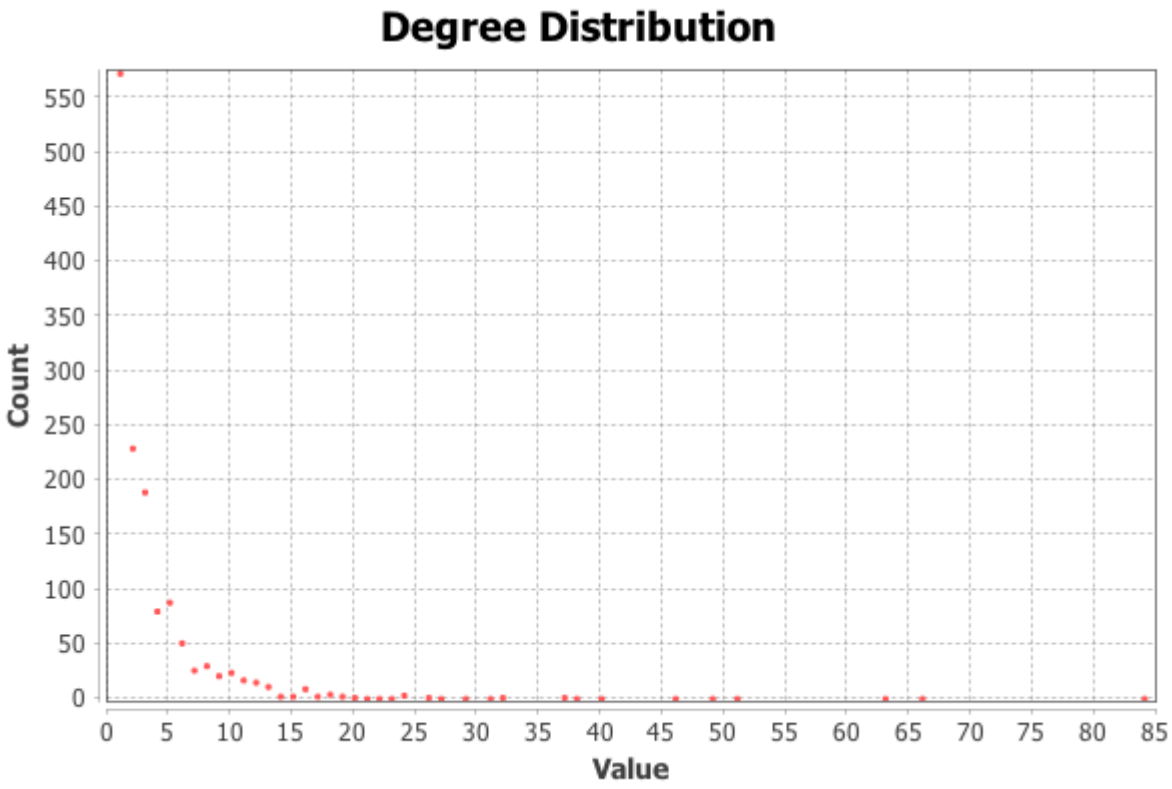


Figure 17: Degree Distribution of Diseasome Dataset

Following this, modularity was run on the dataset. Eight classes were obtained. The node distribution is shown below in Figure 18. Class 6 had the most nodes with 30.73%. Class 1 and 3 followed behind, with 21.42% and 17.69% respectively. The rest of the values are shown in Table 9.

Figure 18: Size Distribution of Modularity Classes in Diseasome Dataset

TABLE 9: MODULARITY CLASS DESCRIPTIONS

Modularity Class (Color)	% of nodes within community
6 (Light Blue)	30.73
1 (Yellow)	21.42
3 (Green)	17.69
7 (Pink)	11.98
4 (Purple)	7.12
0 (Red)	5.36
2 (Grey)	3.81
5 (Blue)	1.90

The visualization was then partitioned by modularity. The resulting figure (Figure 18) is shown below. Class 3, in light blue, has the most nodes and has many different branches as well. Many of these nodes are classified as cancer. This makes sense as there are certainly many different types of cancer as well. This will be explored more in other figures.

