## CSCI479/679 Introduction to Data Mining

### Assignment #4 (100 points): Network Topological Properties
### Deadline: Thursday, November 22nd at 11:59PM

# 1 Description

Enclosed with this assignment, you will find the Yeast protein protein interaction network (HcNetwork.txt). In interaction networks, nodes represent genes and an edge between two genes indicates an interaction between the two genes. Write an R/Python script to compute several of the network topological properties. For more details on how to compute the graph properties, please refer to Chapter 4. Note: You have to write the code and not use any existing graph libraries such as igraph or networkx.

1. **Degree Distribution:** Recall that the degree a node in an undirected network is the number of edges the node has to the other nodes. Let $k_j$ be th number of nodes whose degree is $j$, and probability $P(j)$ is the probability that a node in the graph has degree $j$, i.e., $P(j) = k_j/n$, where $n$ is the number of nodes in the graph.

   Plot the order pairs $(j, k_j)$ on a scatter plot. Next, plot the degree $j$ and the probability $P(k_j)$ on a scatter plot in a log-log scale to see whether the plot looks like a straight line.

2. **Clustering Coefficient:** The Compute clustering coefficient of a node $v_i$ is a measure of the density of edges in the neighborhood of node $v_i$ and is defined as

   $$C(v_i) = \frac{\#\text{ of edges in } G_i}{\text{maximum number of edges in } G_i} = \frac{2 \times m_i}{n_i \times (n_i - 1)}$$

   where $m_i$ is the number of edges in $G_i$ and $n_i$ is the degree of node $v_i$. If a node has a degree less than 2, assume that its local clustering coefficient is 0. The clustering coefficient of a graph $G$ is the average clustering coefficient over all the nodes of the graph.

   Let $C(k)$ denote the average clustering coefficient of all the nodes with degree $k$, Plot the order pairs $(k, C(k))$ on a log-log scatter plot to see whether the plot looks like a straight line.

### 1.1 Hints:

Read the network as a table then create an adjacency matrix for the network. Iterate over the table to populate the matrix. Once the matrix is populated, things will be much easier. Check your work on the network in “toyN.txt”.

## 2 Submission

Your should submit your assignment to the black board. The code file should be named in the following format, useridAssig4.R, useridAssig4.py or useridAssig4.m. Moreover you should submit a report showing the plots for the degree and clustering coefficient distribution for the yeast network.

**Good Luck!**