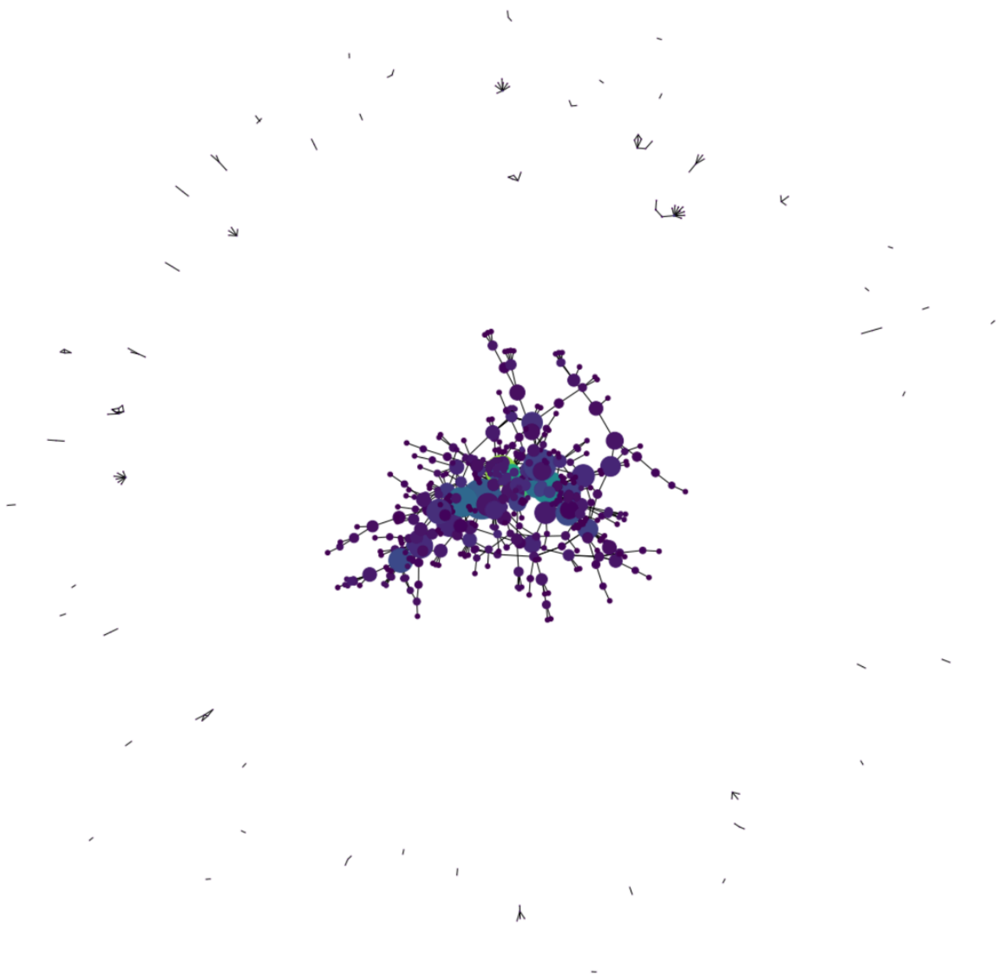
MiniProject_2

Student Name: Ganga Lingden

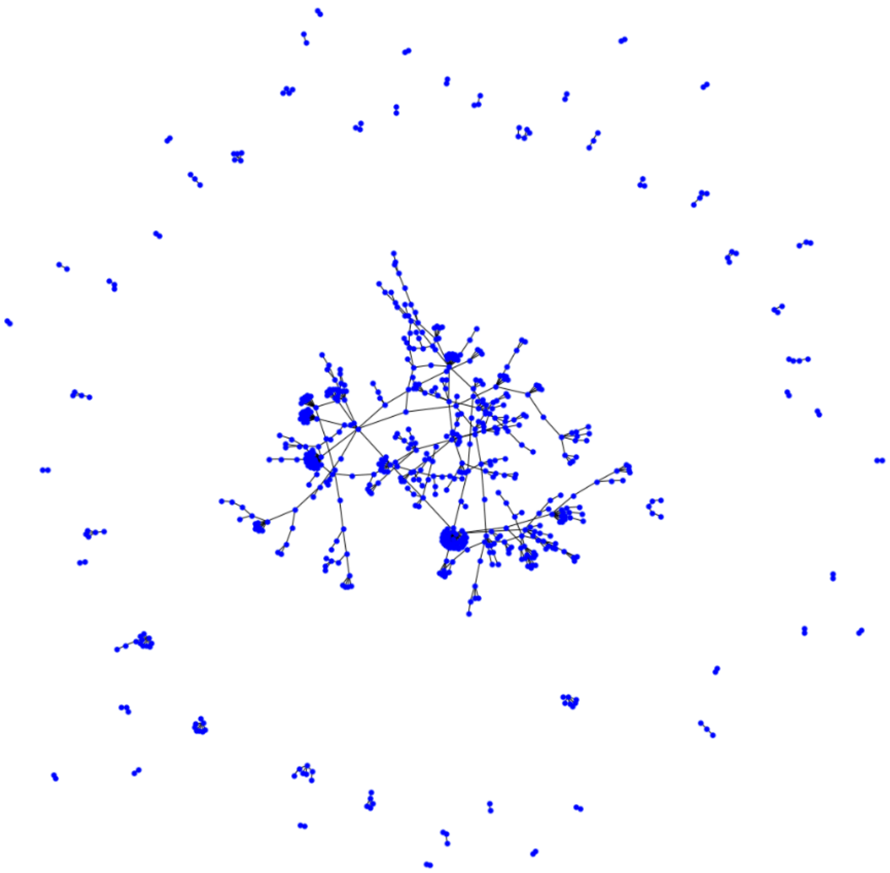
1. Goal: To implement and analyse the protein-protein interactions dataset using Networkx.

2. Procedure: In the section, all the essential steps that were taken during the execution of the project are mentioned step by step. In the project , firstly the dataset (DM-LC.txt) was downloaded and understand about it. The data has three columns, first two indicates protein and the third one as their interaction. So, to related this to the term of graph theory the first two columns are considered as nodes and third one (interaction) was considered as weight/cost of the edge between these two nodes.

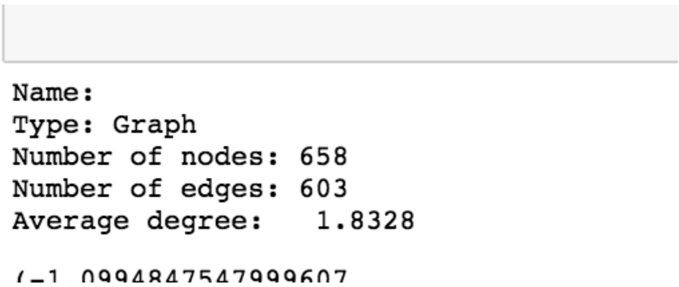
2.1 Create Network : First, the dataset was converted into pandas data frame, which makes easy for creating the network. The network (g) was drawn with the function : **g=nx.Graph()** and then every instances (each row in data frame) of data were added in network using the function: **g.add_weighted_edges_from([(node, next_node,weight)])**. Then a complete undirected network was created. The screen shot is shown below:



2.3 Findings: In the section, the information that was observed in the network (g) are presented. The total number of nodes are 658, edges are 1129 and average degree is 3.4316. This was performed with function: network_info = nx.info(g). The density of the network was found 0.00522 using **density = nx.density(g)** function. And, the minimum spacing tree was calculated using **T = nx.minimum_spanning_tree(g)**. The minimum spacing tree is the subset of the edges of graph that connects all the vertices together, without forming any cycles, but with the minimum possible total edge weight. The screenshot diagram of the minimum spanning tree is shown below:



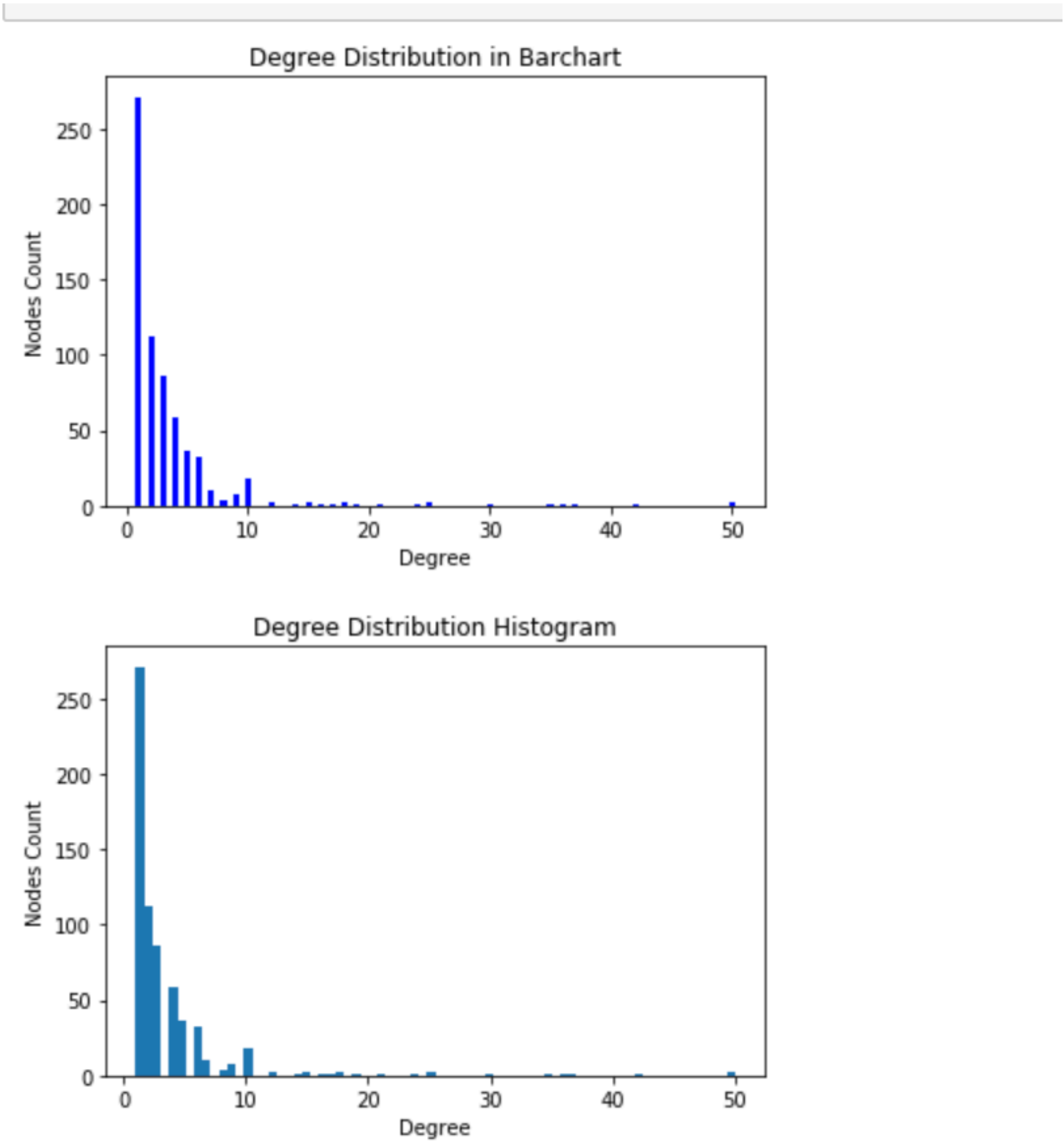
Even thought the minimum spanning tree (mst) looks same as the network (g) here in figure, it is not actual the same. The edges are less in this mst diagram than network (g). The output from the mst info is shown below:



```
Name:
Type: Graph
Number of nodes: 658
Number of edges: 603
Average degree:   1.8328

(-1.0994847547999607
```

And the next part was to draw the degree distribution of the network (g). The degree distribution is the number of nodes have each degree. Basically, it insights or tells us about the how the structure of a network is formed. Below the screen shots that were observed after building degree distribution.