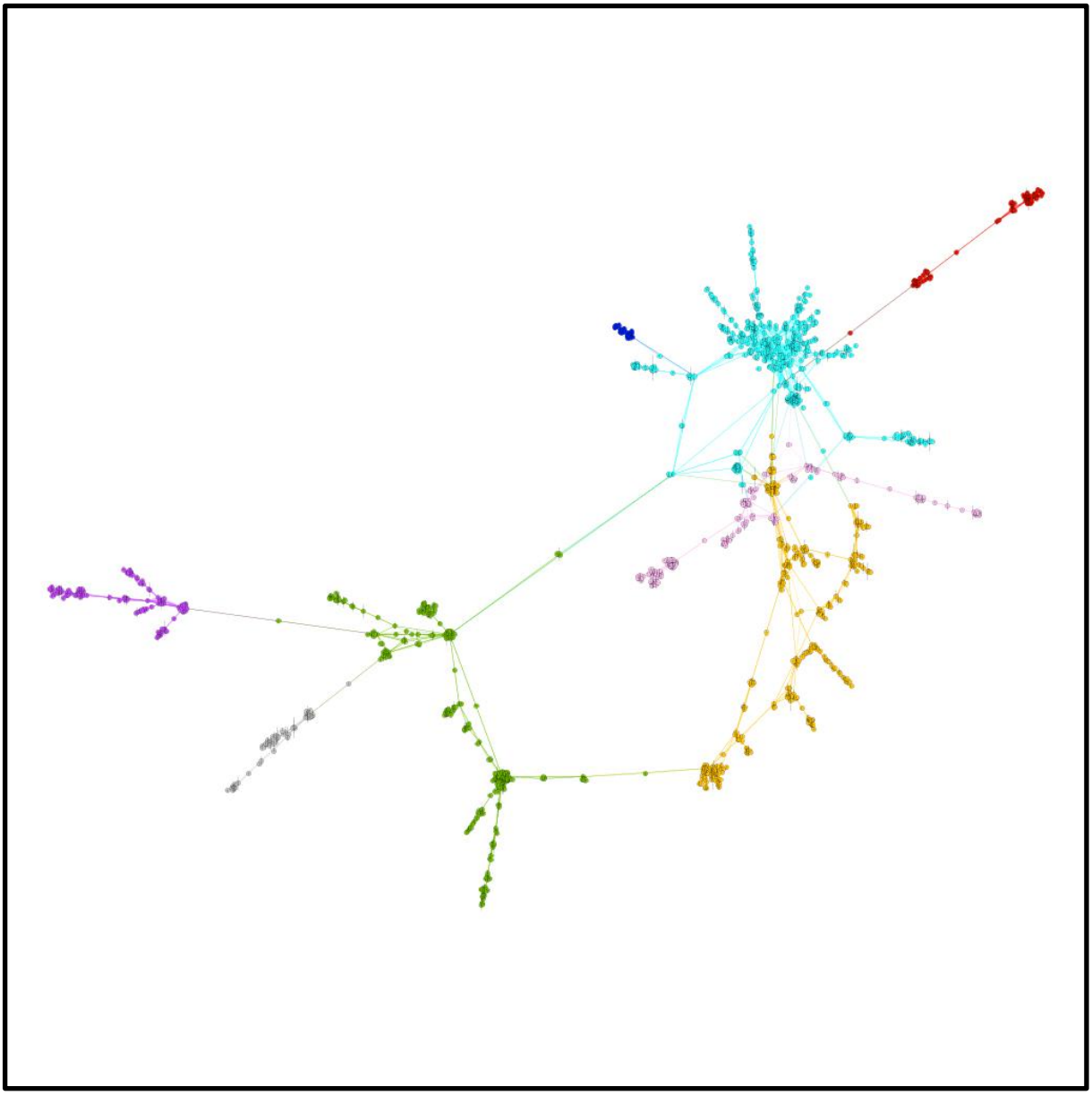


Figure 18: Diseasome Dataset Distribution with ForceAtlas2 Algorithm, Partitioned by Modularity

The above figure shows us the relationships between modularity classes, but it does not tell us too much without any labels. Thus, I added labels. However, with the previous algorithm, ForceAtlas2, the labels were not able to be read, so I used the Fruchterman Reingold algorithm. The Label Adjust algorithm was also used. Next, using the ranking feature, node size was changed to be based on ranking. Labels were also set to be scaled in the same manner. Additionally, to make the visualization easier to comprehend, nodes with a ranking under the unit of 5 were removed using the degree range filter. The result is Figure 19 below.

