CNT5805-20Fall 0058

Group 2: Network Analysis to understand the disease and gene

Team Member:
Anwesh Praharaj
Shriram Sreedhar

Q 1:

- Summarize your project. Include such items as your **motivation** for selecting your topic. What was the source of your dataset (provide a link if it was obtained on-line)?

Motivation to analyze this dataset to understand the relation between disease and gene also can be able to define the similarity between two disease associated with same gene. We have analyzed Disease-gene association network from the SNAP dataset and gene / disease vocabulary from CTD database. Disease are described as MESH (Medical subject heading) code and gene are described as UniProtIDs code. The original dataset has two categories of nodes Disease, Gene and there are links between both the node category which make it bipartite network. Nodes are segregated into 5663 diseases and 17821 genes.

(Disease – gene) : http://snap.stanford.edu/biodata/datasets/10020/10020-DGMiner.html

(Disease & Gene Vocabulary): This has definition of Disease and genes http://ctdbase.org/downloads/;jsessionid=5FE96872EFBC94CA1BE85AD7A84EBA3D#alldiseases

In this project we have analyzed three network disease-gene , disease-disease and gene-gene.

- Explain your purpose (e.g., inform, persuade, educate, entertain, predict, etc.) for analyzing this network. Use at least one of the terms from the preceding sentence for your purpose.

The purpose of our dataset was to understand about the relationships between genetic diseases segregated by disease and gene network. We decomposed the disease-gene network into disease and gene network by priority node but not by biological functional area. This analysis will educate us how to illustrate the topology of a biological network.

- List and explain the research questions of your project

Below are the list of research question and the statistics measurement used to describe the questions in detail. Detailed are explained in statistical analysis report.

Research Question	Measurement	Network
1. What is the depth/degree of the relationship between participating nodes?	Degree Distribution	Disease-Gene
		Disease-Disease
		Gene-Gene
2. Are there identifiable clusters based on which the genes/Disease can be grouped? 3. How to find similarity between gene , Disease and its neighbor?	Leiden Algorithm & HITS	Disease-Gene
	Clustering Co-efficient & Modularity	Disease-Disease
		Gene-Gene
4. Are there identifiable hubs?	HITS (Authorities & Hubs)	Disease-Gene
		Disease-Disease
		Gene-Gene
5. What are the strongest nodes in terms of connectivity and what are the weakest nodes? 6. What is the betweenness of the nodes?	Centrality Measurement	Disease-Gene
		Disease-Disease
		Gene-Gene

Q 2:

- How many nodes and edges are in your graph, is it directed or undirected, weighed or unweighted?

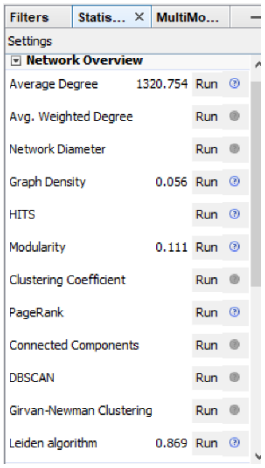
The initial Bipartite graph was an undirected, unweighted having 23486 nodes and 15509619 edges. We decompose the network into two separate monopartite networks. Disease-Disease network was an undirected, unweighted having 910 nodes and 18896 links. Gene-Gene network was an undirected, unweighted having 3594 nodes and 35700 edges.

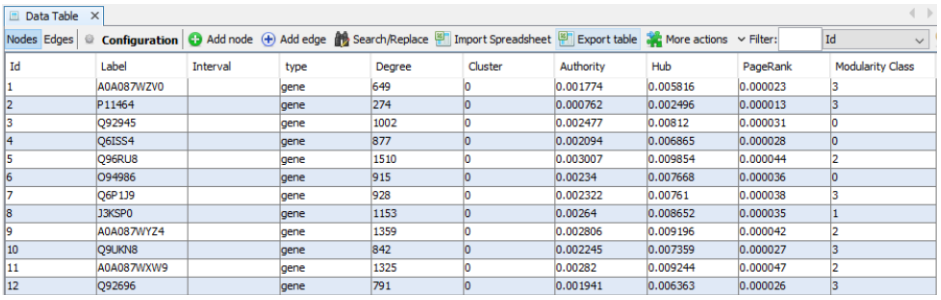
Network	Nodes	Links	Graph Type		Weighted
Disease-Gene	23486	15509619	Undirected	Bipartite	No
Disease-Disease	910	18896	Undirected	Monopartite	No
Gene-Gene	3594	35700	Undirected	Monopartite	No

- Run all the Statistics, Network Overview in Gephi and show a screenshot of it.

Below are the details of Network Overview and Node (Data Laboratory)

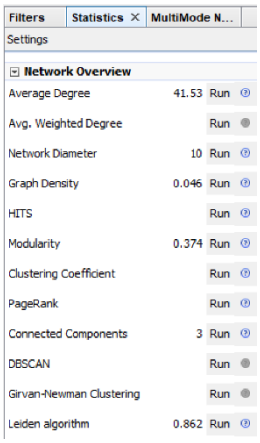
Disease-Gene Network:



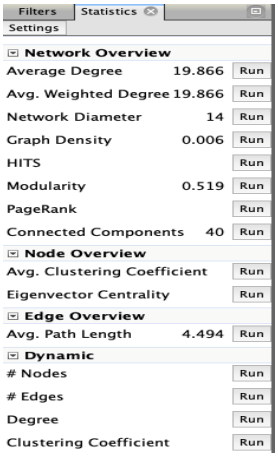


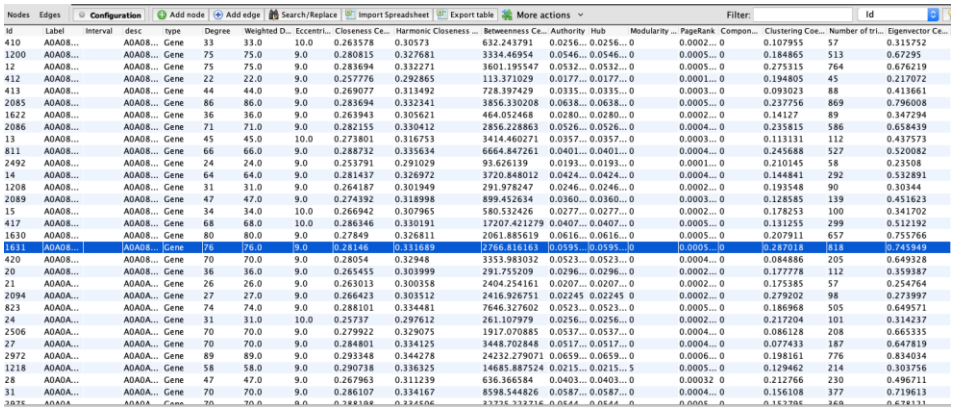
Id	Label	Interval	type	Degree	Cluster	Authority	Hub	PageRank	Modularity Class
1	A0A087WZV0		gene	649	0	0.001774	0.005816	0.000023	3
2	P11464		gene	274	0	0.000762	0.002496	0.000013	3
3	Q92945		gene	1002	0	0.002477	0.00812	0.000031	0
4	Q6ISS4		gene	877	0	0.002094	0.006865	0.000028	0
5	Q96RU8		gene	1510	0	0.003007	0.009854	0.000044	2
6	O94986		gene	915	0	0.00234	0.007668	0.000036	0
7	Q6P1J9		gene	928	0	0.002322	0.00761	0.000038	3
8	J3KSP0		gene	1153	0	0.00264	0.008652	0.000035	1
9	A0A087WYZ4		gene	1359	0	0.002806	0.009196	0.000042	2
10	Q9UKN8		gene	842	0	0.002245	0.007359	0.000027	3
11	A0A087WXW9		gene	1325	0	0.00282	0.009244	0.000047	2
12	Q92696		gene	791	0	0.001941	0.006363	0.000026	3

Disease-Disease Network:



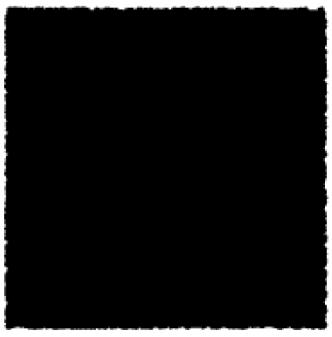
Gene-Gene Network:





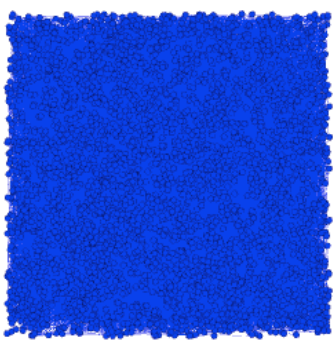
- What can you ascertain from the initial graph you see? Include a diagram of the initial graph in your report

Disease – Gene Network



Disease – Disease Network

Gene – Gene Network



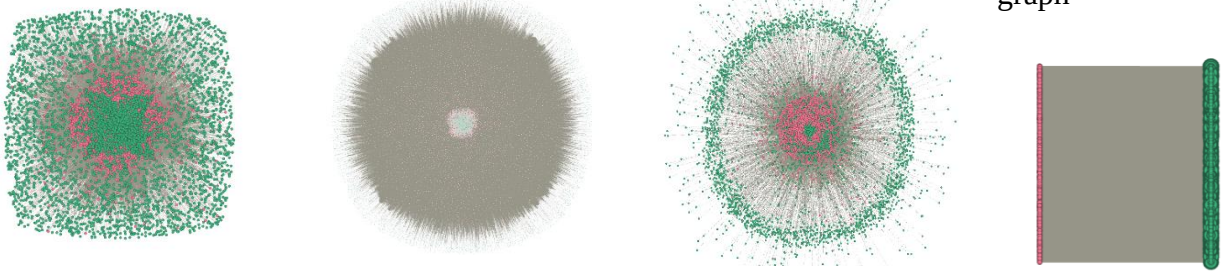
All the three-network graph are undirected network. In Disease-Gene network each node represents either a gene or disease and link signifies the relation between a disease with gene. In Disease-Disease network each node represents disease and links show the relation between two disease. Similarly, for Gene-Gene network each node represents gene and link significance the relation between two gene. These layouts are random layout and here all connection between the nodes are very unclear and cannot be able to identify important or influenced node of the network as all the nodes are of same size, color. The graph won't move in time that means the graph is not dynamic in nature. By looking at the node and edges we can assert that the network is large one. Also, it is very hard to predict any hub or cluster at this point by looking at these initial graphs.

Q 3:

- Show a small screenshot of each one and briefly explain what changed and why?
- At this point, which layout seems most useful and why?

