Problem 1

1(b)

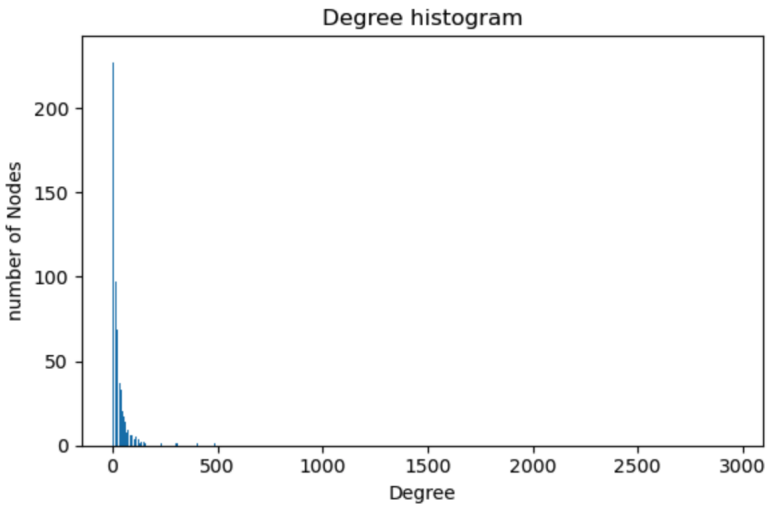

Figure 1: Degree Distribution

1(c)

YPL098C has a higher clustering coefficient which is 0.8. The difference means YPL098C's neighbors are closer to forming a clique(functional module in the PPI network) than YGR296W's.

1(d)

There are 354514 3-cliques in the PPI network.
The global clustering coefficient of this network is 0.3045684470122461.

1(e)

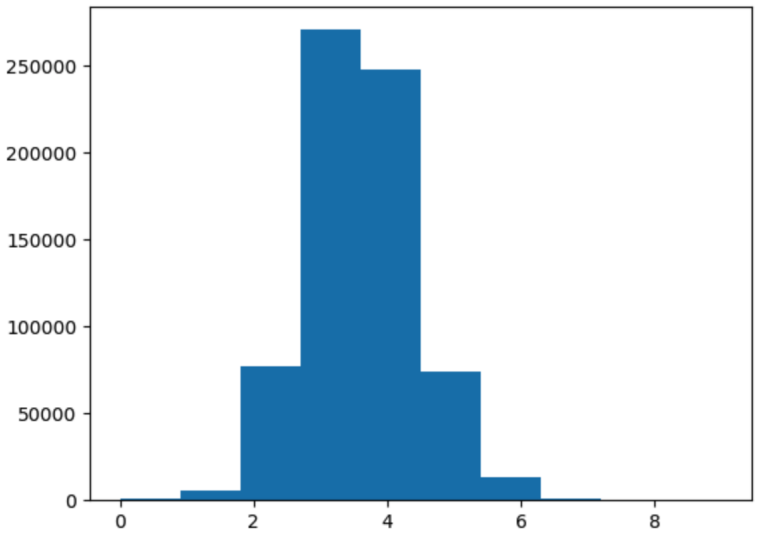

Figure 2: Distribution of Shortest Path Distance

I expected the shortest path distance for a protein-protein interaction network would be high. But it turns out to be less than 10.

1(f)

I first estimate that the diameter of the network must be no more than 7 because the diameter for my sub-graph is 7, maybe 4 or 5. The global clustering coefficient is about 0.3 which is way less than 1. So the density of the graph would not be too high. Then the diameter of the graph would not be too close to the random sub-graph.

Problem 2

2(b)

For the original algorithm, each time I found a label belonging to the direct neighbor of a node I add 1 to its weight.
And for the variation of the original algorithm, each direct neighbor has a weight 1. Then the weight of labels are equally distributed. For example, if one neighbor has 5 labels, then each label weighs 0.2 in this condition. The reason I did it this way is because I think we should treat each node equally, therefore, the weight of each node should be equal.

2(c)

For algorithm in 2(a), the number of successful test is 1422 and the number of unsuccessful test is 3021, which gives the accuracy 32%.
For algorithm in 2(a), the number of successful test is 1493 and the number of unsuccessful test is 2950, which gives the accuracy 33.6%.

2(d)

For multiple labels, I think maintaining a threshold is a good way to do it. Labels whose weights are above the threshold are assigned to the node.
And the performance can be measured by evaluating the number of intersection labels of ground truth and predicted labels.