- **Instructions :**
  You may use software libraries that implement the required functions.
  A few examples software libraries for graphs, are SNAP (a C++ library http://snap.stanford.edu/ written in C++, python interface also provided), igraph (https://igraph.org/, available with R, python and C), NetworkX and Neo4j.
  - While you may discuss with others, your assignment submission must be your own work.
  - You must describe your approach in sufficient detail.
    - This would be helpful for partial credit in case you are not able to complete your implementation in time.
  - Do not just report the final answers.
  - For full credit, you must conduct some analysis of the results and clearly explain any insights obtained from these analyses.

Note: While creating/performing operations on the Graph if your notebook/system crashes due to the memory issue, you can subsample the Graph (delete some of the nodes and edges randomly).

Q1. (5 points)

About Dataset: This is a collection of physical protein-protein interaction networks for a large number of human tissues. Nodes represent human proteins and edges represent tissue-specific physical interactions between proteins.

Download the PPT-Ohmnet networks available at http://snap.stanford.edu/biodata/datasets/10013/10013-PPT-Ohmnet.html (click on the link PPT-Ohmnet_tissues-combined.edgelist.gz to download the data). Identify the largest weakly connected component and do the following on the largest connected component :

(a) Compute the degree distribution and display it graphically. Identify the top 5 proteins which interact with most other proteins.

(b) Compute Pagerank (Proteinrank) for all nodes and calculate the betweenness centrality , closeness centrality , eigenvector centrality for nodes that have top 10 page rank and and that have least 10 page rank and then comment on the values of he betweenness centrality , closeness centrality , eigenvector centrality.

Q2. (5 points)

Node similarity is an important task in graph analysis, many node similarity approaches have been proposed in the literature. In this problem you are asked to compute node similarity using the following approaches for a Drug-Gene graph available at https://snap.stanford.edu/biodata/datasets/10002/10002-ChG-Miner.html (click on link ChG-Miner_miner-chem-gene.tsv.gz) to download the graph and identify the top 10 drug-drug pairs i.e. pairs of drugs that are similar.

(a) (2 points) Compute pairwise similarity for drugs node pairs (i.e. similarity between nodes ***u*** and ***v*** where both are drugs nodes) based on common neighbors. You may use one of the well known methods such as Jaccard, cosine etc. similarity measures.

(b) (3 points) Compute pairwise node similarity using the Sim-rank algorithm Compare and contrast the results using the two similarity measures (Jaccard/cosine etc. and Sim-Rank). Also try to interpret the similarity results from the two similarity measures.

Note : *Node similarity using the Sim-Rank approach is available in NetworkX library. For the second part of the question, you may use either the general version implemented in the software you choose*

References:

https://networkx.org/documentation/networkx-1.7/tutorial/tutorial.html

http://snap.stanford.edu/snappy/index.html