Disease – Gene Network

Fruchterman Reingold Force-Atlas 2 Yifan-hu Event-graph



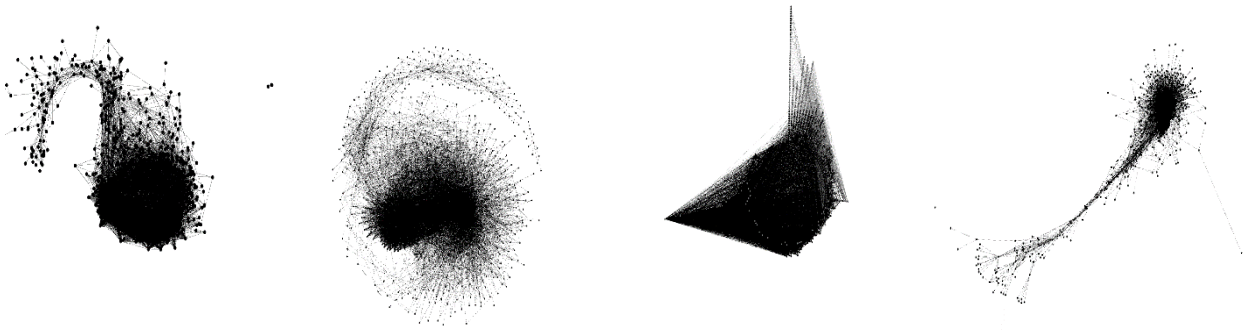
Event-Graph is a plugin in gephi which usually used to represent bipartite network. In case of Disease-Gene network because of more than fifteen million links it is difficult to show the relation between the nodes. Rest all three layouts are force direct graph where because of Higher repulsive strength, the nodes which are not closely related to each other are away from each other. It is very difficult to identify the hub and clusters as the network layout was very dense for Force-Atlas 2 layout. Fruchterman Reingold layout is another forced directed network in which use same attraction and repulsion mechanism to arrange the nodes in network. But the layout is quite different from Force-Atlas 2 layout that is sphere shape. In comparison to Force Atlas and Fruchterman Reingold , YifanHu layout provide much faster result. This is because this layout applies repulsion and attraction in neighborhood rather than entire network.

As a final layout Yifan-hu was considered because it separated the node away in such a way that we can clearly identify the hubs, disease and gene in the network which was easy to interpret network topology while comparing other network layouts. Also, this network is faster and better

precision then other forced directed network because it has lower computational over burden in local machine

Disease- Disease Network

OpenOrd Fruchterman Reingold Radial Axis Yifan-hu

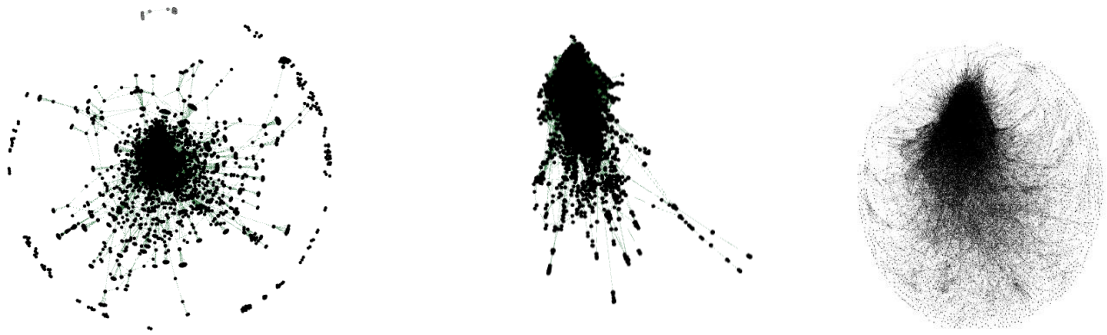


Here four different layouts have used to interpret this network. OpenOrd method is always good in distinguishing the clusters but in this layout it is very difficult to distinguish between communities. Radial Axis layout grouped the nodes (depending on the grouping criteria) and radiated outwards from the central circle. This layout is very good to identify homophily if there are natural grouping between the nodes of the network. In this network radial axis made a giant hub in middle again very difficult to interpret. Yifan-hu was the most fastest forced directed layout which repulse the less connected nodes away from hub. In case of Disease-Disease network Yifan-hu was not able to show the different communities clearly.

As a final layout, Fruchterman Reingold layout was considered as final layout as it created a sphere shape network in which it was easy to identify the clusters , hubs and communities.

Gene-Gene Network

Yifan-hu Force Atlas 2 Fruchterman Reingold



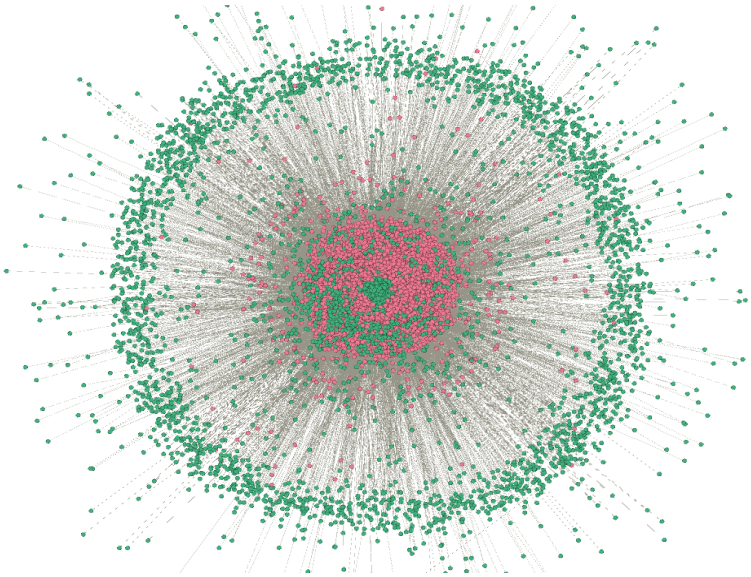
Here three different layouts was used to visualize the Gene-gene network. Yifan Hu layout reduced the complexity of the network because this network had so many nodes and this algorithm organized and visualized the network depicting the relationship between genes in a better way. Force Atlas 2 improved the quality of the network because of repulsion of the nodes as indicated by the Barnes-Hut-Calculation and this therefore reduced the complexity of the algorithm and thus made it easier to visualize the relationship between multiple genes in this network. Frutcherman-Reginold layout represented the relationship between genes as mass particles. This algorithm tried to lower the energy of the physical system which made it easier to view the relationships among genes.

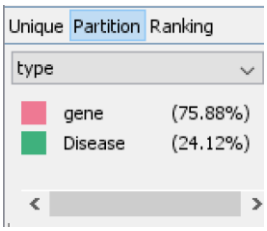
Yifan Hu is most useful layout because it brings together good parts of force directed algorithms multilevel algorithms in order to reduce the complexity of the algorithm. This algorithm worked very well with large networks.

Q4:

- Show a screenshot of this action.
- What have you learned from your new diagram?

Disease – Gene Network

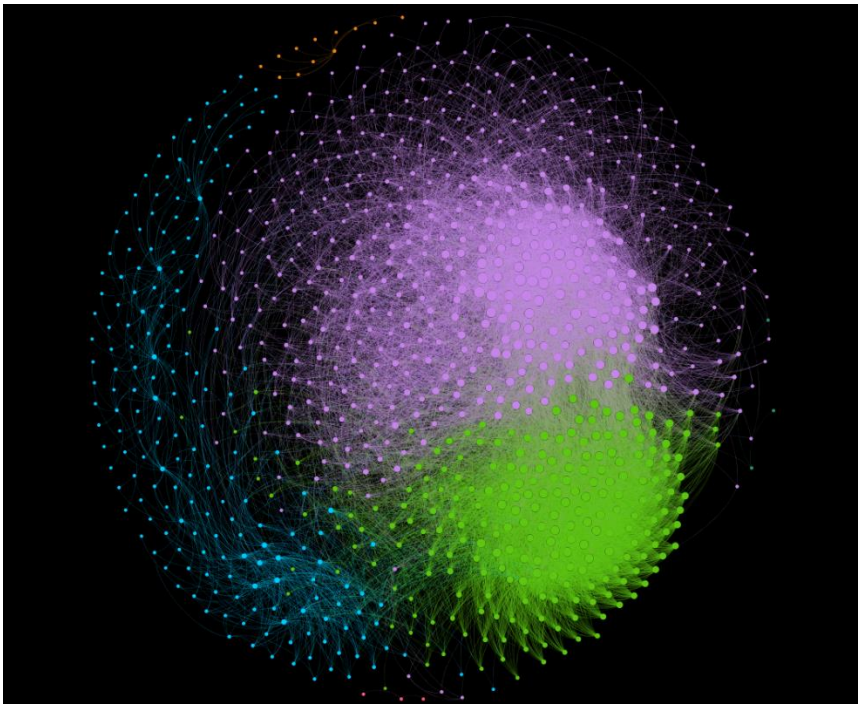


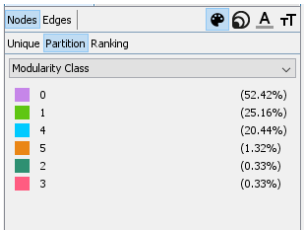


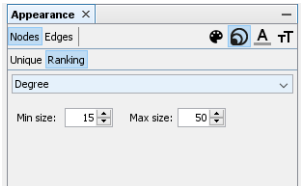
As Disease-Gene network has fifteen million links, it was very difficult to add label and size to the nodes as per degree size. Here the network was portioned by type i.e gene and diseases. The nodes in green color showed all diseases and pink color node as genes.

This layout representation was very useful to identify the gene and disease in the network. Most of the diseases are in outer periphery , at the same time most of the genes are at middle creating a giant network.

