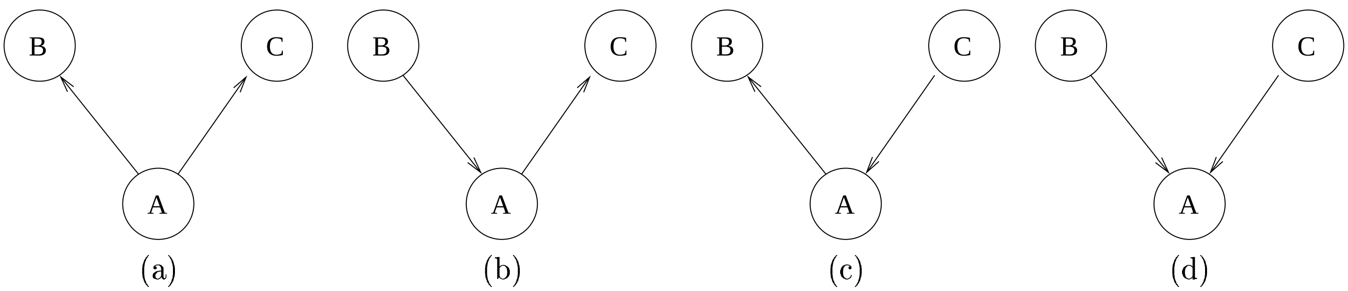
CMSC 828G Homework 3

Data Mining,CMSC 828G, Spring 2002

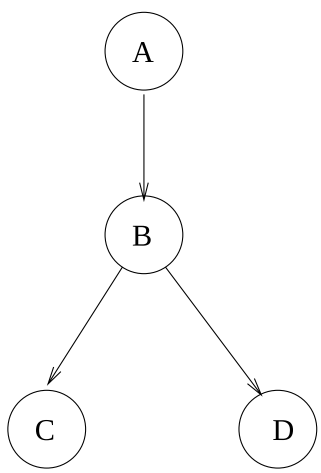
Due Date: Tuesday, April 16, at the start of class

1. Given two Bayesian networks B_1 and B_2 over the same set of random variables, B_1 and B_2 are *I-equivalent* if they encode the same set of d-seperation properties.

   (a) In the figure below, which of the four networks are I-equivalent?

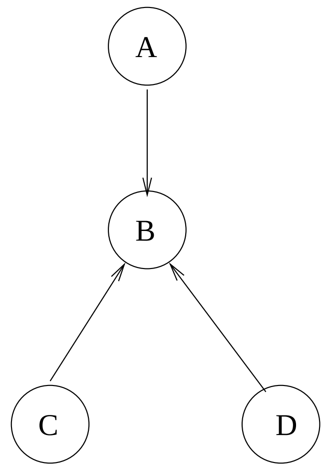


   (b) In the figure below, enumerate all of the conditional independences represented by the network. Each independence should be of the form $I(X, Y \mid \mathbf{Z})$ where X and Y are single variables and $\mathbf{Z}$ is a (possibly empty) set of variables. (If your list contains $I(X, Y \mid \mathbf{Z})$ then it need not contain $I(Y, X \mid \mathbf{Z})$.



   (c) Is there another I-equivalent network to this network? If so, draw it. If not, **briefly** explain why there is no other network over these four variables that can be equivalent (your high-level reasoning should be based on the independencies in the network).

(d) In the figure below, enumerate all of the conditional independences represented by the network.

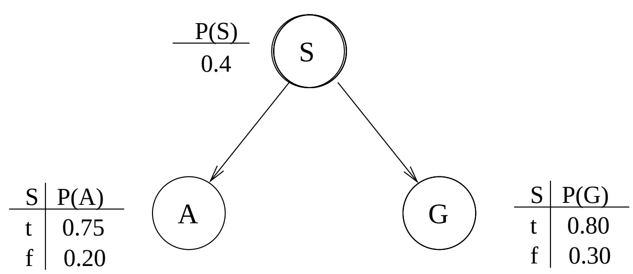


(e) Is there another I-equivalent network to this network? If so, draw it. If not, **briefly** explain why there is no other network over these four variables that can be equivalent (your high-level reasoning should be based on the independencies in the network).

(f) For the following property, either prove (the proof should hold for any probability distribution) or disprove with a counterexample (a counter example may be described by a Bayesian network).

