CIS620 – GRAPHICAL MODELS–SPRING 2009

SOLUTIONS TO HOMEWORKS 1-5.
SOME SOLUTIONS ADAPTED FROM D. KOLLER, N. FRIEDMAN AND C. GUESTRIN.

Homework 1

Problem 2.9. Prove or disprove:

1. $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \;\Rightarrow\; (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$.
Proof:

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = \sum_{\mathbf{w}} P(\mathbf{X}, \mathbf{Y}, \mathrm{w}|\mathbf{Z}) = \sum_{\mathbf{w}} P(\mathbf{X}|\mathbf{Z}) P(\mathbf{Y}, \mathrm{w} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Y} \mid \mathbf{Z}).$$

2. $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \;\;\&\;\; (\mathbf{X}, \mathbf{Y} \perp \mathbf{W} \mid \mathbf{Z}) \;\Rightarrow\; (\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z})$.
Proof:

$$\begin{aligned} P(\mathbf{X}, \mathbf{W} \mid \mathbf{Z}) &= \sum_{\mathbf{y}} P(\mathbf{X}, \mathbf{y}, \mathbf{W} \mid \mathbf{Z}) = \sum_{\mathbf{y}} P(\mathbf{X}, \mathbf{y} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}) \\ &= \sum_{\mathbf{y}} P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{y} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}). \end{aligned}$$

3. $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \;\;\&\;\; (\mathbf{Y} \perp \mathbf{W} \mid \mathbf{Z}) \;\Rightarrow\; (\mathbf{X}, \mathbf{W} \perp \mathbf{Y} \mid \mathbf{Z})$.
Proof:

$$\begin{aligned} P(\mathbf{X}, \mathbf{W}, \mathbf{Y} \mid \mathbf{Z}) &= P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{Y} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}). \\ P(\mathbf{X}, \mathbf{W} \mid \mathbf{Z}) &= \sum_{\mathbf{y}} P(\mathbf{X}, \mathbf{W}, \mathbf{y} \mid \mathbf{Z}) = \sum_{\mathbf{y}} P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{W}, \mathbf{y} \mid \mathbf{Z}) \\ &= \sum_{\mathbf{y}} P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}) P(\mathbf{y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{W} \mid \mathbf{Z}). \end{aligned}$$

combining the two parts gives us $P(\mathbf{X}, \mathbf{W}, \mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z}) P(\mathbf{X}, \mathbf{W} \mid \mathbf{Z})$.

4. $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \;\;\&\;\; (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W}) \;\not\Rightarrow\; (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$.
Counter-example: Consider the following BN:

$$X \rightarrow Z \leftarrow U \rightarrow W \leftarrow Y$$

D-separation gives us: $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}), (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W}), \neg(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$, since the path between X and Y is activated when both Z and W are observed.

Problem 3.1. Provide an example of a distribution $P(X_1, X_2, X_3)$ where for each $i \neq j$, we have that $(X_i \perp X_j) \in \mathcal{I}(P)$, but we also have that $(X_1, X_2 \perp X_3) \notin \mathcal{I}(P)$.
Solution: The simplest example uses exclusive-or ($\oplus$). Let $X_1, X_2 \sim \text{Bernoulli}(\frac{1}{2})$ with $X_3 = X_1 \oplus X_2$. It is easy to show that $P(X_i, X_j) = \frac{1}{4} = P(X_i)P(X_j)$ (but needs to be shown with a table). Since X_3 depends deterministically on X_1 and X_2 it cannot be the case that $(X_1, X_2 \perp X_3)$.

Problem 3.4. Positive V-Structure Interactions
A slightly different formulation than the book:

Let X, Y, Z be binary random variables with joint distribution given by the graphical model $X \rightarrow Z \leftarrow Y$. We define the following:

$$a \equiv P(X = t);\ b \equiv P(X = t|Z = t);\ c \equiv P(X = t; |Z = t; Y = t)$$

1. For all the following cases, provide examples of conditional probability tables (CPTs) (and compute the quantities, a, b, c), which make the statements true: (a) $a > c$ (b) $a < c < b$ (c) $b < a < c$.
2. Think of X and Y as causes and Z as a common effect, and for all the above cases summarize (in a sentence or two) why the statements are true for your examples.
Solution

(a). Here is a possible solution:

X	$P(X)$
f	0.5
t	0.5

Y	$P(Y)$
f	0.5
t	0.5

Z	$P(Z \mid X = f, Y = f)$	$P(Z \mid X = t, Y = f)$	$P(Z \mid X = f, Y = t)$	$P(Z \mid X = t, Y = t)$
f	0	0.5	0.5	1
t	1	0.5	0.5	0

$a = P(X = t) = 0.5;\ \ c = P(X = t|Y = t, Z = t) = 0$

Therefore we have $P(X = t) > P(X = t|Y = t, Z = t)$. This model has negative correlation on both edges. That is, if X is true, Z is more likely to be false. This effect therefore propagates backwards as well, if Z is true, X is less likely to be true.