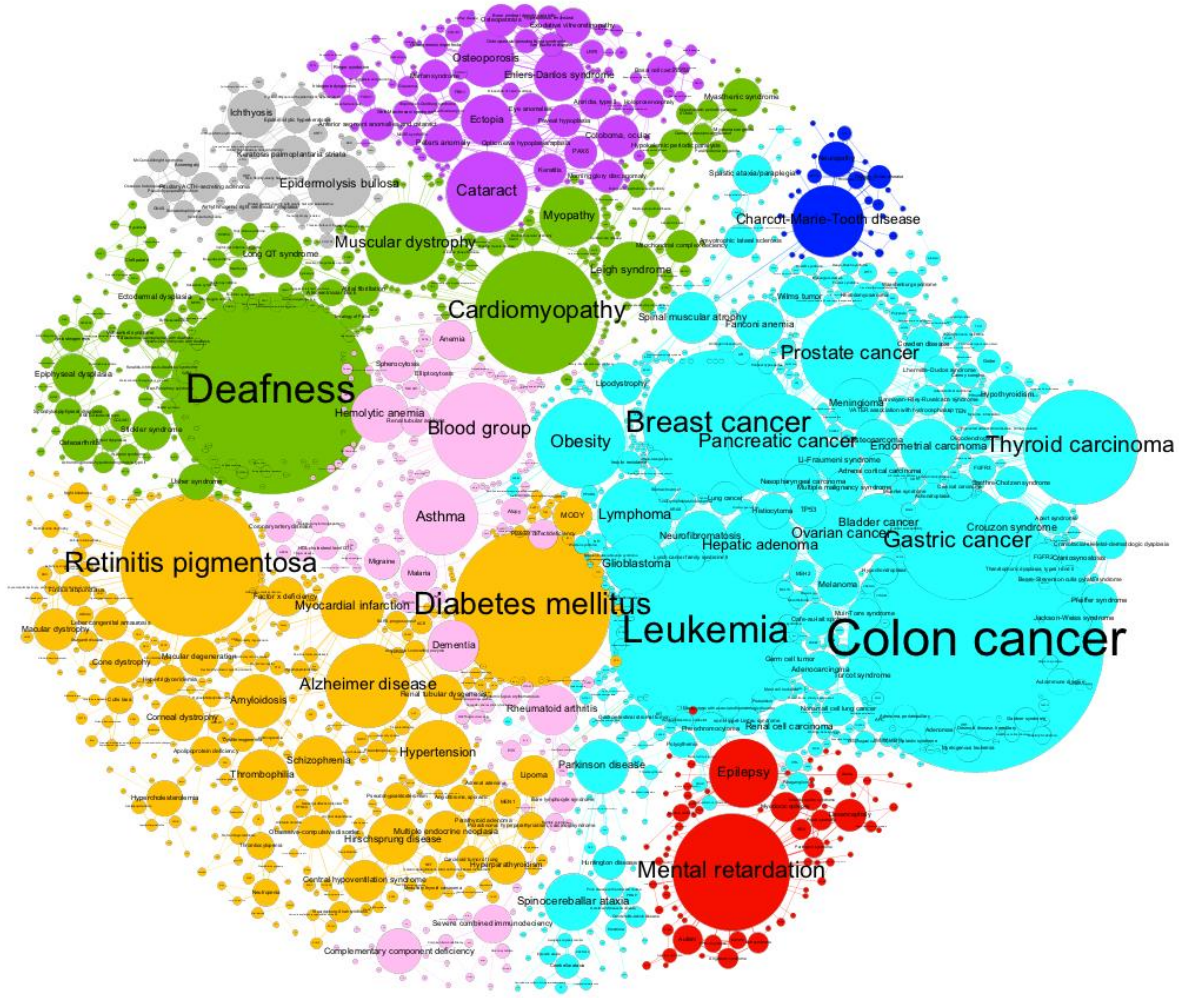


Figure 19: Diseasome Dataset Distribution with Fruchterman Reingold Algorithm, Partitioned by Modularity

The same colors are used in this figure as in the previous figure. The high number of nodes in Modularity Class 6 becomes immediately apparent. The many different forms of cancer also come to attention immediately. Lastly, it is interesting to see other factors such as obesity also appear in class 6.