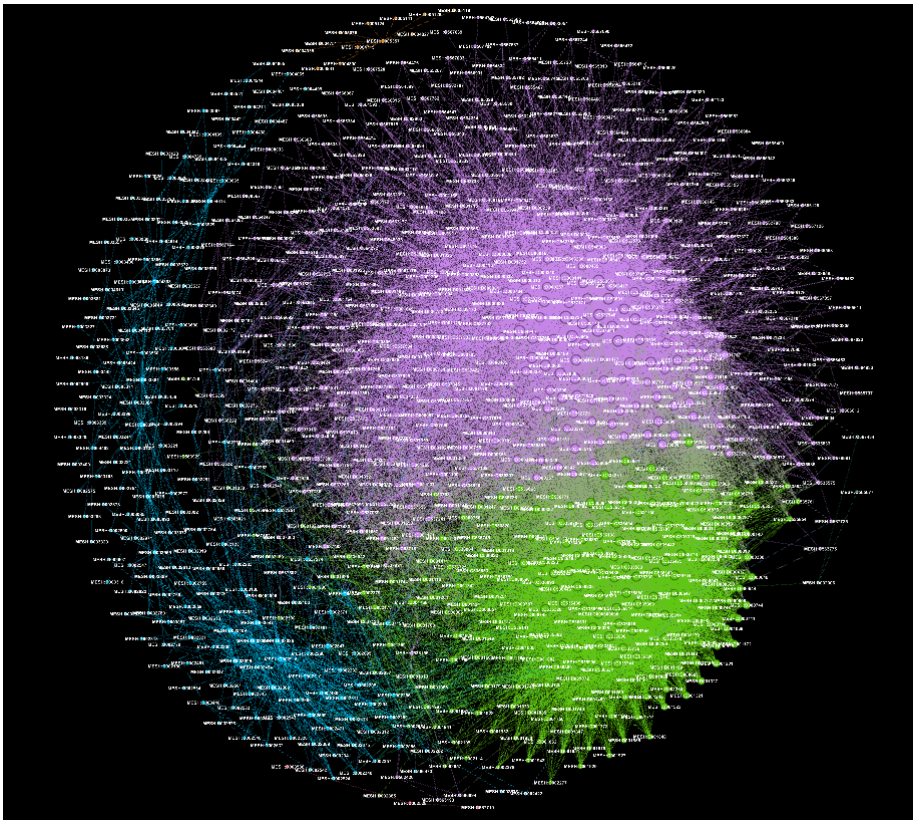
Disease – Disease Network: -







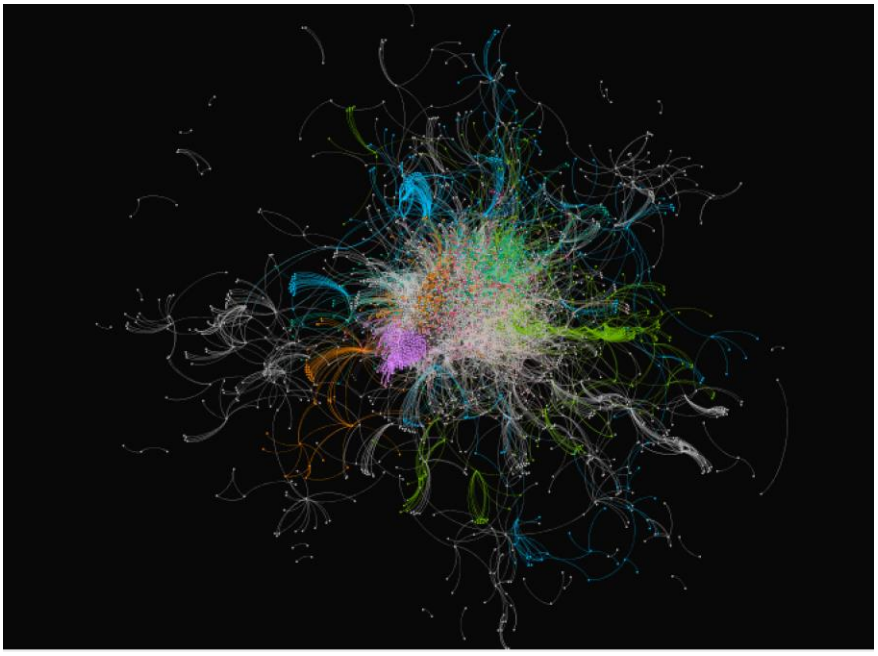
Applied partition by modularity group and changed the Node size in a range between fifteen – fifty. This graphical represenation gave a better visualization of different communities in this network.



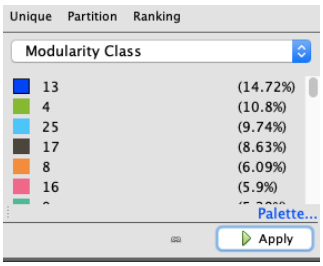
Here applied label name in each node which represented the disease name. Because of huge data it is difficult to read the names in graph.

This network visualization was very useful to detect communities in the Disease-Disease network.

Gene – Gene Network:



-



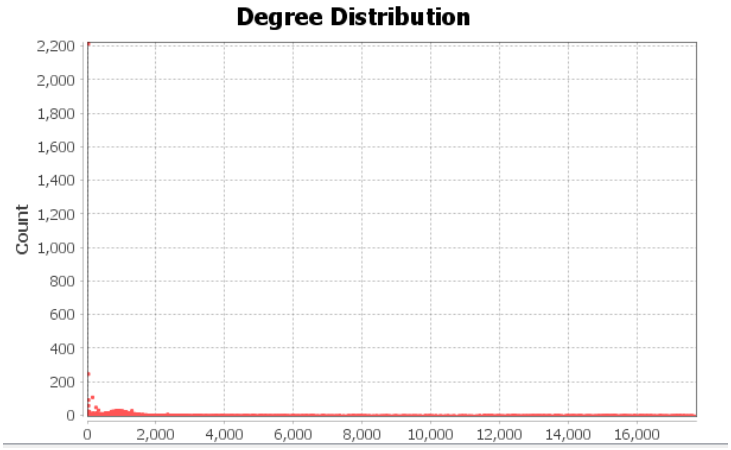
Applied color code by modularity class which helped to identify different communities in gene-gene network.

Q5.

- Explain what statistical results you found. Show graphs, plots, and numbers generated by your software and explain what impact it had on your analysis.

5.1 Disease-Gene Network: -

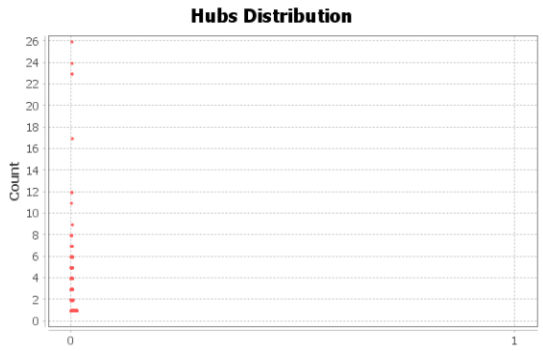
Degree Distribution:



Degree of distribution can be analyzed by Average degree of Network (Scatter Plot) . Average degree is nothing but average number of unweighted connections across a network. In the graph X axis shows the degree and Y axis shows number of Nodes. We can conclude that highest number of Nodes are in degree 1. Degree distribution follow nearly power law distribution and scale free network.

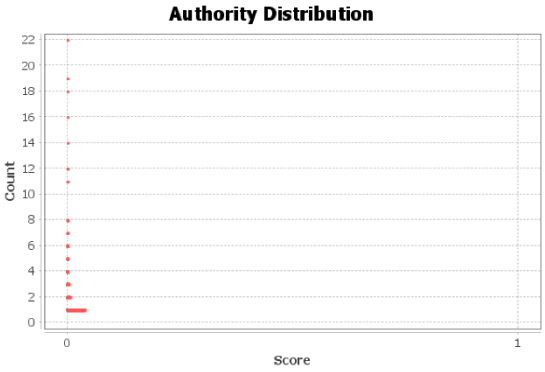
HITS Analysis:

For undirected network HITS algorithm provide hub and authority score for each node. This analysis provides the nodes which are more likely to create hub in this network. Below are the top 10 Disease and gene from Bipartite network. This helped us ranking the nodes (Disease and Gene) which used for filtering purpose.



Disease	Hub Score
MESH:D009336	0.011289
MESH:D056486	0.011289
MESH:D007674	0.011282
MESH:D007249	0.011282
MESH:D011297	0.01128
MESH:D015431	0.011275
MESH:D007859	0.011273
MESH:D008569	0.011273
MESH:D005234	0.011272
MESH:D006973	0.011272

Gene	Hub Score
P01375	0.011274
P42574	0.011268
I6LPK7	0.011246
B5MC21	0.01124
P10415	0.011237
P01100	0.011216
A2A2V4	0.011201
P28482	0.011198
P04040	0.011196
P05412	0.011195



Disease	Authority
MESH:D009336	0.036999
MESH:D056486	0.036999
MESH:D007249	0.036977
MESH:D007674	0.036975
MESH:D011297	0.03697
MESH:D015431	0.036953
MESH:D008569	0.036947
MESH:D007859	0.036946
MESH:D005234	0.036944
MESH:D006973	0.036944

Gene	Authority
P01375	0.00344
P42574	0.003438
I6LPK7	0.003431
B5MC21	0.003429
P10415	0.003429
P01100	0.003422
A2A2V4	0.003418
P28482	0.003416
P04040	0.003416
P05412	0.003416

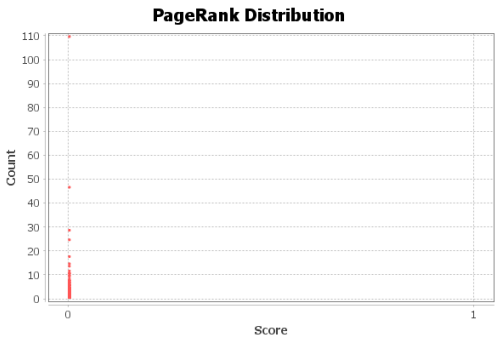
PageRank distribution:

PageRank algorithm helped identifying influencer in a network. node that received more links is likely to be an important node in the network. This is used to rank the important nodes in this bipartite network.

Below are the top 10 disease and gene from page rank score. As this is bipartite network page rank is very low for each nodes.

Epsilon = 0.001
Probability = 0.85

Results:



Disease	PageRank
MESH:D007249	0.000575
MESH:D009336	0.000566
MESH:D056486	0.000565
MESH:D008569	0.000556
MESH:D001943	0.000553
MESH:D011297	0.000553
MESH:D002471	0.000551
MESH:D007859	0.00055
MESH:D006965	0.00055
MESH:D007674	0.000549

Gene	PageRank
P01375	0.00013
P02458	0.000116
P01584	0.000115
P42574	0.000112
B5MC21	0.000109
A0A087WT22	0.000103
P04040	0.000101
P08684	0.000101
P01579	0.000099
P10145	0.000098

Liden Algorithm:

Regular clustering coefficient algorithm is not applicable for bipartite network as nodes can not make triangle in the network. For this network Liden algorithm has used as this consider rectangle

as cluster. In this network there were 4030 cluster. As this algorithm is new to us , did not use this statistics while ranking the important nodes.

Configuration

Algorithm	Leiden
Quality Function	Constant Potts Model (CPM)
Resolution	0.01
Number of iterations	10
Number of restarts	1
Random seed	0

Results

Quality	0.8694155904150013
Number of clusters	4030

Disease	Cluster
MESH:D005401	4013
MESH:D029503	4010
MESH:D007645	3998
MESH:C536990	3996
MESH:C537240	3995
MESH:C537081	3994
MESH:C567489	3993
MESH:C535396	3992
MESH:C567383	3991
MESH:D049068	3990

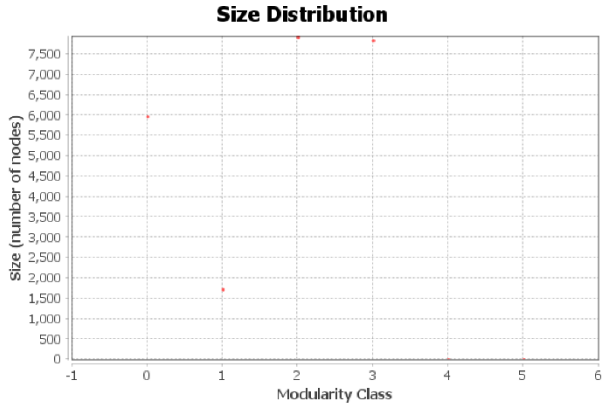
Gene	Cluster
A6NJZ3	4029
H3BQB6	4028
Q8NGF6	4027
Q5T870	4026
Q9H343	4025
Q8N349	4024
Q6UWN8	4023
G3V1C2	4022
Q96RI8	4021
Q9GZY0	4020

Modularity:

Modularity Class determined the number of communities present within a Network using cluster analysis approach. This can be used to measure clustering in a network. There were 6 modularity classes.

Results:

Modularity: 0.111
Modularity with resolution: 0.111
Number of Communities: 6



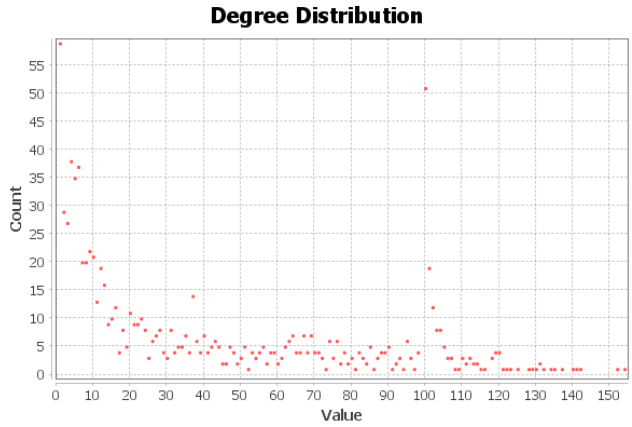
5.2 Disease-Disease Network:

Degree Distribution:

Below graph represent power law distribution which concluded that this network is a scale free distribution. More than 55 diseases have degree value 1 and less number of diseases have more than 135 degree value.

