[Problem 1 – Separation and factorization in MRF – NOT GRADED]

a. Consider the graph below. Which of the following separation relationships hold in this graph? If a separation is not true, give an open path.

\[
\begin{align*}
BC & \perp E \mid A \\
A & \perp B \\
B & \perp F \mid CE
\end{align*}
\]

b. Write the general factored expression of a joint distribution \( P_V \) over \( V = \{A, B, \ldots, F\} \) that factors according to \( G \).

c. Choose one of the true statements in a. Prove that the corresponding independence statement holds in \( P_V \) by using the factored expression of \( P_V \) from b.

d. We say that graph \( G = (V, E) \) an I-map for a distribution \( P_V \) iff any separation (or D-separation) that is true in \( G \) represents an independence relation that holds in \( P_V \). In other words, what is asked below is if \( P_V \) factors according to \( G \).

Is the graph \( G \) an I-map for the following distribution?

\[
P_V = \exp[w_{ABC}f_{ABC} + w_{AD}f_{AD} + w_{EF}f_{EF}]
\]  

(1)

where \( w_X \) are real parameters and \( f_X \) are functions of the variables in \( X \subseteq V \). Explain why yes or no.

e. Is the graph \( G \) an I-map for the following distribution?

\[
P_V = \exp[w_{ABC}f_{ABC} + w_{ABD}f_{ABD} + w_{AED}f_{AED} + w_{ACD}f_{ACD} + w_{CE}f_{CE} + w_{EF}f_{EF}]
\]  

(2)

where \( w_X \) are real parameters and \( f_X \) are functions of the variables in \( X \subseteq V \). Explain why yes or no.

[Problem 2 - More I-maps – NOT GRADED]

This exercise exemplifies how directed and undirected graphical representations are not equivalent.

Let \( D_1 \) and \( U_2 \) be the graphs below
a. Write all the independencies in $D_1$. List all the V-structures in $D_1$.

b. Write all the independencies in $U_2$.

c. Draw an undirected graph $U_1$ which is an I-map of $D_1$. In other words, $U_1$ must capture as many as possible of the independencies in $D_1$, and have no extra independence relation besides those in $D_1$.

List all the independencies in $D_1$ which are not in $U_1$.

d. Draw a directed graph $D_2$ which is an I-map of $U_2$. List all the independencies in $U_2$ which are not in $D_2$. List all the V-structures in $D_2$.

e. Draw two graphs $U_3, D_3$ on the five nodes $A, B, C, D, E$, so that the first graph is undirected, the second graph is directed, and the two graphs have the same list of independence relations. (That is, the two graphs are perfect maps of each other).

Try to find the most interesting, non-trivial example you can. In particular, the graphs cannot be: all disconnected, fully connected, or chains.

**Problem 3 - D-separation – NOT GRADED**

a. Are there any converging arrows on the path $ABCG$? Enumerate them.

Are there any converging arrows on the path $HCFEI$? Enumerate them.

Enumerate all the V-structures at node $C$. 

2
b. Which of the following independence statements is true? For each false statement, give an open path.

\[
\begin{align*}
B & \perp H \mid E \\
B & \perp I \mid A,D,E \\
A & \perp E \mid G \\
C & \perp J \\
A & \perp I \\
E & \perp G \mid C
\end{align*}
\]

c. Find all edges that can be inverted, so that the newly obtained DAG is equivalent to this DAG?

d. Write the expression of a joint distribution \( P_V \) over \( V \) that factors according to this graph.

3. Verify that \( B \perp I \mid EH \) in \( P_V \) using marginalization in the factored form of \( P_V \) and the factorization lemma.

**Problem 4 – A graph of tree-width 2**

The graph \( G \) below has **treewidth 2**. The treewidth of a graph is one less than the size of the maximum clique in that graph.

![Graph G](image)

a. Find an orientation for the graph \( G \), which produces no V-structures. Denote the resulting DAG by \( G' \). (Be careful!)

b. Write the general factored form of a distribution \( P \) for which the undirected graph is an I-map.

c. Write the general factored form of a distribution \( P' \) of which the DAG you found in question a. is an I-map.

d. The two factorizations in b,c must be equal. Find a way to group the factors in \( P' \) to obtain the factorization in \( P \). The grouping may not be unique.