I mentioned previously that the dataset came with a number of labels. I wanted to compare these labels to the above figure, partitioned by modularity. Hence, the same figure above has now been recolored with the original labels provided by the dataset. The new visualization along with a table with the labels are below. Every category with at least 2% nodes has a color on the visualization; all other categories are grey.

TABLE 10: ORIGINAL CATEGORY LABELS AND CORRESPONDING DETAILS

Category (Color)	**% of nodes within community**
Gene (Orange)	63.64
Cancer (Light Blue)	6.2
Neurological (Light Green)	3.88
Multiple Categories (Red)	3.81
Ophthalmological (Blue)	2.75
Hematological (Red)	2.4
Metabolic (Pink)	2.26
Other (Grey)	15.06

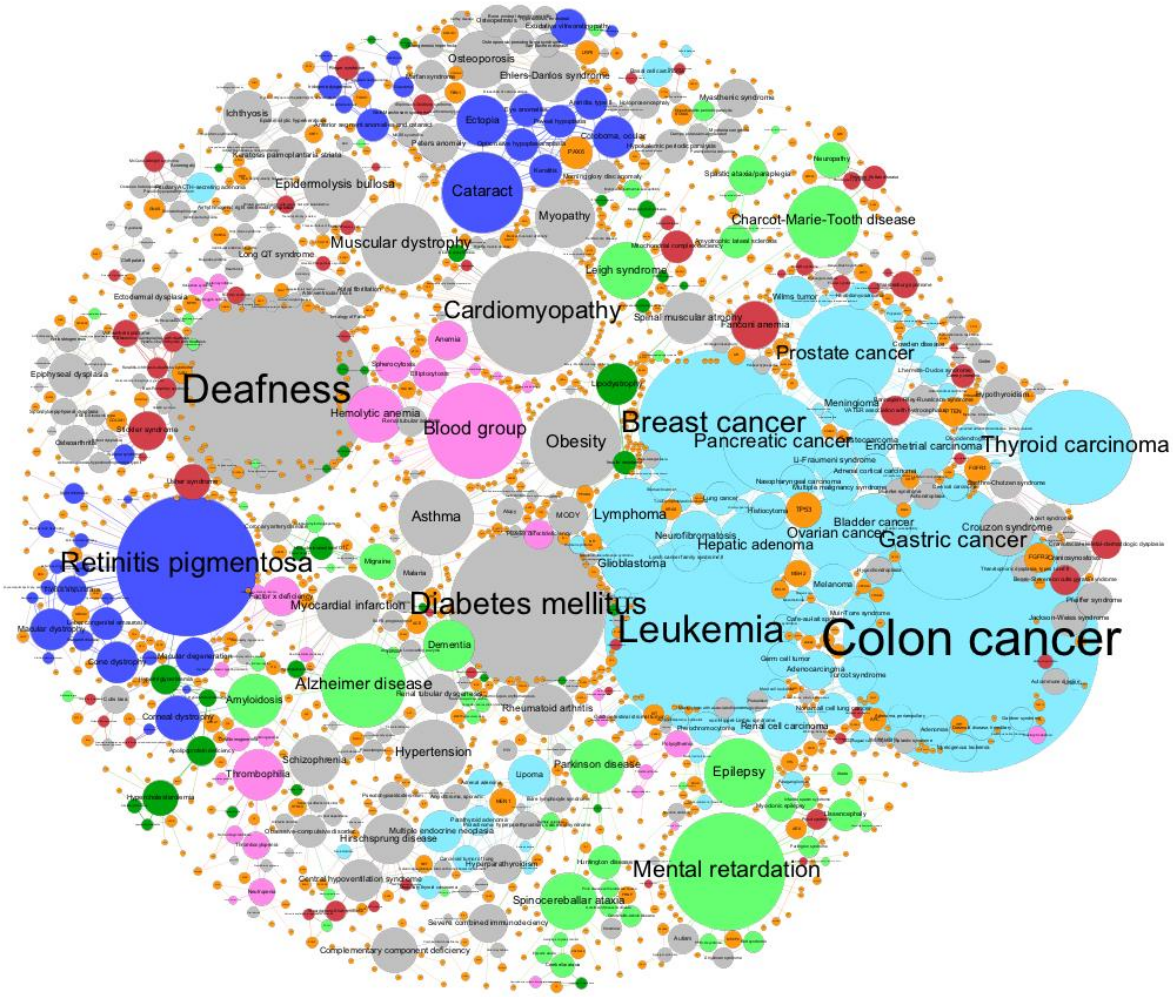


Figure 20: Diseasome Dataset Distribution with Fruchterman Reingold Algorithm, Partitioned by Original Label

As suggested earlier, there is a high correlation between modularity class 6 and the cancer category. Also of note is the distribution of gene nodes and hematological nodes throughout the visualization. The genes may be different, but we see that there are different genes correlated with all diseases. Like genetics, we see that blood also plays a critical role in a variety of functions in the body.