Results:

Average Degree: 41.530



Network Diameter:

this network has diameter value as 10, which mean that it would take maximum 20 steps to traverse the graph between its two most distant points or nodes. Gephi used different centrality measurement to calculate the network diameter.

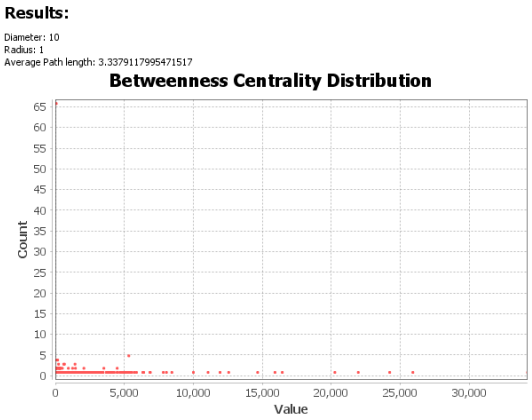
Modularity:

Results:

Modularity: 0.374
Modularity with resolution: 0.374
Number of Communities: 6

For Disease-Disease network gephi divided into 6 communities. Nodes that are highly connected were combined in a common cluster group.

Betweenness Centrality:

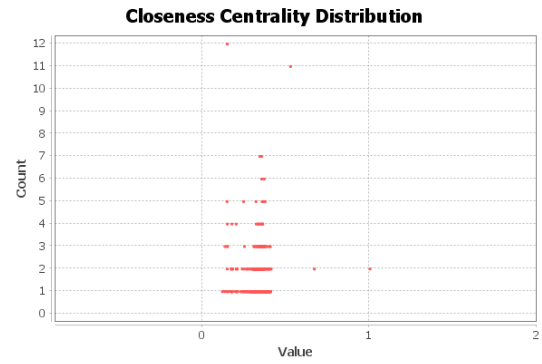


This centrality measurement determined important diseases which appeared most of the times in the shortest path between other diseases. Here these diseases are treated as a gatekeeper or the bridge between two diseases.

Here are the top 10 diseases having higher betweenness centrality.

Disease	Centrality Score
MESH:D003635	34170.789
MESH:C537251	25850.80882
MESH:D002759	24171.68661
MESH:D003109	21892.13352
MESH:D003715	20195.71759
MESH:D003881	16385.57582
MESH:D003095	15858.54343
MESH:D002916	14601.38808
MESH:D003616	12497.57049
MESH:C536141	11874.85334

Closeness centrality:



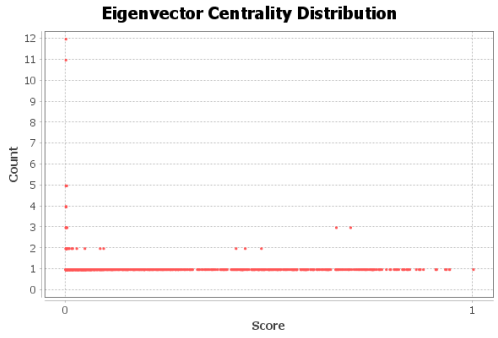
This centrality measurement helped us to find out which all are the diseases are closely connected with other disease. This analysis helped us to find which two diseases are more likely to occur.

Below are the top 10 diseases which are most relevant in this analysis and having lowest score.

Disease	Centrality Score
MESH:D004605	0.114763
MESH:D004769	0.127587
MESH:D003330	0.129584
MESH:D003551	0.129622

MESH:D003554	0.130378
MESH:D003677	0.130378
MESH:D003681	0.130378
MESH:D003428	0.142675
MESH:D003288	0.142949
MESH:D003316	0.142949

Eigenvector centrality:

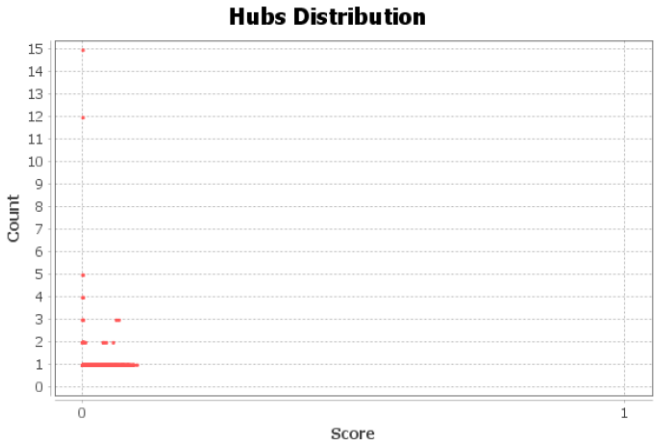


This centrality measurement determined the diseases which hold higher influence in this Network, in other words important diseases in our data set. The centrality score is proportional to the sum of the scores of all nodes multiply by weights which are connected to it. Also, when nodes are highly connected to other nodes with high levels of influence, the result will be a high-level of Eigenvector centrality. Below were the top 10 diseases having highest centrality score. This was found that all these diseases are part of one of the hub in the network.

Disease	Centrality Score
MESH:C538178	1
MESH:C537844	0.941728
MESH:C536357	0.941043
MESH:C537948	0.933664
MESH:C536914	0.931269
MESH:C538011	0.90959
MESH:C536495	0.908401
MESH:C563739	0.9075
MESH:C563614	0.877281
MESH:C536392	0.869671

Hubs Distribution:

This statistic provided a hub score for each node and helped to identify the important diseases which all are part of the hubs. Below are the top 10 diseases having highest hub score in this network.



Disease	Hub Score
MESH:C538178	0.0997
MESH:C536357	0.093241
MESH:C537844	0.093231
MESH:C537948	0.093109
MESH:C536914	0.092866
MESH:C538011	0.090719
MESH:C536495	0.089846
MESH:C563739	0.089798
MESH:C563614	0.08725
MESH:C537268	0.086329

Graph Density:

Graph Density Report

Parameters:

Network Interpretation: undirected

Results:

Density: 0.046

This analysis interpreted roughly how much percentage of nodes are connected in this network. This network has graph density 4% (0.046) which is low density when considered the whole network.

Clustering co-efficient

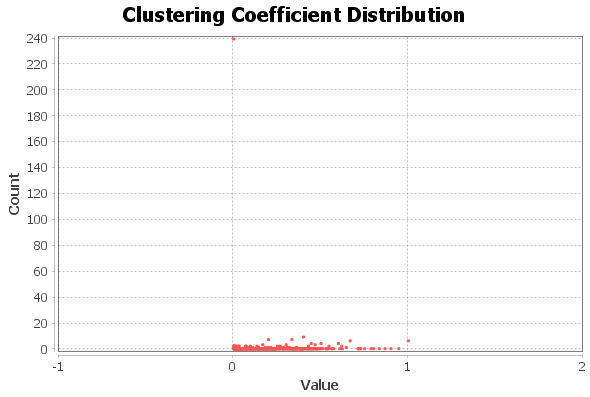
This value is calculated by the number of closed triangles (triplets) relative to the potential number of triangles (triplets) available in the network. In other words, connections between the nodes that form complete triangles. This network has very less clustering coefficient 0.149. Below are the Disease which made perfect cluster as 1. From below graph it can be concluded that there are around 240 diseases having 0 clustering co-efficient.

Parameters:

Network Interpretation: undirected

Results:

Average Clustering Coefficient: 0.149
Total triangles: 71002
The Average Clustering Coefficient is the mean value of individual coefficients.

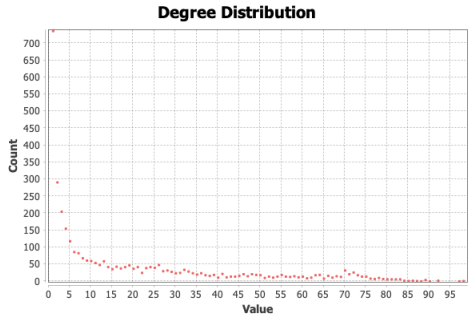


Disease	Clustering Co-efficient
MESH:C564133	1
MESH:C564567	1
MESH:C565193	1
MESH:C565780	1
MESH:C567582	1
MESH:C567502	1

5.3 Gene-Gene Network :

Results:

Average Degree: 19.866



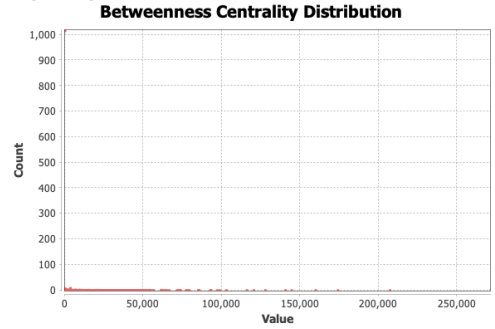
Degree distribution for this network followed power law distribution. So, this was a scale free network. There were more than 700 nodes having degree value around 1.

Parameters:

Network Interpretation: undirected

Results:

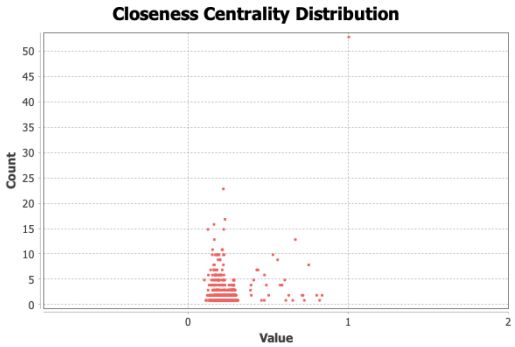
Diameter: 14
Radius: 1
Average Path length: 4.494287543156982



This centrality measurement illustrates the important genes which comes multiple times in the shortest path of other genes. These genes act as bridge between other genes.

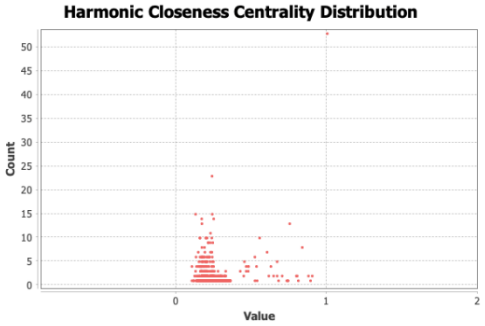
Below are the top 10 Gene in network.

