PPI networks and Cytoscape
Due 11/22/2016 at 10pm.

A Protein-Protein Interaction (PPI) network is a graph where nodes represent proteins and an edge between two nodes represents interacting proteins (either physically or functionally).
You will write procedures that compute the following global and local properties of a PPI network:

Degree distribution
Clustering Coefficient
Node closeness-centrality
Edge Betweenness-centrality

A detailed description of those properties, already presented in class, follows:

Degree distributions
Recall that the degree of a node in a network with *n* nodes is the number of edges incident on (i.e., connected to) that node. We define p_k to be the number of nodes in the network that have degree k. The degree distribution of a network is a histogram or table of the frequencies p_k for all k.

Clustering coefficient
In many networks, if node A is connected to B, and B is connected to C, then it is highly probable that A also has a direct link to C. This phenomenon can be quantified using the clustering coefficient. Given node *u* with degree k_u, it is defined as

$$ClusterC(u) = 2 \times n_u / k_u \times (k_u\text{-}1)$$

where n_u is the number of links connecting the k_u neighbors of node *u* to each other. In other words, $ClusterC(u)$ gives the number of 'triangles' that go through node u, whereas $k_u \times (k_u\text{-}1)/2$ is the total number of triangles that could pass through node *u*, should all of node u's neighbors be connected to each other.

The average clustering coefficient, AVERAGE_C, characterizes the overall tendency of nodes to form clusters or groups. An important measure of the network's structure is the function AVERAGE_C(k),which is defined as the average clustering coefficient of all nodes of degree k.

Closeness centrality of a node
Closeness-centrality is a measure of node centrality and uses information about the length of the shortest paths within a network; it uses the sum of the shortest distances of a node to all other nodes. The closeness-centrality is defined as the reciprocal of this sum:
.

$$ClosenessC(u) = \frac{1}{\sum_{v \in V} length\_shortest\_path(u, v)}$$

Betweenness centrality of an edge
In networks, the greater the number of paths in which an edge participates, the higher the importance of this edge for the network. Thus, assuming that the interactions follow the shortest paths between two nodes, it is possible to quantify the importance of an edge in terms of its betweenness centrality defined as:

$$BTW\_C(e) = \sum_{i,j} s(i, e, j)/s(i, j)$$

where the sum Σ is over all pairs i, j of distinct nodes, $s(i, e, j)$ is the number of shortest paths between nodes i and j that pass through edge e, and $s(i, j)$ is the total number of shortest paths between i and j.

INPUT DATA
In this assignment you analyze the Protein-Protein Interaction (PPI) graph of the herpes Kaposi virus. The file kshv.sif (in T-square resources) contains such a graph in sif format. Each line of the file represents an edge and looks like the one below:

```
kshv_ORF53 1.0   kshv_ORF45
```

In the above example, the edge connects the two proteins `kshv_ORF53 and kshv_ORF45.` The intermediate value 1.0 on the same line is the weight of the edge. In your file all weights are 1.

IMPLEMENTATION
Use an *adjacency matrix* representation of the graph, i.e. a $n \times n$ *matrix* ADJ with ADJ[u,v]=1 if there is an edge between nodes u and v, 0 otherwise. Use hashes to map the protein ids into the indexes of the matrix. Write a procedure **Create-Adjacency-Matrix that** takes as input argument the .sif representation of the graph and generates ADJ.

OUTPUT

Your procedure prints:

1) a table of the degree frequencies, where each line consists of a value of k and the corresponding p_k
2) the average clustering coefficient AVERAGE_C of the network
3) for all k, the average clustering coefficient AVERAGE_C(k) of all nodes of degree k
4) the top 5 nodes (print the name of the proteins) with highest cluster coefficient.
5) the top 5 nodes (print the name of the proteins) with the highest closeness centrality measure

Electronically submit the following:
a) A cytoscape drawing of the graph kshv.sif in which the label of an edge is the betweenness centrality value of that edge. Submit the drawing in a separate file called last_name_graph_betweenness.
c) the file containing your perl program, called last_name_program4.pl

GRADING.

The assignement is worth 100 points. The computation of the betweenness centrality measure is worth 20 points.