Next, the PageRank algorithm was utilized. Interestingly, the Leukemia node had the highest PageRank with a score of .0109, despite the colon cancer node having the highest ranking by a significant margin. The Ego Network function was used with three levels of connection to the Leukemia node. I then used the Noverlap algorithm to generate this visualization. The resulting visualization (Figure 21) follows.

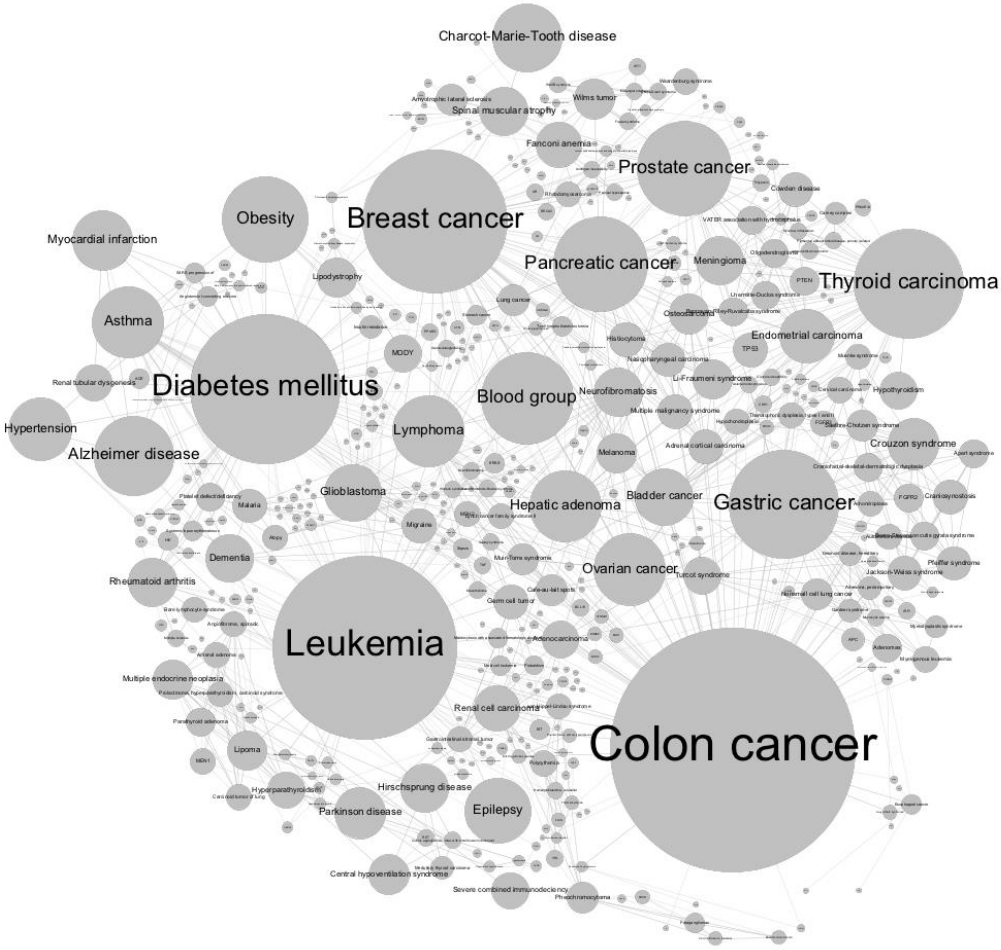


Figure 21: Egocentric Network of the Node with the Highest Degree (Leukemia)

On the same visualization, I added the previously used colors to partition by modularity once again. These changes are shown in Figure 22.