Gene	Centrality Score
A0A087X169	271240.8869
A0A0A0MTL8	207013.8007
A0A087WW00	173887.1367
A0A087X0I6	159696.0108
A0A0A0MR60	144393.4026
A0A0G2JNB1	140452.7872
B1ALF6	127619.1252
P46782	120280.1664
A0PJE2	115897.7833
H9NIL4	102857.6364



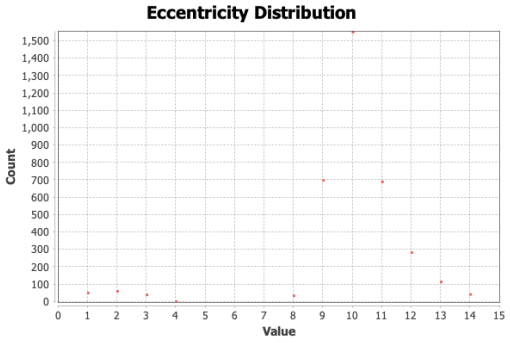
The data that is depicted in this graph was spread out evenly and this makes it easier to visualize the data. Nodes that are more close together have a low closeness score and have the shortest distance to other nodes. Nodes that are less close together on the other hand have a high closeness score and have the longest distance to other nodes. Below are the top 10 genes having low centrality score.

gene	centrality score
Q9NR99	0.097914
Q9H221	0.097914
Q9HCB6	0.097914
Q9GZT9	0.097914
Q9H6X2	0.097914
Q8IZD4	0.108504
Q9NRR4	0.108538
Q8WZA1	0.115643
Q9H9S5	0.115643
P39900	0.120585

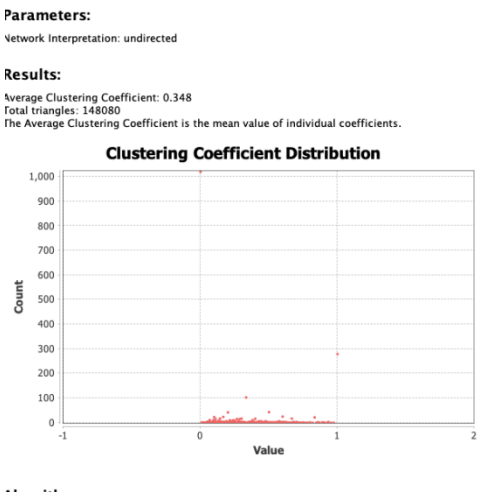


This distribution is also known as valued centrality and was used to solve the problem that the original formula did when it dealt with unconnected graphs. This distribution came originated from the field of social network analysis. Furthermore, the Harmonic Closeness Centrality distribution sums the inverse of a relationship of a node to all other nodes and helps dealing with infinite values. Below are the top 10 genes having Centrality score 1. In total there are more than 50 genes having centrality score as 1.

Gene	Centrality Score
Q03591	1
Q9UHC1	1
F5H6I3	1
Q9NRE1	1
P23409	1
Q96H96	1
H3BS19	1
Q9UJ55	1
P50219	1
P30301	1

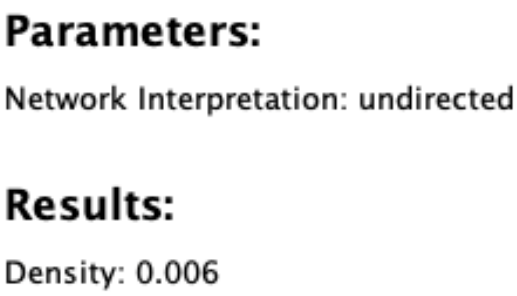


This distribution helps in calculating the length of the longest short path starting from that node. The maximum eccentricity is the diameter of the graph and the minimum graph eccentricity is the radius of the graph



Below are the top 10 genes which created complete triangle having clustering coefficient as 1 and the average clustering co-efficient was 0.348 .

Gene	Clustering co-efficient
Q8N183	1
Q86Y39	1
Q8TB37	1
Q5TEU4	1
Q7Z2E3	1
Q5T2R2	1
Q92556	1
Q6UUV9	1
Q15722	1
Q9NPC1	1



Density of this network was very low.

Q6.

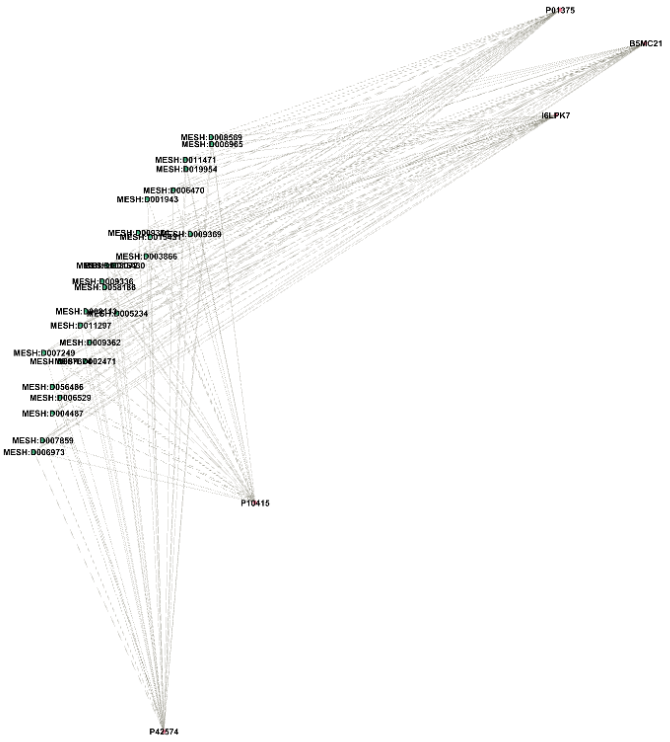
6.1 Disease-Gene Network:

To find top 20 priority gene and disease, filtering was done by Degree, Authority score and PageRank.

Gene	PageRank	Degree	Authority Score
P01375	0.00013	2801	0.00344
P01584	0.000115	2616	0.00339
P42574	0.000112	2754	0.003438
B5MC21	0.000109	2690	0.003429
A0A087WT22	0.000103	2475	0.003397
P04040	0.000101	2582	0.003416
P08684	0.000101	2444	0.003411
P01579	0.000099	2516	0.003411
P10145	0.000098	2496	0.003393
P02768	0.000096	2480	0.003393
P10415	0.000095	2593	0.003429
I6LPK7	0.000094	2599	0.003431
A0A0C4DFU1	0.000094	2401	0.003401
P08183	0.000092	2424	0.003399
P14780	0.000092	2410	0.003416
P28482	0.000091	2549	0.003416
P27361	0.00009	2532	0.003411
A2A2V4	0.000088	2475	0.003418
P13500	0.000088	2500	0.003411

Label	Degree	Authority	pageranks
MESH:D007249	17693	0.036977	0.000575
MESH:D009336	17683	0.036999	0.000566
MESH:D056486	17679	0.036999	0.000565
MESH:D008569	17550	0.036947	0.000556
MESH:D011297	17580	0.03697	0.000553
MESH:D001943	17550	0.036932	0.000553
MESH:D002471	17497	0.036868	0.000551
MESH:D007859	17536	0.036946	0.00055
MESH:D006965	17508	0.036902	0.00055
MESH:D007674	17576	0.036975	0.000549
MESH:D003072	17486	0.036919	0.000548
MESH:D015431	17527	0.036953	0.000547
MESH:D048629	17325	0.036719	0.000547
MESH:D011471	17411	0.03682	0.000545
MESH:D009369	17333	0.036818	0.000544
MESH:D008175	17237	0.036677	0.000544
MESH:D009374	17443	0.036861	0.000543
MESH:D005234	17494	0.036944	0.000542
MESH:D008113	17398	0.036822	0.000542

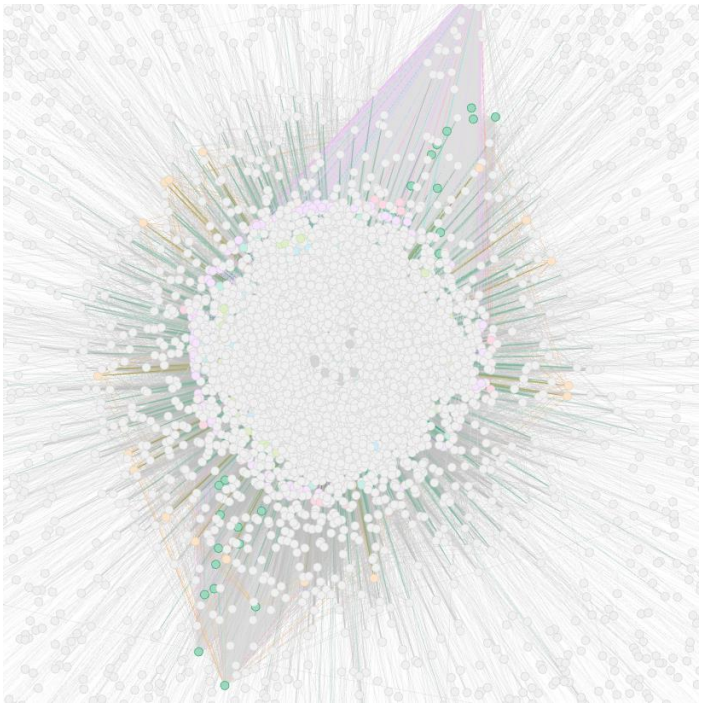
Filter to find top Nodes as per authority score



Disease	Authority Score
MESH:D009336	0.011289
MESH:D056486	0.011289
MESH:D007674	0.011282
MESH:D007249	0.011282
MESH:D011297	0.01128
MESH:D015431	0.011275
MESH:D007859	0.011273
MESH:D008569	0.011273
MESH:D005234	0.011272
MESH:D006973	0.011272

Gene	Authority Score
P01375	0.011274
P42574	0.011268
I6LPK7	0.011246
B5MC21	0.01124
P10415	0.011237
P01100	0.011216
A2A2V4	0.011201
P28482	0.011198
P04040	0.011196
P05412	0.011195

Filter top nodes per page Rank



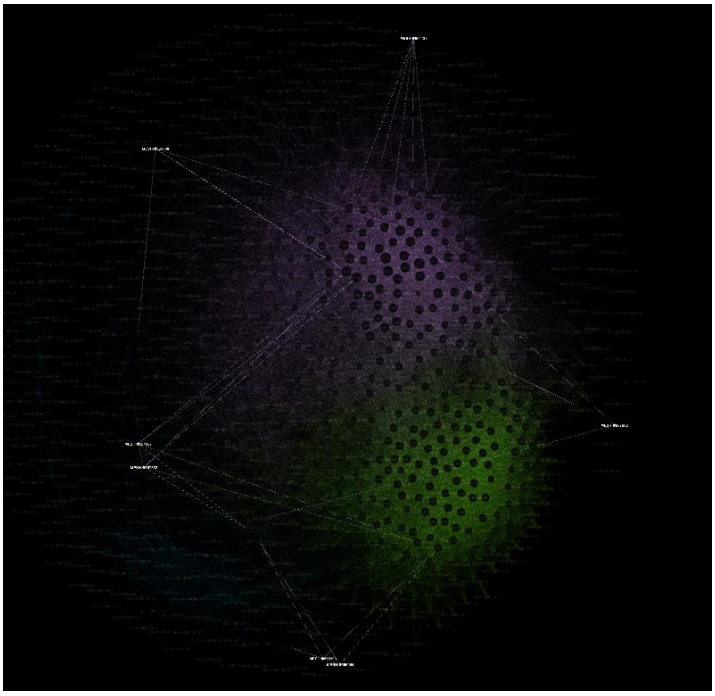
Disease	PageRank
MESH:D007249	0.000575
MESH:D009336	0.000566
MESH:D056486	0.000565
MESH:D008569	0.000556
MESH:D001943	0.000553
MESH:D011297	0.000553
MESH:D002471	0.000551
MESH:D007859	0.00055
MESH:D006965	0.00055
MESH:D007674	0.000549

Gene	PageRank
P01375	0.00013
P02458	0.000116
P01584	0.000115
P42574	0.000112
B5MC21	0.000109
A0A087WT22	0.000103
P04040	0.000101
P08684	0.000101
P01579	0.000099
P10145	0.000098