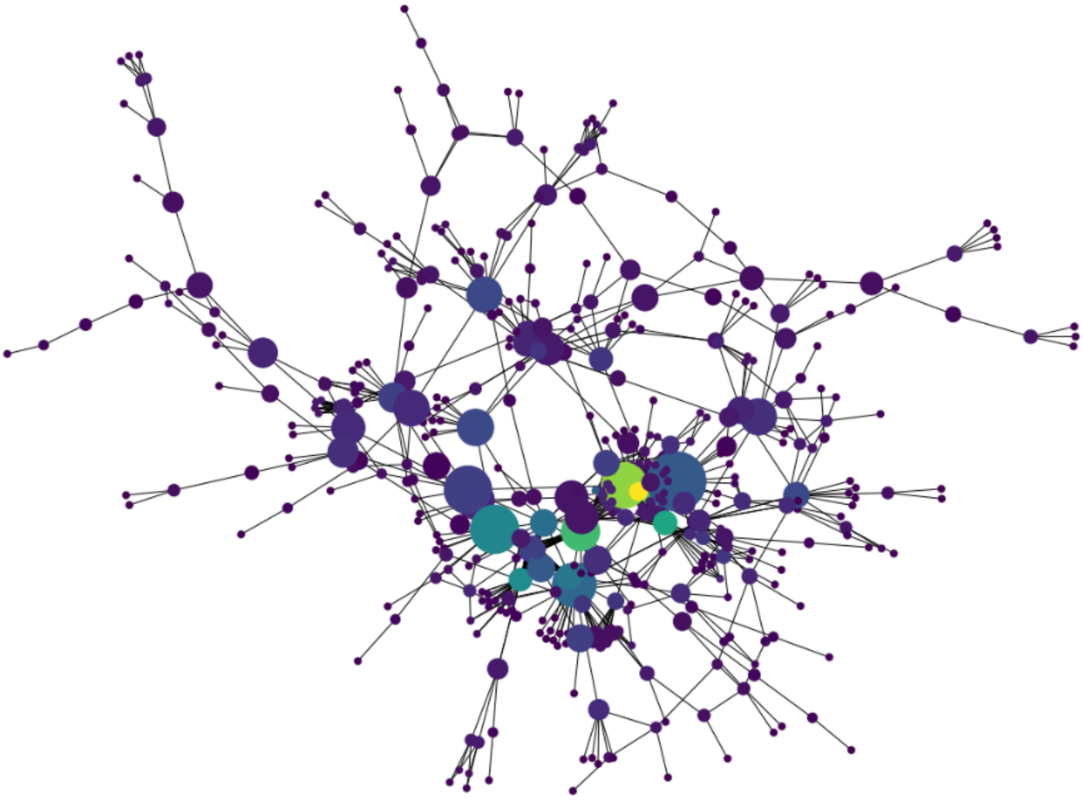
Both the above figure represent the same meaning , but the first one is the bar chart showing the node number with respect to degree. And the last one is the histogram that was built with the frequency of degree in network (g). The bin size in this histogram was used as 'auto'. Both of the figure depicts the same result. From degree distribution figures, it can say that the majority of nodes(proteins) are with very few degree while fewer nodes have the highest degree(50),which is also called as hubs of the networks. The degrees in the network (g) are range from 1-50 inclusively.

Last part was to find the largest component (LC) in the network (g). The LC was found by comparing the all the components and find the largest one in term of their size(length), i.e assuming the largest the component has high number of nodes. This is the code used for LC.

components = nx.connected_components(g)

largest_component = max(components, key=len)

And, LC is drawn as the subgraph of the network and the below is the diagram that depicts the LC in the network (g).



The diameter of LC was 18, which tells that this is the longest path of all the calculated shortest path in the network between one node to far another node. This depicts the linear size of the network (g). The centre of LC was found **['W02D3.9']** protein and the number of clique communities with 3 nodes was 31.

For finding the protein that has the biggest effect in the network while chaining its status, the three centrality measures (degree of centrality, Eigenvector centrality and betweenness centrality) were performed. In the LC, the protein - '**F58A4.3'** was observed for degree of centrality and eigenvector centrality while the protein - '**W02D3.9'** was found as the betweenness centrality. The degree of centrality tells only about how many edges connected to a node i.e. node of the highest degree network, no matter how importance are neighbouring nodes. It doesn't take care about it. Unlike degree of centrality, eigenvector centrality measures the node with highest degree along the importance of neighbouring nodes. The betweenness centrality shows the important node where the majority information passes through it . Thus, **'W02D3.9'** is the protein in LC that has the maximum influence in the network when changes happen in it because it is in critical node where the interaction (information) between one to another group of nodes (proteins) in LC pass through it. Thus, incase it goes inactive there is no flow of information between group of nodes in network.