CS 476/676, Spring 2016 Problem Set #2a (Due by 11:59pm on Monday, Mar. 7)

---

1 Instructions

This assignment contains a few Bayesian network problems, continued from Homework 1b.

1.1 What to Hand In

All of your submission files should be handed in as a single pdf named `hw2a-username1-username2.pdf`, where the `usernames` have been replaced with the JHED IDs of your group members. (Your group is allowed to have between 2 members; see Section 1.2 for details on collaboration.)

Hand in the pdf file by creating a private note to the instructors on Piazza with the title *Submission 2a from* `<list of names of team members>`, and attaching your pdf file to that note. The note should be submitted to the `submission2a` folder.

1.2 Submission Policies

Please note the following:

- **Collaboration:** Please work in groups of size 2 people. The homeworks are a way for you to work through the material you're learning in this class on your own. But, by working in a group, and debugging each other's solutions, you'll have a chance to learn the material in more depth. The recommended format for tackling these problem sets is the following. Write a high level sketch of the solution for all of the problems on your own. Meet as a group to brainstorm your solutions and converge on a solution as a group. It is important that you have a good understanding of how you'd have approached the problem independently before discussing your solution with the other group members. Developing this intuition will serve you well in the final exam where you will be required to work on your own. Persuant to your group meeting, write up the solutions on your own. Thereafter, meet as a group to clean up and submit a final write up as a group. By now, each of you should have a solid understand of the concepts involved, and by meeting as a group, you've had a chance to see common ways in which one can make mistakes. Submit your final solution as a final writeup for the group. Your submission should include the names of every team member. Also, name your file as `hw2a-username1-username2.pdf`.
- **Late Submissions:** We allow each student to use up to 3 late days over the semester. You have late days, not late hours. This means that if your submission is late by any amount of time past the deadline, then this will use up a late day. If it is late by any amount beyond 24 hours past the deadline, then this will use a second late. **If you jointly submit an assignment as a team, then every team member will lose late days if the assignment is submitted late.** If you collaborate with team members but independently submit your own version, then late hours will only apply to you.

2 Bayesian Network Problems (Continued from HW1b) [48 points]

This section contains some questions on Bayesian networks, in the style of 2.1 and 2.2 on Homework 1b. Please refer back to that assignment for details on the formatting of the `network` and `cpd` files.

2.1 Network Manipulation [20 points]

One operation on Bayesian networks that arises in many settings is the marginalization of some node in the network. Consider the network below, which models the flu prevalence in a population. (This is a different network than what you designed in Homework 1b.)

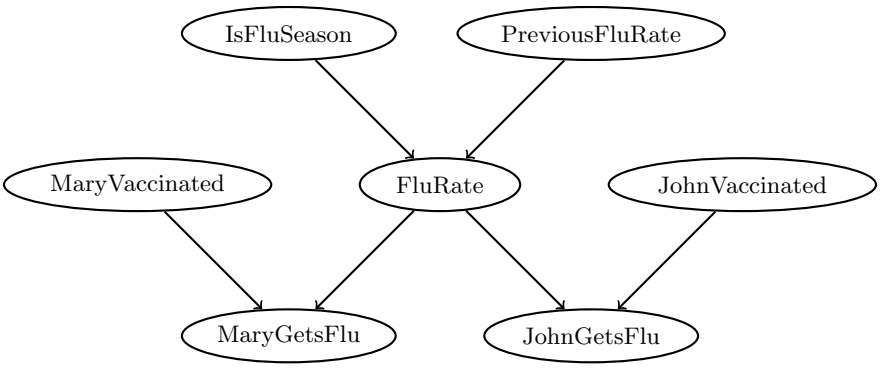


Figure 1: Network $\mathcal{F}$.

2.1.1 Deliverables [14 points]

Draw a new network that is a minimal I-map over the marginal distributions $P_{\mathcal{F}}$(IsFluSeason, PreviousFluRate, MaryVaccinated, JohnVaccinated, MaryGetsFlu, JohnGetsFlu). Be sure to maintain all dependencies in the original graph.

