COMP 150NS: Network Sciences
Tufts University

Prof. Lenore Cowen
Fall 2020

HW 3: due Thursday, October 1

For this assignment, submit your answers for both problems as a single pdf filenamed hwk3.pdf, alongside any code you may have written, using provide with the command:

```
homework% provide comp150ns hwk3 hwk3.pdf code-you-wrote
```

where code-you-wrote is replaced with any source files you may have written.

1. For this problem, you will need to download the yeast PPI physical interaction network available as the PPI data set on the course site (http://www.cs.tufts.edu/comp/150NS/private/datasets.html)

   The file includes 76025 yeast protein protein interactions (PPI) among 5001 proteins. The file is tab-delimited. Each row corresponds to one interaction and has three columns: the first and the second columns are the two interacting proteins, while the third entry is an number in (0, 1] that shows the confidence of the existence of the interaction.
   First, you will construct a simple undirected, unweighted graph from this protein-protein interaction network. Namely, each graph node corresponds to a protein in the protein interaction network, and an undirected edge exists between two nodes if there is an interaction between the two proteins.
   After you construct the graph, please use it to answer the following questions, some of which will require additional graph analysis methods.

   (a) Many computational biologists have concluded that PPI networks are scale-free. Compute the degree diatribution for the 5001 nodes and submit it as a histogram with your program.

   (b) Compute the local clustering coefficient for each node in the graph. Submit a tab-delimited file containing two columns, where the first column is the protein ID and the second column is the clustering coefficient. Both YGR296W and YPL098C have 5 interacting partners; which one has the higher clustering coefficient? Explain briefly what the difference means.

(c) Implement a function that counts the number of triangles, or 3-cliques, in the graph. A k-clique is defined as a graph with k nodes with an edge between every pair of nodes. In general, we consider that a clique forms a functional module in the PPI network; thus, it will be interesting to find all of them. How many 3-cliques are there in the PPI network? What is the global clustering coefficient of this network?

(d) In some applications, we seek to find "close" pairs of proteins based on the network and study the similarity between such "close" pairs. A simple way to define the "closeness" is to use shortest path distance.
In order to estimate the path length distribution for this graph, sample 1000 nodes at random from the graph, and compute the distribution of shortest path distances between these 1000 nodes. Please submit a figure showing the distribution of shortest path distances you observe. How does this compare to your expectations for a protein-protein interaction network?

(e) Estimate the diameter of the network based on your answer to the previous question. How does what you found relate to the results of the previous question?

2. For this problem, you will need the PPI network from the previous problem, and you will need to download the known functional labels from the top level of MIPS posted here included with the PPI data set.

(a) Implement the majority vote algorithm. Given a yeast protein, have each of its labeled neighbors vote once for each of their labels, and assign that protein the single label that got the most votes (if there is a tie, break it numerically by MIPS number, assigning it to the lower-numbered MIPS term).

(b) Implement a different algorithm, or a variation of this algorithm to accomplish the same task. Explain carefully what you did, and why.

(c) Measure the performance of both the majority vote algorithm and your algorithm in *leave one out cross validation,* where a protein is marked as correctly labeled if it assigned one of its correct labels.

(d) In the previous question, you were asked to predict a single correct functional label, but some proteins have been assigned multiple labels. Say a few words about how you might extend the majority vote algorithm to allow the assignment of multiple labels. How might you measure performance in this setting?