(b). Here is a possible solution:

X	P(X)
f	0.5
t	0.5

Y	P(Y)
f	0.5
t	0.5

Z	$P(Z \mid X = f, Y = f)$	$P(Z \mid X = t, Y = f)$	$P(Z \mid X = f, Y = t)$	$P(Z \mid X = t, Y = t)$
f	1	0.5	0.5	0
t	0	0.5	0.5	1

$a = P(X = t) = 0.5$

$$\begin{aligned}
b = P(X=t|Z=t) &= \frac{P(X=t,Z=t)}{P(Z=t)} \\
&= \frac{P(X=t)\sum_{Y} P(Y)P(Z=t|X=t,Y))}{\sum_{X}\sum_{Y} P(Z=t|X,Y)P(X)P(Y)} \\
&= \frac{0.5\times(0.5\times 0.5+0.5\times 1)}{0.5^3+0.5^3+0.5^2} \\
&= 0.75
\end{aligned}$$

$$\begin{aligned}
c = P(X=t|Y=t,Z=t) &= \frac{P(X=t,Y=t,Z=t)}{P(Y=t,Z=t)} \\
&= \frac{P(X=t)P(Y=t)P(Z=t|X=t,Y=t)}{P(Y=t)\sum_{X} P(X)P(Z=t|X,Y=t)} \\
&= \frac{0.5}{0.5\times 0.5+0.5\times 1} \\
&= 0.6667
\end{aligned}$$

Then we have $a < c < b$

This model has puts positive correlation on both edges. That is, if X is true, Z is more likely to be true (similarly for Y). Therefore, we have $a < b$ since knowing Z is true makes it more likely that X is true (postively correlated). However, if we also know that Y is true, Z is “explained away” by Y, therefore reducing the probability that X is true. Therefore, we have $c < b$. However, since it is still positively correlated, it should never fall below the base probability of X, therefore we have $a < c$

(c). Here is a possible solution:

X	P(X)
f	0.5
t	0.5

Y	P(Y)
f	0.5
t	0.5

Z	$P(Z \mid X=f,Y=f)$	$P(Z \mid X=t,Y=f)$	$P(Z \mid X=f,Y=t)$	$P(Z \mid X=t,Y=t)$
f	0	1	0.9	0
t	1	0	0.1	1

$a = P(X=t) = 0.5$

$$\begin{aligned}
b = P(X=t|Z=t) &= \frac{P(X=t,Z=t)}{P(Z=t)} \\
&= \frac{P(X=t)\sum_{Y} P(Y)P(Z=t|X=t,Y))}{\sum_{X}\sum_{Y} P(Z=t|X,Y)P(X)P(Y)} \\
&= \frac{0.5 \times (0.5 \times 1)}{0.5^2 + 0.5^2 \times 0.1 + 0.5^2} \\
&= \frac{0.25}{0.525} \\
&= 0.476
\end{aligned}$$

$$\begin{aligned}
c = P(X=t|Y=t,Z=t) &= \frac{P(X=t,Y=t,Z=t)}{P(Y=t,Z=t)} \\
&= \frac{P(X=t)P(Y=t)P(Z=t|X=t,Y=t)}{P(Y=t)\sum_{X} P(X)P(Z=t|X,Y=t)} \\
&= \frac{0.5}{0.5 \times 0.1 + 0.5 \times 1} \\
&= 0.909
\end{aligned}$$

Therefore we have $b < a < c$

The correlation mechanism is more complicated here. Basically $Z = X \wedge Y$ (NOT-XOR) except when $X = f, Y = t$ then $Z = X \wedge Y$ with high probability. Now, if $Z = X \wedge Y$, knowing $Y = t$ should not change the probability of X, however the exceptional case ($X = f, Y = t$) increases the probability that X is false. Also, we have $P(X = t|Y = t, Z = t)$ is extremely high as, if Y are true, X should be true to allow Y to be true as $Z = X \wedge Y$.