6.2 Disease-Disease Network:

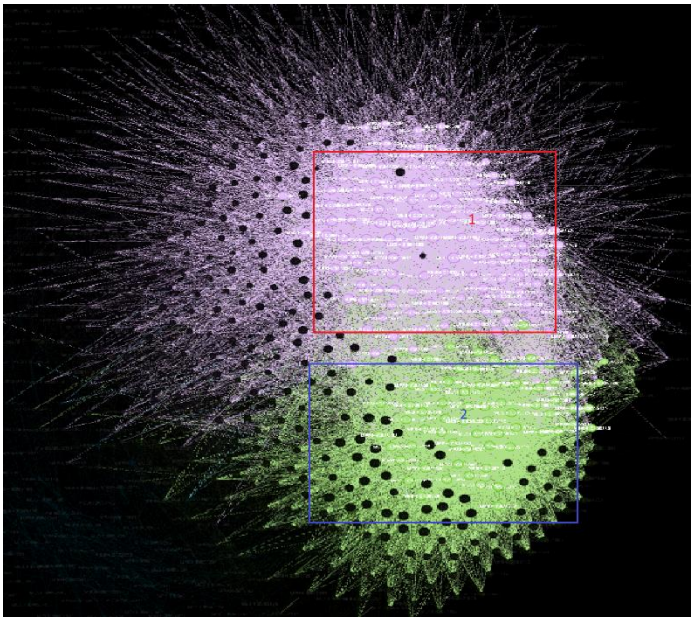
Filter top nodes by Clustering coefficient :

Disease	Clustering Co-efficient
MESH:C564133	1
MESH:C564567	1
MESH:C565193	1
MESH:C565780	1
MESH:C567582	1
MESH:C567502	1



All the Diseases had clustering coefficient as 1 were away from the hub of the network

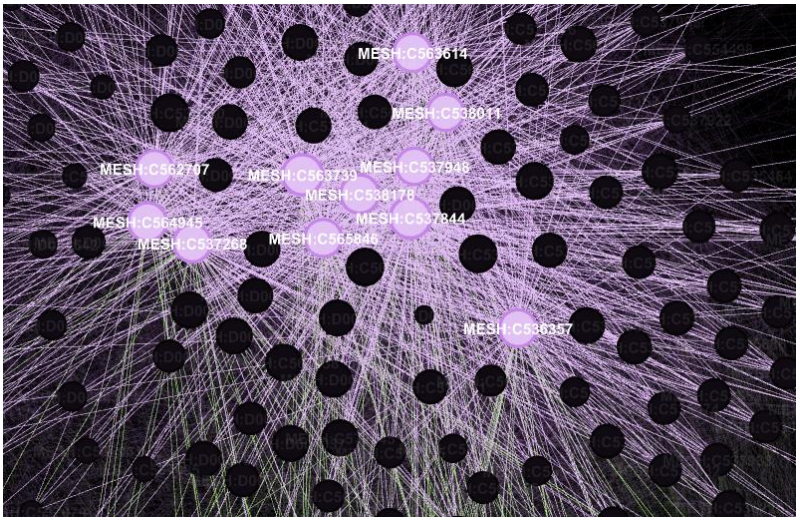
Filter nodes by hub score :



Disease	Hub Score
MESH:C536357	0.093241
MESH:C536495	0.089846
MESH:C536914	0.092866
MESH:C537844	0.093231
MESH:C537948	0.093109
MESH:C538011	0.090719
MESH:C538178	0.0997
MESH:C563614	0.08725
MESH:C563739	0.089798

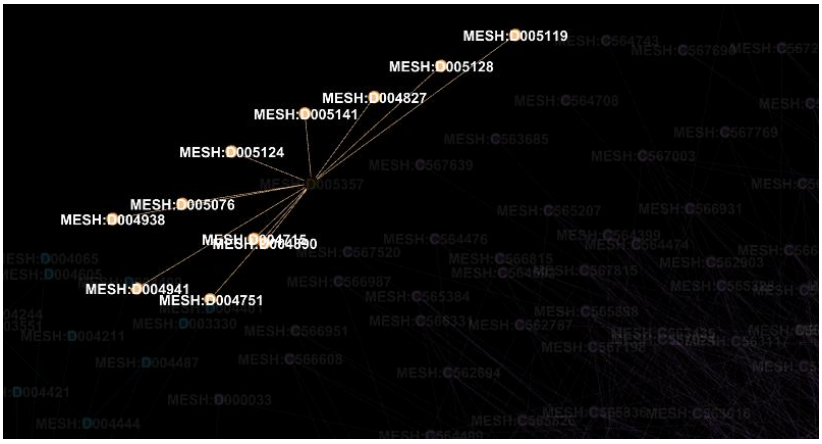
There were two major hub in his network which can be identified by filtering top 20 diseases having higher hub scores.

Filter by degree distribution :

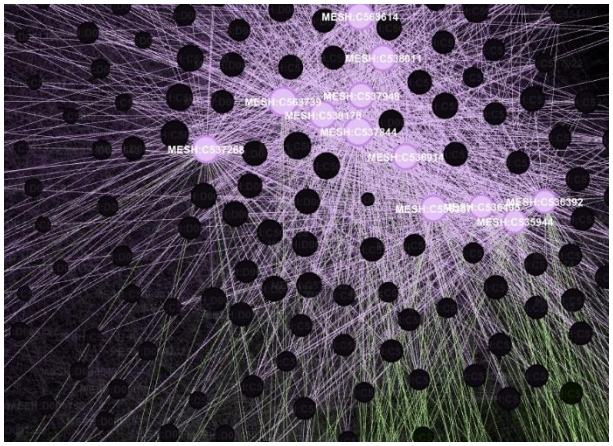


Disease	Degrees
MESH:C537844	154
MESH:C563739	152
MESH:C538178	142
MESH:C564945	141
MESH:C563614	140
MESH:C537268	137
MESH:C565846	135
MESH:C536357	134
MESH:C562707	132
MESH:C537948	131

All the diseases having higher degree distribution can be found in network hubs. Below graph showed the top 10 diseases having degree as one. These lowest degree diseases are away from the network hub.

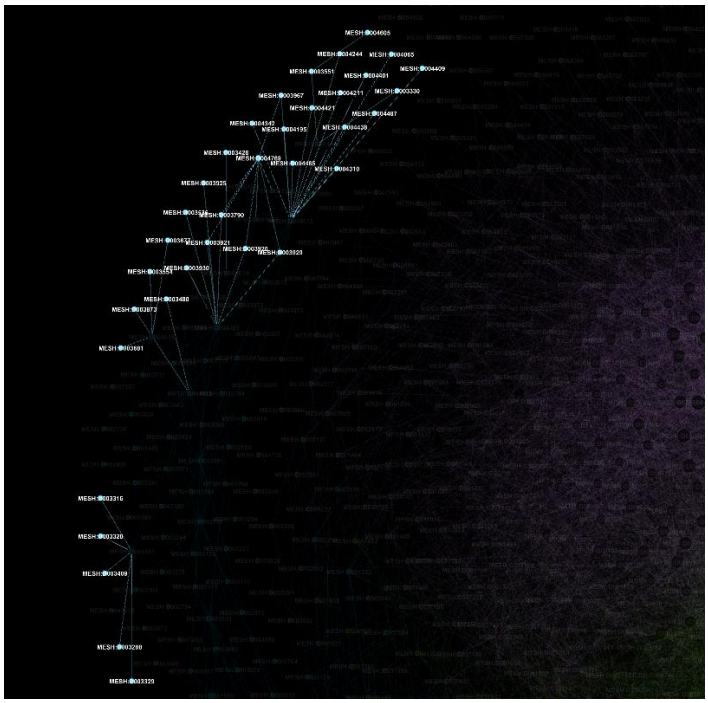


Filter by Eigenvector centrality :



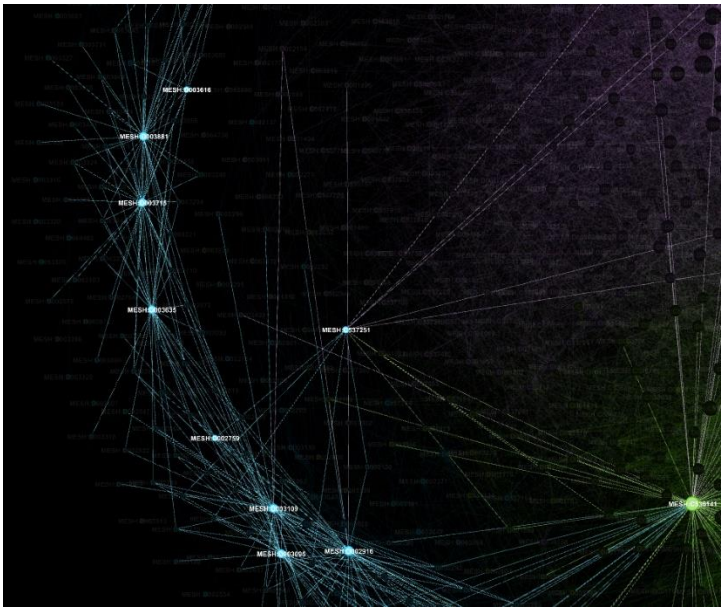
Disease	Centrality Score
MESH:C538178	1
MESH:C537844	0.941728
MESH:C536357	0.941043
MESH:C537948	0.933664
MESH:C536914	0.931269
MESH:C538011	0.90959
MESH:C536495	0.908401
MESH:C563739	0.9075
MESH:C563614	0.877281
MESH:C536392	0.869671

Filter by closeness centrality :



Disease	Centrality Score
MESH:D004605	0.114763
MESH:D004769	0.127587
MESH:D003330	0.129584
MESH:D003551	0.129622
MESH:D003554	0.130378
MESH:D003677	0.130378
MESH:D003681	0.130378
MESH:D003428	0.142675
MESH:D003288	0.142949
MESH:D003316	0.142949

Filter by betweenness centrality:

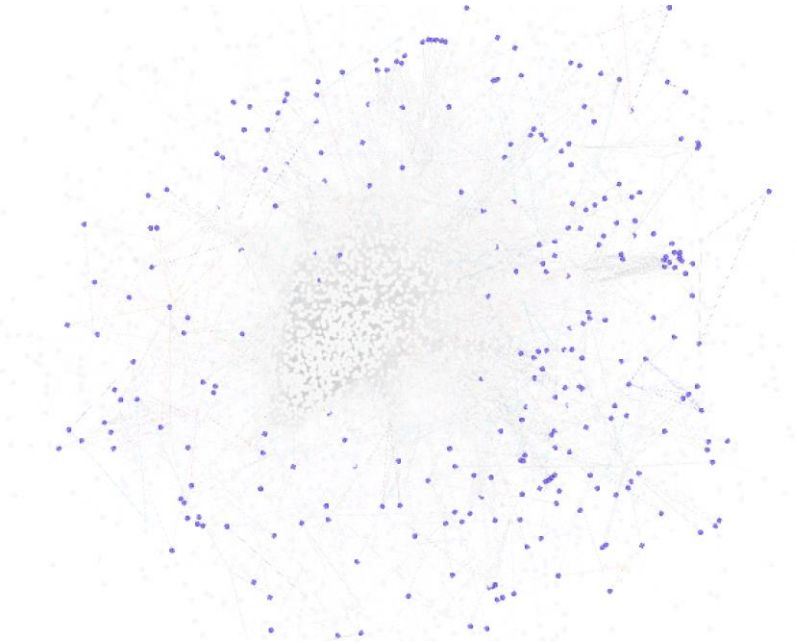


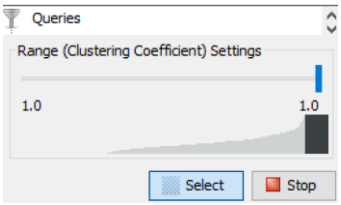
Disease	Centrality Score
MESH:D003635	34170.789
MESH:C537251	25850.80882
MESH:D002759	24171.68661
MESH:D003109	21892.13352
MESH:D003715	20195.71759
MESH:D003881	16385.57582
MESH:D003095	15858.54343
MESH:D002916	14601.38808
MESH:D003616	12497.57049
MESH:C536141	11874.85334

6.3 Gene-Gene Network:

Filter by Clustering co-efficient

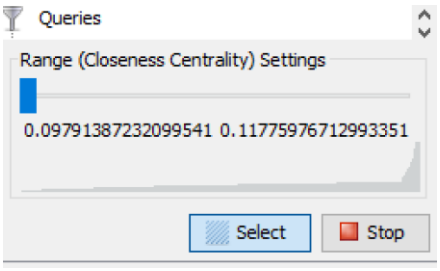
Showing all the nodes in green having clustering co-efficient as 1.



