1) **Giant component in the Erdős-Renyi \((G_{n,p})\) random graph.** Consider a \(G_{n,p}\) random graph. In this problem we will study conditions for the emergence of a giant component, that is a connected component whose size grows in proportion to \(n\). To that end, denote by \(u\) the average fraction of vertices in the random graph that do not belong to the giant component. Thus if there is no giant component in the graph we will have \(u = 1\), and if there is a giant component then \(u < 1\). Equivalently, one can think of \(u\) as the probability that a randomly chosen vertex in the graph does not belong to the giant component.

A) For a vertex \(i\) not to belong to the giant component it must not be connected to the giant component via any other vertex. That means that for any other vertex \(j \neq i \in V\), either (a) \(i\) is not connected to \(j\) by an edge, or, (b) \(i\) is connected to \(j\) but \(j\) itself is not a member of the giant component. Using this line of reasoning, show that the probability of not being connected to the giant component via vertex \(j\) is \(1 - p + pu\).

B) Argue that the probability \(u\) of not being connected to the giant component via any of the \(n - 1\) other vertices in the network is implicitly defined by the expression

\[
u = \left[1 - \frac{\bar{d}}{n-1}(1-u)\right]^{n-1}
\]

where \(\bar{d} = \mathbb{E}[d_i] = (n-1)p\) is the average degree of the \(G_{n,p}\) graph.

C) Using that \(\ln(1-x) \simeq -x\) as \(x \to 0\), show that as \(n \to \infty\) then \(\ln u = -\bar{d}(1-u)\); or equivalently

\[
u = e^{-\bar{d}(1-u)}, \quad \text{as } n \to \infty.
\]

D) Let \(s = 1 - u\) be the probability that a randomly chosen vertex in the graph belongs to the giant component. The result in part C) implies that

\[
s = 1 - e^{-\bar{d}s}.
\]

Unfortunately, while a very simple relationship there is no closed-form solution for \(s\). Still, you can get a feeling for its behavior and its solution(s) by superimposing plots of \(f_1(s) = s\) and \(f_2(s) = 1 - e^{-\bar{d}s}\). For the range \(0 \leq s \leq 1\) (recall \(s\) is a probability), use Matlab to generate such superimposed plots of \(f_1(s)\) and \(f_2(s)\), for \(\bar{d} = 0.5, \bar{d} = 1,\) and \(\bar{d} = 1.5\). How can you identify the solution(s) to (1) from your plot? Comment on the number of solutions as \(\bar{d}\) changes.

E) From your study of the plots in part D) or otherwise, show that the value \(\bar{d}^*\) given by

\[
\frac{\partial(1 - e^{-\bar{d}s})}{\partial s} \bigg|_{s=0} = 1
\]

defines a transition between two regimes, where the number of solutions of (1) changes. Hence conclude that for \(\bar{d} \leq \bar{d}^*\) there is no giant component, while the random graph can have a giant component only if \(\bar{d} > \bar{d}^*\).

2) **Partial correlations.** Recall that for \(S_m = \{k_1, \ldots, k_m\}\), we defined the partial correlation of vertex attributes \(X_i\) and \(X_j\), adjusting for \(X_{S_m} = [X_{k_1}, \ldots, X_{k_m}]^\top\), as

\[
\rho_{ij|m} = \frac{\sigma_{ij|m}}{\sqrt{\sigma_{ii|m}\sigma_{jj|m}}}.
\]
Here, $\sigma_{ii|S_m}, \sigma_{jj|S_m}$ and $\sigma_{ij|S_m}$ are the diagonal and off-diagonal elements, respectively, of the $2 \times 2$ partial covariance matrix

$$\Sigma_{11|2} := \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, \tag{4}$$

where $\Sigma_{11}, \Sigma_{22},$ and $\Sigma_{12} = \Sigma_{21}^\top$ are defined through the partitioned covariance matrix

$$\text{cov} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \text{where } W_1 = [X_i, X_j]^\top \text{ and } W_2 = X_{S_m}. \tag{5}$$

Note that for $m = 0$, partial correlations reduce to the standard Pearson correlations.

For the partial correlations in (3), there exist recursive relationships among coefficients at adjacent orders $m - 1$ and $m$ that allow for their efficient calculation. In this problem you will derive this relationship for the case of $m = 1$ and three random variables, say $X, Y$ and $Z$. Specifically, you are asked to derive an expression relating $\rho_{XY|Z}$ to the Pearson correlations $\rho_{XY}, \rho_{XZ}$, and $\rho_{YZ}$.