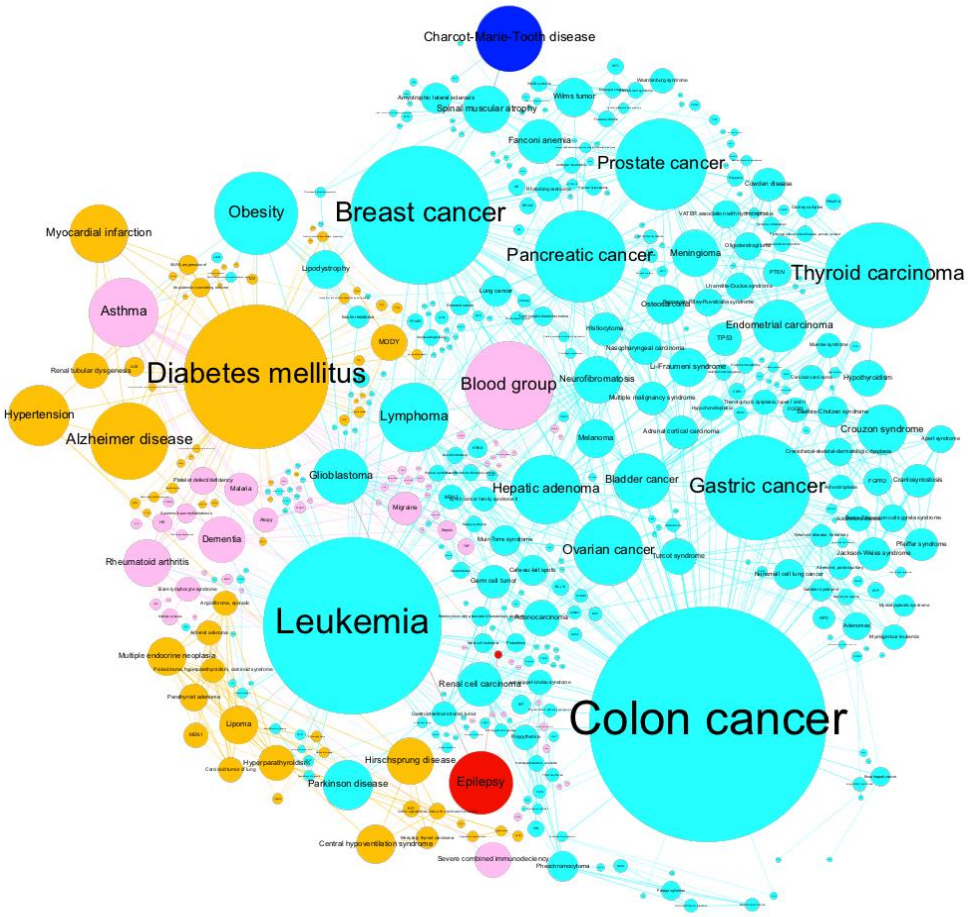


Figure 22: Egocentric Network of the Node with the Highest Degree (Leukemia), Partitioned by Modularity

III. Discussion

I found this dataset to be very interesting. Scientists have learned so much about genetics and their connection to diseases in recent years, but there is still so much we do not know. Biological networks analyzed with tools like Gephi can definitely provide more insight into the connections between genes and diseases.

In this analysis, the data was examined on a large scale to see general trends, but more specific analysis of individual mutations and their relationships with different diseases would certainly be possible and potentially extremely beneficial.

It was observed that the cancer nodes had high rankings and were mostly all found in the same modularity community. It was also observed that blood and genetics are related to many different diseases found in all of the modularity classes. Lastly, it was also interesting to see that Leukemia had some relationship with nearly all other cancer nodes, as well as a number of other nodes in different classes. Some of these seemingly unrelated nodes include obesity, hypertension, diabetes, and Alzheimer's. There are certainly so many genetic factors that underlie all of these diseases, and network analysis is a great tool to study these relationships.

References

[1] B. Rozemberczki and R. Sarkar, "Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. 2020.," LastFM Asia Social Network, 2020. [Online]. Available: https://snap.stanford.edu/data/feather-lastfm-social.html. [Accessed: 27-Jun-2021].

[2] R. Lambiotte, J.-C. Delvenne, M. Barahona Laplacian Dynamics and Multiscale Modular Structure in Networks 2009

[3] Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, in Proceedings of the seventh International Conference on the World Wide Web (WWW1998):107-117

[4] D. Nations, "What Exactly Is Last.fm, Anyway?," *Lifewire*, 07-Mar-2020. [Online]. Available: https://www.lifewire.com/what-is-last-fm-3486395. [Accessed: 29-Jun-2021].