$[I(X;Y|Z) \wedge I(X;Y|W)] \rightarrow I(X;Y|Z,W)$

2. A cmsc828g student notices that people that drive SUVs (S) consume large amounts of gas (G) and are involved in more accidents than the national average (A). They have constructed the following Bayesian network:



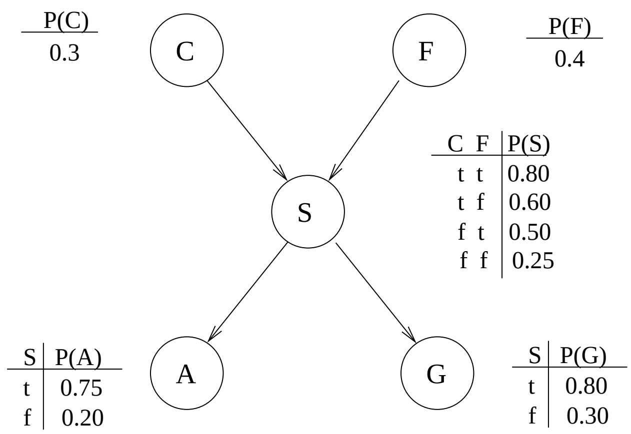
(a) Compute $P(A)$ in two ways:

i. By generating the *entire joint distribution* over these variables and explicitly summing the appropriate entries.

ii. Using the variable elimination algorithm

(b) Using conditional independence, compute $P(\bar{g}, a \mid s)$ and $P(\bar{g}, a \mid \bar{s})$. Then use Bayes rule to compute $P(s|\bar{g}, a)$.

(c) The enterprising cmsc828g student notices that there are two types of people that drive SUVs, people from California (C) and people with large families (F). After collecting some statistics, the student arrives at the following form for the Bayesian network:

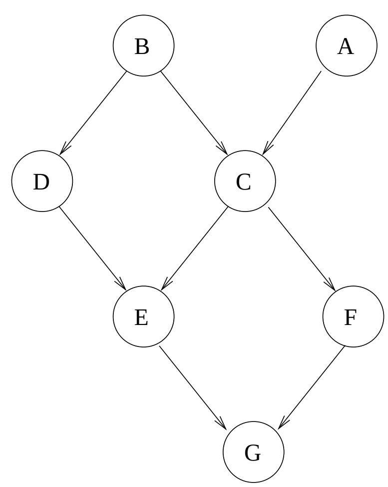


Using the chain rule from probability compute the probability $P(\bar{g}, a, s, c, \bar{f})$.

(d) Without explicitly generating the entire probability distribution, compute $P(s)$ and $P(s \mid c)$ and $P(s \mid \bar{c})$. (Hint: consider different values of F.)

(e) Show how you can compute the value of $P(G)$ in the complete distribution without doing any additional work (i.e., based solely on the work you have already done). Explain.

3. Given the following Bayesian network:



(a) Suppose we have the query P(G) and the following elimination ordering: first A, then B, C, D, E, F. Show what the variable elimination algorithm will do on this graph by listing for each variable eliminated, the new factor generated and the factors used to generate the new factor. The factors should be written in the form $f_i(X_1, \ldots, X_k)$ for some set of variables $X_1, \ldots, X_k$.

(b) Elimination ordering matters to the complexity of inference. Assume all variables have the same domain size $m > 2$. Find an elimination ordering for the same query whose time complexity as a function of m is worse than the ordering given above. As in the above problem, show what the factors generated at each step.

(c) Sketch a *s*imple BN over n random variables where there is one elimination ordering for a query $P(X)$ where the complexity of inference is linear in n and another elimination ordering for the same query in which the complexity of inference is exponential in n. Show the network and describe the different orderings.

4. In this problem, you are asked to learn and explore a Bayesian network. The steps involved include:

- Choose a Bayesian network learning tool. Three possibilities are:
  - Data Digest BN 5, `http://www.data-digest.com/home.html`
  - WinMin, `http://research.microsoft.com/~dmax/WinMine/tooldoc.htm`
  - B-course, `http://b-course.cs.helsinki.fi/`

  You may choose another tool, as long as it supports the operations required to complete the question.

- Choose a data set. You may use the data set for your project (or some subset), or a data set you have found on the web. Please do not use a data set that is included in the demo for the tool you are using (however, you may use a data set that is in the demo for one of the other tools!)

(a) Learn a Bayesian network. Turn in a printout of the Bayesian network learned.

(b) See if you can find a better scoring Bayesian network by either adding in constraints or modifying the Bayesian network by hand. Describe the changes that you made and what effect they had on the learned network. Turn in a copy of the best scoring network that you could find.

(c) Explore the Bayesian network that you have learned. Describe 4-5 inferences that you can make in the model. Describe both quantitatively, illustrate quantitatively and evaluate whether this seems like a reasonable relationship to you or whether the relationship seems dubious. For example, if I have a medical data set, which includes a random variable for patient age and whether or not the patient has had a stroke, one inference I may be able to make in the learned model is the following:

qualitative: if the patient is older, the probability of stroke is higher heart.
quantitative: if the patient is aged 25-30, the probability of stroke is 0.001; if the patient is aged 65-70, the probability of stroke is 0.02.
evaluation: this seems like a reasonable probabilistic relationship between the random variables.

At least two of your inferences should be between nodes that are node directly connected.

(d) Mail an electronic version of your best Bayesian network, along with the name of the tool that you used to construct it, to the TA, eiman@cs.umd.edu.