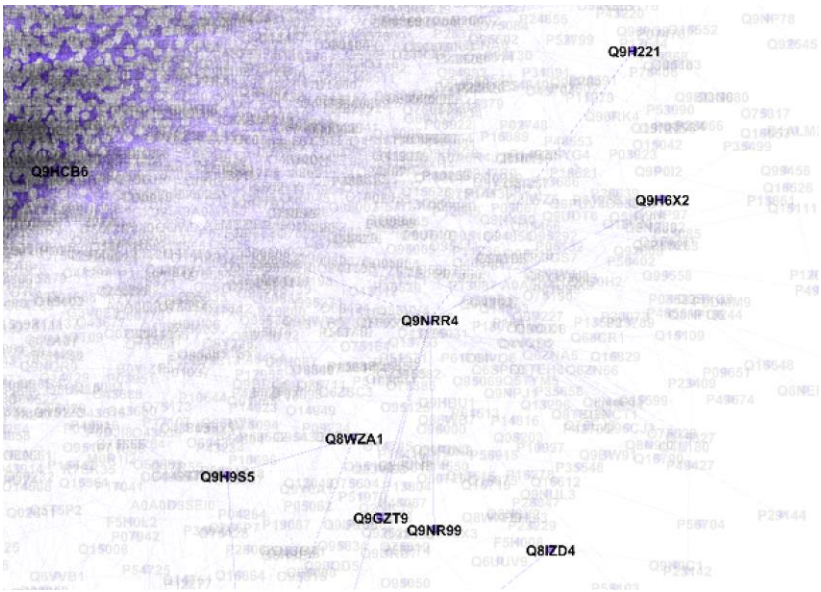


Closeness centrality

Filter by top 10 genes having lowest centrality score.



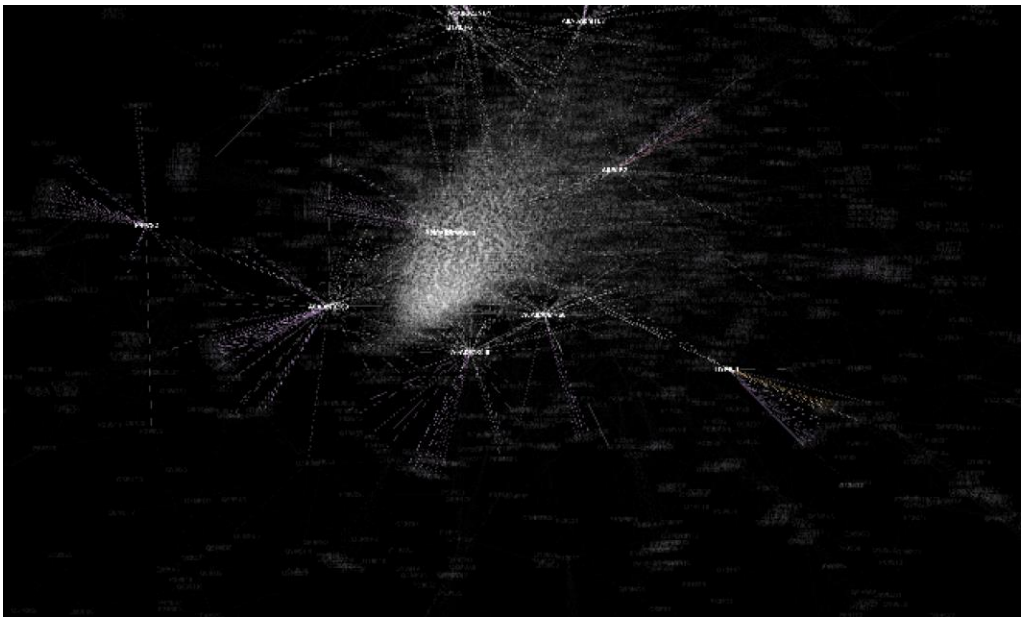
gene	centrality score
Q9NR99	0.097914
Q9H221	0.097914
Q9HCB6	0.097914
Q9GZT9	0.097914
Q9H6X2	0.097914
Q8IZD4	0.108504
Q9NRR4	0.108538
Q8WZA1	0.115643
Q9H9S5	0.115643
P39900	0.120585



Betweenness centrality:

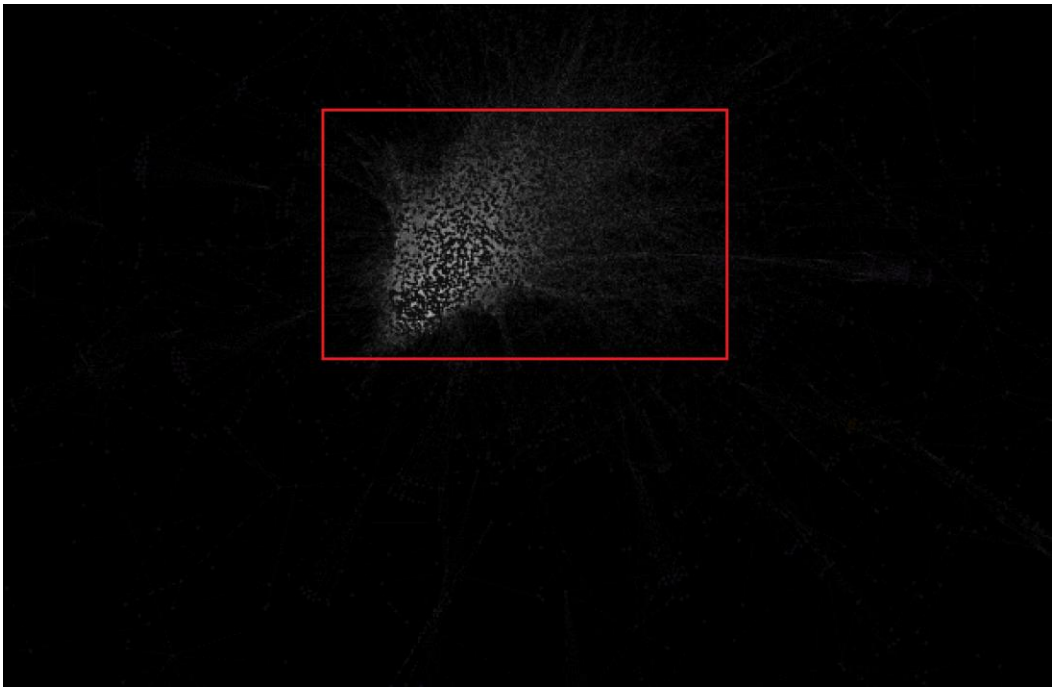
Filter by top 10 genes having highest centrality score.

Gene	Centrality Score
A0A087X169	271240.8869
A0A0A0MTL8	207013.8007
A0A087WW00	173887.1367
A0A087X0I6	159696.0108
A0A0A0MR60	144393.4026
A0A0G2JNB1	140452.7872
B1ALF6	127619.1252
P46782	120280.1664
A0PJE2	115897.7833
H9NIL4	102857.6364



Filter by hub score

All the nodes having highest hub score are highlighted in screenshot. In this graph there is one hub.

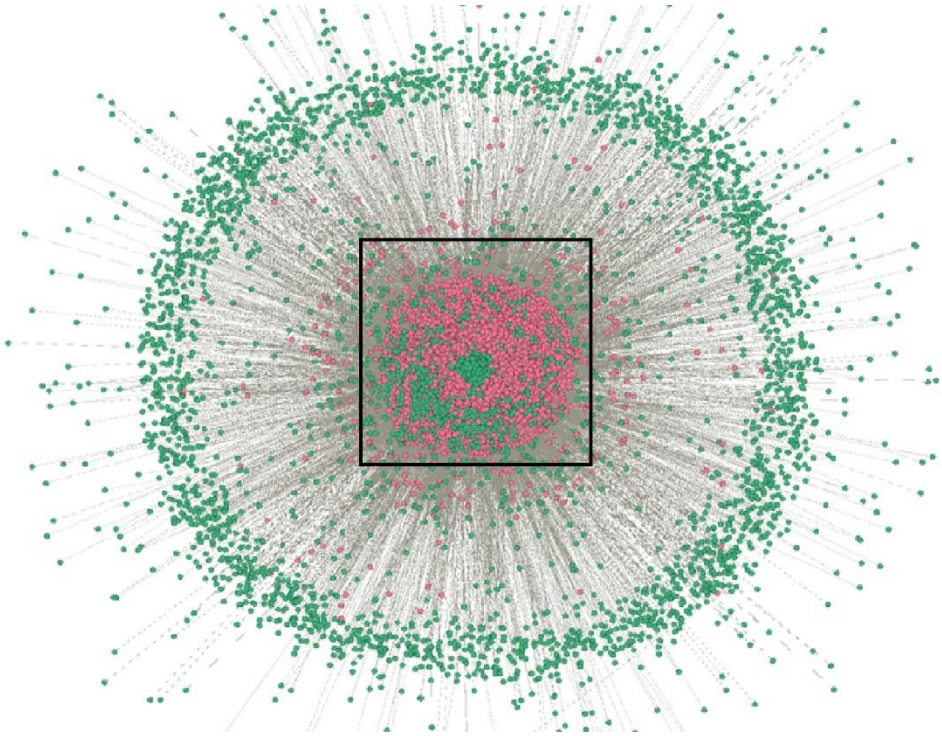


Q7:

Disease-Gene Network : -

Giant Component : -

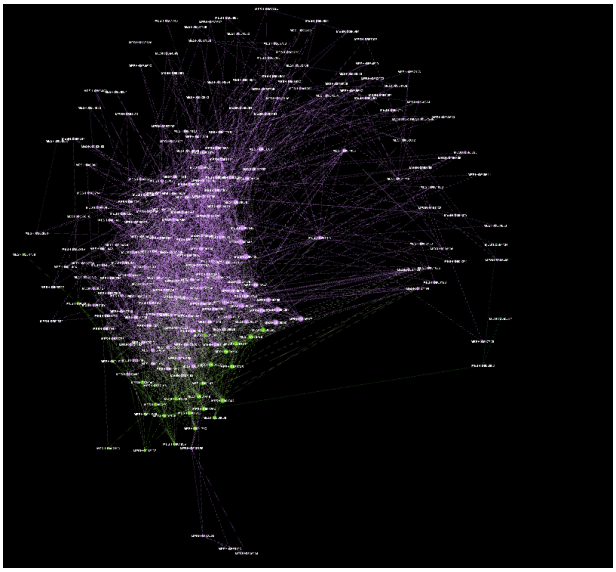
In a network, giant component is where all the nodes are tightly coupled with each other to make a giant structure. In Disease-Gene network there were a giant component as below image.