2.1.2 Analytical Questions [6 points]

Generalize the above procedure to come up with a node-elimination algorithm for Bayesian networks. Given a graph $\mathcal{G}$ and a node to eliminate E, the algorithm should produce a new graph $\mathcal{G}'$ that is a minimal I-map over the marginal distributions $P_{\mathcal{G}}(\mathcal{X} - \{E\})$.

2.2 Network Queries [16 points]

Let's consider the sensitivity of a particular query $P(X|\boldsymbol{Y})$ to the CPD of a particular node Z. Let X and Z be nodes (which are not directly connected) and $\boldsymbol{Y}$ be a set of nodes. We say that Z has a *requisite CPD* for answering the query $P(X|\boldsymbol{Y})$ if there are two networks $\mathcal{B}_1$ and $\mathcal{B}_2$ that have identical graph structure $\mathcal{G}$ and identical CPDs everywhere except at the node Z, and where $P_{\mathcal{B}_1}(X|\boldsymbol{Y}) \neq P_{\mathcal{B}_2}(X|\boldsymbol{Y})$; in other words, the CPD of Z affects the answer to this query.

This type of analysis is useful in various settings, including determining which CPDs we need to acquire for a certain query.

Suppose we modify $\mathcal{G}$ into a graph $\mathcal{G}'$ which is identical to $\mathcal{G}$ except it contains a new "dummy" node $\hat{Z}$ which is a parent of Z (thereby altering the CPD of Z). One way to test whether Z is a requisite probability node for $P(X|\boldsymbol{Y})$ is to test whether $\hat{Z}$ has an active trail to X given $\boldsymbol{Y}$ in $\mathcal{G}'$. If so, then altering Z's CPD can affect $P(X|\boldsymbol{Y})$. If not, then altering Z's CPD cannot affect $P(X|\boldsymbol{Y})$.

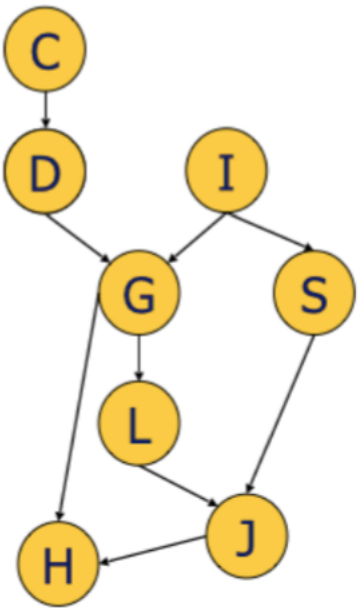
2.2.1 Analytical Questions

1. **[8 points]** Prove that, if there is an active trail from $\hat{Z}$ to X given $\boldsymbol{Y}$, then altering Z's CPD can affect $P(X|\boldsymbol{Y})$.

2. **[8 points]** Prove that, if there is no active trail from $\hat{Z}$ to X given $\boldsymbol{Y}$, then altering Z's CPD cannot affect $P(X|\boldsymbol{Y})$.

2.3 Variable Elimination [12 points]

Here is the Bayesian Network we discussed in class.



1. **[8 points]** For the two variable elimination orderings given below, provide a valid clique tree. Reason that your resulting clique trees are valid.

   Assume {J, S} are your query nodes.

   a. C-D-I-H-G-L
   b. G-L-C-D-I-H

2. **[4 points]** Consider the sub-network over the variables, {C,D,G,I}. Here, we can compute $P(G)$ as $\sum_{C,D,I} P(C)P(D|C)P(G|I,D)P(I)$. This can be written in its more efficient form as $\sum_{D,I} P(G|I,D)P(I) \sum_{C} P(C)P(D|C)$. We refer to this form as the sum-product form.

   Now, consider the following (valid) clique tree for inference in the network shown above. Here {G,H,J} is the root node and messages are sent upstream to {G,H,J}. Write down the distribution over the query variables in the sum-product form as defined by the message passing operations shown in this clique tree.