A) For convenience, we express the covariance matrix $\Sigma$ of $[X, Y, Z]^\top$ in block standardized form as

$$\Sigma = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}. \tag{6}$$

Here the horizontal and vertical lines indicate block partitioning as in (5). Use (4) to derive a closed-form expression for the partial covariance matrix $\Sigma_{XY|Z}$.

B) Combine the elements of the matrix you obtained in part A) according to (3), to show that

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ} \rho_{YZ}}{\sqrt{(1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2)}}. \tag{7}$$

3) Regression-based inference of Gaussian graphical models. Let $[X, Y, Z]^T$ be a Gaussian random vector (e.g., comprising attributes in 3 vertices of a network graph) with zero mean and covariance matrix $\Sigma$, where $\Sigma$ is defined in (6). Consider the task of optimally predicting $Z$ as a linear combination of $X$ and $Y$, using mean-squared error (MSE) as criterion. That is, consider the MSE minimization problem

$$\min_{\beta_{ZX}, \beta_{ZY}} \mathbb{E} \left[ (Z - \beta_{ZX} X - \beta_{ZY} Y)^2 \right]. \tag{8}$$

A) Show that under our model assumptions, the cost function in (7) takes the form

$$1 + \beta_{ZX}^2 + \beta_{ZY}^2 - 2\beta_{ZX} \rho_{Z} - 2\beta_{ZY} \rho_{Y} + 2\beta_{ZX} \beta_{ZY} \rho_{XY}. \tag{9}$$

(Hint: What are the entries of the covariance matrix $\Sigma$?)

B) Differentiating (8) with respect to $\beta_{ZX}$ and $\beta_{ZY}$ and setting the resulting expressions equal to zero, show that the optimum predictor $[\hat{\beta}_{ZX}, \hat{\beta}_{ZY}]^\top$ satisfies the following linear system of equations

$$\begin{pmatrix} \rho_{ZX} \\ \rho_{ZY} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{XY} \\ \rho_{YX} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_{ZX} \\ \hat{\beta}_{ZY} \end{pmatrix}. \tag{10}$$

C) Solve (9) to show that

$$\hat{\beta}_{ZX} = \frac{\rho_{ZX} - \rho_{ZY} \rho_{XY}}{1 - \rho_{XY}^2} \quad \text{and} \quad \hat{\beta}_{ZY} = \frac{\rho_{ZX} \rho_{XY} - \rho_{XY} \rho_{Z} \rho_{Y}}{1 - \rho_{XY}^2}. \tag{11}$$

Using your results in Problem 2), compare this solution with the partial correlations $\rho_{ZX|Y}$ and $\rho_{ZY|X}$. Hence argue that $\hat{\beta}_{ZX} = 0$ if and only if $\rho_{ZX|Y} = 0$, and likewise for $\hat{\beta}_{ZY}$. 
4) **Partitioning a network of US political blogs.** In this problem we will study a network of Internet blogs on the subject of US politics, with the goal of partitioning the graph into liberal (i.e., Democrat) and conservative (i.e., Republican) ‘blogger communities’.

Data on the 2004 US Election’s political blogosphere was compiled by L. Adamic and N. Glance in 2005, and can be obtained e.g., from Mark Newman’s network collection at [http://www-personal.umich.edu/~mejn/netdata/](http://www-personal.umich.edu/~mejn/netdata/). Here we use an undirected version of the original directed graph, where edges correspond to hyperlinks between blogs. The network comprises $N_v = 1490$ blogs (vertices), and a binary attribute associated to each vertex indicates political leaning according to: 0 (liberal) and 1 (conservative).

A) Download the Matlab version of the political blogosphere dataset political_blogs.mat from the class website, or directly from:

[http://www.ece.rochester.edu/~gmateosb/ECE442/Homework/political_blogs.mat](http://www.ece.rochester.edu/~gmateosb/ECE442/Homework/political_blogs.mat)

Load the dataset in Matlab. Matrix $A \in \{0, 1\}^{1490 \times 1490}$ is the graph’s adjacency matrix, whereas the vector $\text{nodes} \in \{0, 1\}^{1490}$ contains the binary vertex attributes indicating political leaning of each blog.

B) Compute the degrees $d_v$ of all nodes $v \in V$ and the total number of edges $N_e$. Hence compute the modularity matrix $B$ for the network.

C) Implement the spectral modularity maximization algorithm to perform a bisection of the graph, i.e., split the vertices into two communities.

D) Use the ground-truth community membership in the vector $\text{nodes}$ to determine the percentage of correctly classified blogs.

E) *(Extra credit, try it out if time allows.)* Compute the graph Laplacian $L$. Repeat part D) after bisecting the graph using the spectral algorithm we derived to solve the relaxed ratio-cut minimization problem.