Using the grouping you found, show that the potentials of \( P \) have a probabilistic interpretation. Since there will be many potentials, it is sufficient to find a probabilistic interpretation for one potential containing variable \( A \) and for one containing the variable \( E \).

e. Show step by step the variable elimination procedure for computing \( P_{A,C=0,D=0,E=0} \) using the factorization of \( P \) obtained in c. Choose an elimination ordering that does not create new factors (a.k.a potentials) involving more than 2 variables.
f. Each potential represents a (conditional) probability table; find the probabilistic semantics for each of the potentials created during this elimination. Some of the eliminations will lead to potentials equal to constants. Be sure to identify these (so, in practice, these eliminations can be skipped). Explain why and when this happens.

g. Now show how to get \( P_{A|C=0, D=0, E=0} \) using the Sum-Product algorithm.

**Problem 5 – Exponential Random Graph Models**

The file `stat-authors-last-all.txt` contains the co-authorship network extracted from the UW Statistics Technical Reports published between 1 January, 2000 and January 27, 2015. Each row corresponds to a technical report, and lists the authors of the report by last name, separated by a space, with the most recent report first. For example, the row `Scrucca Raftery` corresponds to TR 629 by Luca Scrucca and Adrian Raftery.

The other files posted are obtained from this list, e.g. an index file where each author is assigned a number, and a file where each row represents a report with the authors encoded by numbers.

There is also a list `stat-faculty.txt` that lists essentially who of these authors are or were faculty in Statistics. We will consider that the authors belong to one of two clusters: faculty or non-faculty.

Your task is to read these data and build the co-authorship network that it represents.

a. Construct the graph representation of the data, where \( V = \{1, 2, \ldots, n\} \) is the set of authors \( n = 191 \) is given in the author index file. Edge \( Y_{ij} = 1 \) iff \( i \) and \( j \) are among the authors of at least one technical report. For example, the set of authors 12 13 2 appears more than once in the data; in the graph, \( Y_{12,13} = Y_{12,2} = Y_{13,2} = 1 \).

Optional but highly recommended: plot the graph! Extra credit will be given if the plot is nice and legible.

Make a graph of the node degrees in this network. (The degree of a node is the number of edges incident to the node. Sort the degrees from highest to lowest. Color or mark the faculty nodes distinctly from the non-faculty nodes. Label the three nodes with the highest degree. Also, give the number of leaves (nodes of degree 1).

Answer: is the graph connected? if not list the two nodes of highest degree\(^1\) in each of the connected components.

b. Model this network as an Erdos-Renyi graph. Show briefly but rigorously\(^2\) how you estimate the parameter \( \theta \) (by the Maximum Likelihood method). What is the numerical value of \( \theta \) that you obtain?

c. Derive the expression for the average degree of a node in an ER model, and compute this value for the co-authorship network and the \( \theta \) obtained in b.

d. Derive the expression for the expected number of edges in \( G \) under the ER model, and

---

\(^1\)Unless a component is a singleton node!

\(^2\)This means: give an algorithm in pseudocode and mathematical equations. Explain briefly where the equations come from. In general, everything that is “mainstream statistics or probability or math” need not be proved. For an example to the point, you don’t need to prove what is the ML estimate of the parameter of a Bernoulli distribution, but you need to state the formula.
compute this value for the co-authorship network and the $\theta$ obtained in b.

e. For a single node $i$, derive the expression and compute the CDF of the distribution of its degree $d_i$ in the ER model (i.e. compute $P_\theta(d_i \leq m)$, for $m = 0, 1, \ldots, n$.

Plot this CDF on the same graph as the empirical CDF obtained from the actual node degrees.

f. Now consider that the data come from a SBM model, where $\beta_{kl}, k, l = (0, 0), (1, 1), (0, 1)$ represents the log-odds of an edge between cluster $k$ and cluster $l$, where 0 denotes non-faculty and 1 denotes faculty. Show briefly but rigorously how you estimate the parameters $\beta_{kl}$. What are the numerical values that you obtain?

c'. Repeat question e for this model: derive and compute the average degrees for $d_1, d_0$ for faculty and non-faculty respectively, and compare with the empirical values.

e'. Repeat question e for this model: derive and compute the CDF of $d_i$ when $i$ is of class 0 and 1, respectively. Compare with the empirical CDF's for these quantities. Show these plots on the same graph.

g. Now consider a subgraph $G'$ of $G$. We will take this subgraph to be the collaboration network for the period 2000-2007 (i.e TR numbers smaller and equal to TR 523 – Handcock Gile). The “old” data starts at line 103 in stat-authors-all.txt and stat-authors-all.dat.

Recalculate all the numerical estimates from questions a–e’ and make the corresponding plots for this network.

Compare with the results for $G$. Are the parameter estimates the same?

Extra Credit: use statnet or another software and fit an ERGM model to the two networks. Compare the model predictions with the actual data.