Disease-Disease Network : -

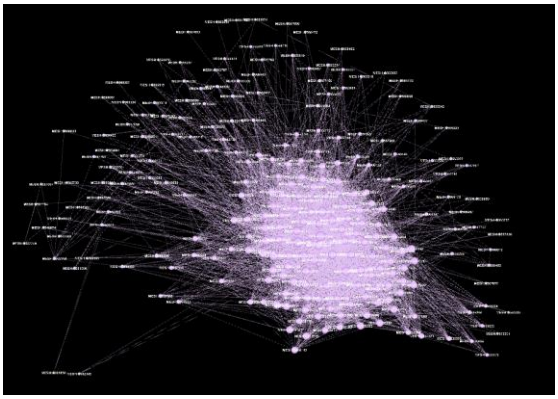
Communities:-

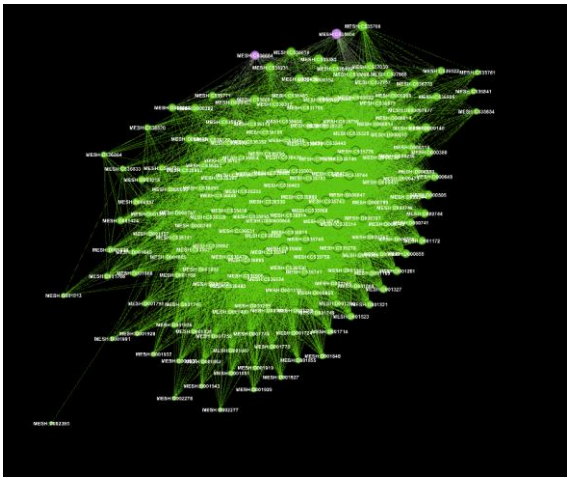
In this network there were 6 communities and gephi was able to club all the diseases depending on centrality measurement and clustering.



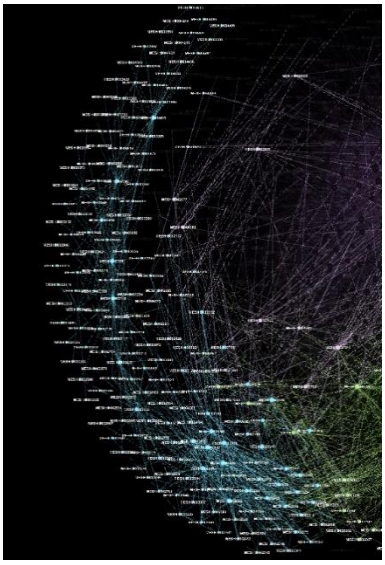
These communities have most of the disease having higher clustering coefficient.

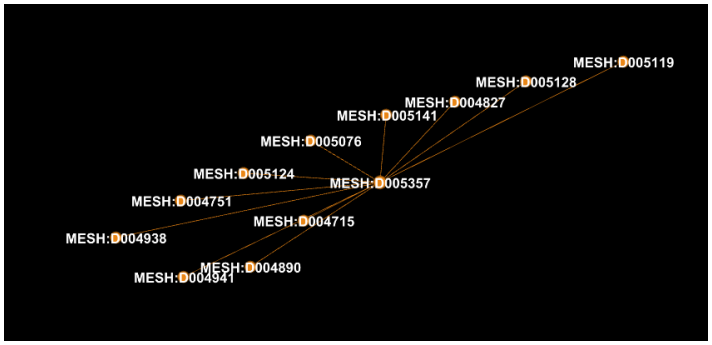
Below two communities represent the hub of the network where higher degree disease are close to each other.





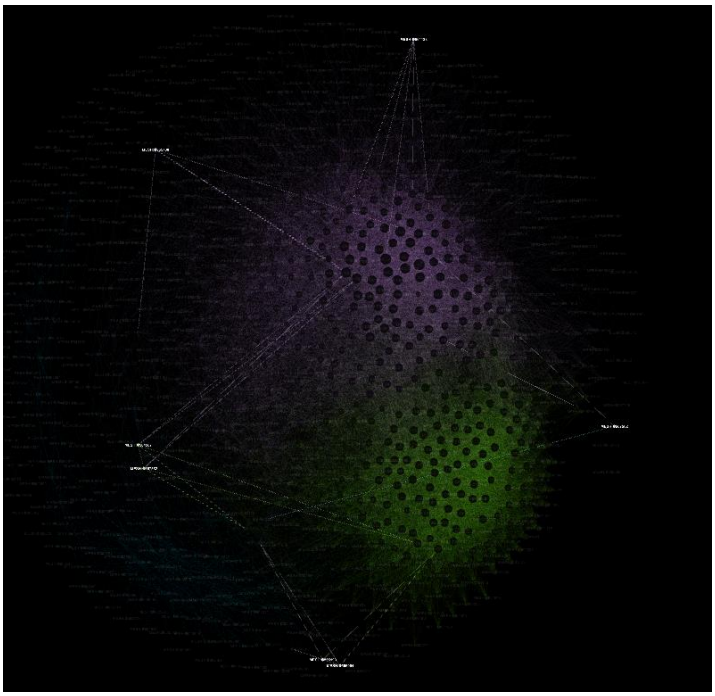
All below three communities have lower degree diseases and right two communities have diseases connected to each other but separated from other communities.





Homophily: -

Homophily is an extreme form of clustering, where nodes are highly interconnected within the group but largely disconnected from the other communities. In this network all disease which form perfect cluster are not highly disconnected from other groups. There were no such homophily present in this network.



Here all these clusters are connected to each other not disconnected from other communities.

Q8:

- What concrete results did you achieve in doing your project? Address this in terms of your research questions

The initial dataset got from SNAP database was a bipartite network. Because of huge link and node size it was very difficult to analyze the relationship between disease and gene. Because of that the huge network was decomposed into two separate monopartite network by filtering priority gene and diseases. Node degree, authority score and PageRank was used to rank each node (disease & gene) then filter out top nodes from each category. SQL database and python language were used to prepare disease-disease network by taking the common diseases linked with top 20 gene and similarly gene-gene network was formed by taking common gene linked with top 20 diseases. Below are the top 20 gene and diseases extracted from this giant network.

Gene	PageRank	Degree	Authority Score
P01375	0.00013	2801	0.00344
P01584	0.000115	2616	0.00339
P42574	0.000112	2754	0.003438
B5MC21	0.000109	2690	0.003429
A0A087WT22	0.000103	2475	0.003397
P04040	0.000101	2582	0.003416
P08684	0.000101	2444	0.003411
P01579	0.000099	2516	0.003411
P10145	0.000098	2496	0.003393
P02768	0.000096	2480	0.003393
P10415	0.000095	2593	0.003429
I6LPK7	0.000094	2599	0.003431
A0A0C4DFU1	0.000094	2401	0.003401
P08183	0.000092	2424	0.003399
P14780	0.000092	2410	0.003416
P28482	0.000091	2549	0.003416
P27361	0.00009	2532	0.003411
A2A2V4	0.000088	2475	0.003418
P13500	0.000088	2500	0.003411

Label	Degree	Authority	pageranks
MESH:D007249	17693	0.036977	0.000575
MESH:D009336	17683	0.036999	0.000566
MESH:D056486	17679	0.036999	0.000565
MESH:D008569	17550	0.036947	0.000556
MESH:D011297	17580	0.03697	0.000553
MESH:D001943	17550	0.036932	0.000553
MESH:D002471	17497	0.036868	0.000551
MESH:D007859	17536	0.036946	0.00055
MESH:D006965	17508	0.036902	0.00055
MESH:D007674	17576	0.036975	0.000549
MESH:D003072	17486	0.036919	0.000548
MESH:D015431	17527	0.036953	0.000547
MESH:D048629	17325	0.036719	0.000547
MESH:D011471	17411	0.03682	0.000545
MESH:D009369	17333	0.036818	0.000544
MESH:D008175	17237	0.036677	0.000544
MESH:D009374	17443	0.036861	0.000543
MESH:D005234	17494	0.036944	0.000542
MESH:D008113	17398	0.036822	0.000542

For both disease-disease and gene-gene network below statistical analysis was used to answer the research questions. In statistical analysis section all statistical measurement was explained in details by filtering top 10 disease and genes.

Research Question	Measurement	Network
1. What is the depth/degree of the relationship between participating nodes?	Degree Distribution	Disease-Gene
		Disease-Disease
		Gene-Gene
2. Are there identifiable clusters based on which the genes/Disease can be grouped? 3. How to find similarity between gene , Disease and its neighbor?	Leiden Algorithm & HITS	Disease-Gene
	Clustering Co-efficient & Modularity	Disease-Disease
		Gene-Gene
4. Are there identifiable hubs?	HITS (Authorities & Hubs)	Disease-Gene
		Disease-Disease
		Gene-Gene
	Centrality Measurement	Disease-Gene

5. What are the strongest nodes in terms of connectivity and what are the weakest nodes? 6. What is the betweenness of the nodes?		Disease-Disease Gene-Gene

This network framework can be used to analyze the relationship between for any group of disease related to a specific gene or group of genes related to specific disease.

- Address your contemplation about the project. Include, for example, other data that could have made your study more relevant, the difficulty in dividing the workload between team members, data preparation, what was unexpected, etc.

  - In the initial dataset all the diseases were represented by MESH code and genes by UniProtIDs code. We found it very difficult to understand the category of disease or genes due to lack of prior knowledge related to biological network. Large amount of time was used to do data preparations and decompose the giant bipartite network into two monopartite networks.
  - In this project we got a chance to learn more about bipartite network topology. We too investigated couple of tools like Cytoscape, Gephi, Pajek and language like NetworkX using python.
  - As a team Anwesh Praharaj has worked on data preparation, disease-gene and disease-disease analysis network where Shriram Sreedhar has worked on gene-gene network analysis. Since we have been working remotely, the co-ordinations and distributions were a bit tough.

  Future Work:

  Many different experiments and analysis have been left for the future due to lack of time (i.e Analysis with huge amount of data are usually very time consuming and taking even days to decompose into smaller dataset)

  - Different machine learning algorithm can be used to rank the nodes in disease-gene network which may improve the efficiency of our decompose mechanism into two separate networks (Disease and Gene).
  - Weighted edges dataset could be considered to identify how likely two diseases or genes are associated with each other.
  - Proper human readable labeling for disease and gene could be used to segregate different category of disease or gene into one group for analysis purpose.

Bibliography

Masuda N, Sakaki M, Ezaki T, Watanabe T. Clustering Coefficients for Correlation Networks. Front Neuroinform. 2018;12:7. Published 2018 Mar 15. doi:10.3389/fninf.2018.00007

Ostroumova Prokhorenkova, Liudmila & Samosvat, Egor. (2014). Global Clustering Coefficient in Scale-Free Networks. 10.1007/978-3-319-13123-8_5. Selim, H., Zhan, J. Towards shortest path identification on large networks. J Big Data 3, 10 (2016). https://doi.org/10.1186/s40537-016-0042-7

Rui Zhang, Lei Li, Chongming Bao, Lihua Zhou and Bing Kong, "The community detection algorithm based on the node clustering coefficient and the edge clustering coefficient," Proceeding of the 11th World Congress on Intelligent Control and Automation, Shenyang, 2014, pp. 3240- 3245, doi: 10.1109/WCICA.2014.7053250.