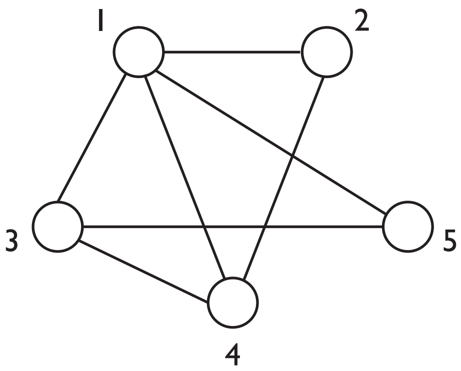
Homework 1
COMP 572: Bioinformatics: Networks

Please submit a zip file with a pdf of all your answers (including figures), as well as your Jupyter notebook (in Python 3) or R markdown file for the programming components of the assignment. A friendly reminder with respect to the collaboration policy for this course, as stated in the syllabus: You are allowed to work in groups on the homework. However, each student must write up their own homework report (code and answers), and the write-up should mention each person you talked to about the homework.

1 Theoretical part

1.1 A simple graph (*25 pts*)



(i) Write down the adjacency matrix representation of this graph. (*3 pts*)

(ii) Calculate the following for each vertex in the graph: (*7 pts*)

- degree
- eigenvector centrality
- betweenness centrality

(iii) Discuss what information each type of centrality you calculated in (ii) gives you about the individual vertices and how they differ. (*5 pts*)

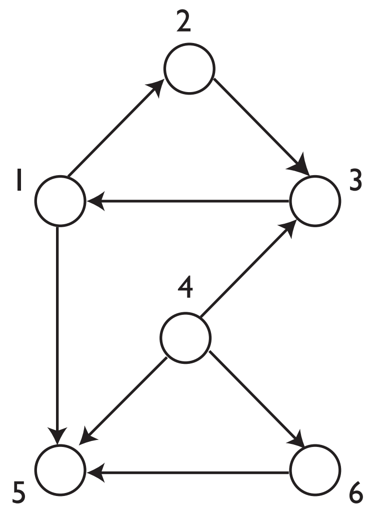
(iv) Draw a bipartite graph based on the above graph, where you are only allowed to remove edges, given the following partition of vertices: $\{\{1,2,3\},\{4,5\}\}$. (*10 pts*)

(a) Draw the one-mode projections of this bipartite graph.

(b) Recompute the betweenness centralities for the new graph.

(c) Is it possible to derive a generalized solution for the betweenness centralities of complete bipartite graphs? If yes, please provide the mathematical derivation and result.

1.2 A directed graph (*25 pts*)



(i) Provide the following representations for the graph: (*5 pts*)

- adjacency matrix
- adjacency list
- edge list

(ii) A^k, where $k \in \mathbb{N}$, has interesting properties that relate to path lengths. (*5 pts*)

(a) Calculate A^2 using the adjacency matrix A you wrote down above.

(b) Draw the graph representation of A^2 if this were the adjacency matrix.

(c) Comment on what each element of A^2 represents.

(iii) Now suppose this graph is undirected. (*15 pts*)

(a) Calculate A^2 and comment specifically on what A^2_{ii}, $i \in V$ denotes.

(b) Use this result to prove that $\sum_{i=1}^{n} k_i = 2m$ for undirected graphs, where k_i is the degree of node i, and m is the total number of edges in the graph. (Use only the A^2 matrix for this proof.)

(c) Often, in biological networks, it is important to identify local network motifs. Use the concept derived above to prove the following:

$$\#\triangle = \frac{1}{6}\mathrm{tr}(A^3),$$

where tr is the matrix trace operation (sum of elements along the diagonal).

2 Programming part

Please use Jupyter notebooks (Python 3 only) or R markdown for the programming parts of the assignment. Remember to put axes labels for any plots you generate, and also comment your code and make sure everything runs without errors before submitting! Also note that you are allowed to use packages, including ones that can handle network data, in python / R for the programming part. For example, people commonly use networkx in python and igraph in R.

2.1 Exploratory analysis of a human PPI network (*50 pts*)

There are several major data repositories that catalog discovered PPIs (predominantly by expert curation of biomedical literature). One such database is www.thebiogrid.org, which typically releases updates every month. It is always important to make note of which release your analyses are based off of, since these are changing over time.

1. Download the human (*Homo sapiens*) interactions from release 4.4.205 (the most current release at the time this homework was written), and put it into a representation of your choice. (*5 pts*)

   Note: BioGRID provides the same interactions in multiple formats, and after parsing, you should come to the same answers regardless of data format. Genes also have multiple different identifiers. For the purposes of our exploratory analysis here, use the 'Official Symbol.'

2. The BioGRID dataset you downloaded provides both *physical* and *genetic* interaction information. We will be looking at only the physical interactions for the purposes of this analysis, and furthermore, be considering whether the interactions come from *low throughput* or *high throughput* experimental assays (you can ignore any edges that are ambiguously annotated).

   Compile a table for these 3 networks (low throughput only, high throughput only, all physical interactions) with the following information: (*15 pts*)

   - Number of nodes
   - Number of edges (interactions)
   - Number of different publications that these interactions were curated from
   - Number of connected components
   - Network density
   - Average degree
   - Protein(s) with the highest degree (including both the protein name and the degree)
   - Protein(s) with degree ≥ 3 that have the highest clustering coefficient (including both the protein name and clustering coefficient)

   Note: Self-loops should not be considered, and duplicated edges should be collapsed. Think carefully about whether this is a *directed* or *undirected* network.

3. Plot the degree distributions of these 3 networks in the same plot. Are they similar or different? Discuss and provide potential explanations. (*10 pts*)

4. In network science literature, there are many centrality measures that have been introduced to determine important nodes in a protein-protein interaction graph, importance here meaning hubs,

essential proteins, protein-bridges, etc. Here, we will examine the node centrality profiles of the low throughput protein-protein interaction network, in order to detect how proteins are clustered based on these measures. Since we do not have any particular labels to assess the importance of each centrality measure, we will instead use dimensionality reduction techniques to study the network. (*20 pts*)

(a) For each node of the low throughput network, calculate:

- degree centrality
- eigenvector centrality
- PageRank
- harmonic centrality
- betweenness centrality

Discuss why we choose the harmonic centrality variant of the closeness centrality measure for this particular network.

(b) Compute the 5x5 Pearson's correlation coefficient matrix. Comment on the values of the matrix. Plot the scatterplot of the harmonic centrality and another centrality measure of your liking. What do you observe? How does it contrast to the correlation value for this particular pair of centrality measures?

(c) Perform Principal Component Analysis (PCA) on the node centrality measures. Plot the scree plot and the cumulative scree plot. How much variance is explained by the first 2 PCs in each case?

Note: Recall that it is important to perform standardization (scaling / normalization) of variables before PCA, since PCA can be sensitive to variance.

(d) Study the loadings (coefficients) of the first 2 PCs gained from the PCA Analysis. Based on the loadings, can you guess what each PC pays attention to? Based on the loadings, the correlation analysis and the scatter plot, roughly, in how many categories would you classify the nodes in the network, based on the centrality measures chosen?