HOMEWORK 2

Problem 3.11. Marginalization

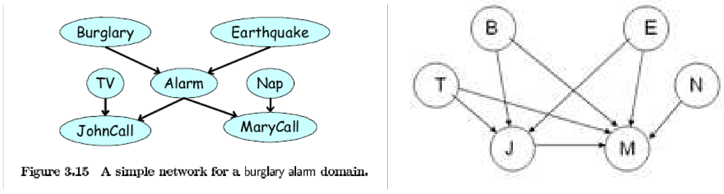

Figure 3.15 A simple network for a burglary alarm domain.

FIGURE 1. Left: original Network G. Right: marginalized Network G'

To construct a minimal I-map for the marginalized network, we should preserve all independencies that exist in G when the variable being marginalized, A, is unobserved. We will see what active trails pass though A and try to preserve them. For B, there is an active trail from it to J, $B \rightarrow A \rightarrow J$, so we must add an edge $B \rightarrow J$ in G', otherwise we are asserting an independence $B \perp J$ that does not exist in G. Similarly, we need to add an edge from each parent to each child of A. There is also an active trail in G between J and M, $J \leftarrow A \rightarrow M$, therefore we must add an edge between J and M (either direction is OK), but we will choose $J \rightarrow M$. Moreover, observing J, there is an active trail from T to M that passes through A, $T \rightarrow J \leftarrow A \rightarrow M$, however, in G', this trail will be blocked when J is observed. Moreover, while there exists an alternative trail in G' that utilizes the new added v-structure between T, B via $(T \rightarrow J \leftarrow B \rightarrow M)$ or the other v-structure T, E via $(T \rightarrow J \leftarrow E \rightarrow M)$, observing both B and E, in addition to J, will block these trails in G' but not in G. Therefore, we must add an edge between T and M, otherwise, we are asserting an independence $(T \perp M|J, B, E)$ that does not exist in G. Moreover, the active trail in G between N and J, when M is observed, is still active in G' due to the new v-structure between N and J. No other edges are needed. (It should be noted that if we had chosen to direct the edge between J and M as $J \leftarrow M$, then we should have added the edge $N \rightarrow J$ instead of $T \rightarrow M$.)

General Case

To construct G', we will simply follow the minimal I-map construction algorithm discussed in class. First, order the variables according to their topological order in G. Without loss of generality, we rename the variables according to their topological order: $X_1, \cdots X_{i-1}, X_i, \cdots X_n$, where X_i is the variable to be marginalized. Note that using the above order, running the minimal I-map algorithm will exactly result in G due to the local Markov assumption. Second, add these variables to G' using the order above, while skipping X_i, and use the independencies that can be read from G about P to select a parent set for each variable. For the variables $X_1 \cdots X_{i-1}$, adding each of them to G' results in the same parent set as in G: that is, $\text{Pa}^{g'}(X_j) = \text{Pa}^{g}(X_j), j < i$. Where $\text{Pa}^{g}(X)$ and $\text{Pa}^{g'}(X)$ is the parent set of variable X in G and G' respectively. In fact, we can generalize the above relationship for any variable X_j such that $X_i \notin \text{Pa}^{g}(X_j)$. That is, $\text{Pa}^{g'}(X_j) = \text{Pa}^{g}(X_j), X_i \notin \text{Pa}^{g}(X_j)$. To see why this is true for $j > i$, if X_i was not selected as a parent for X_j, then $X_i \perp X_j|\text{Pa}^{g}(X_j)$, moreover $X_j \perp X_k|\text{Pa}^{g}(X_j), X_k \in \{X_1 \cdots X_{j-1}\} - \text{Pa}^{g}(X_j)$. Since skipping X_i will not change the previous assertions, the parent set of X_j will remain the same in G' as it still satisfies the local Markov assumption. Therefore, the only variables that will be affected by skipping X_i are those variables that has X_i as a parent in G, i.e., the children on X_i in G.

For each child of X_i, we need to find a new parent set. Lets consider one of these children and call it X_c. X_c must retain its old parent set in G, other than X_i, since from G, we know that $\neg(X_k \perp X_c), X_k \in Pa^{g}(X_c)$. Therefore, we just need to replace X_i with a set of variables that acts as a surrogate for X_i in *parenting* X_c. Using the local Markov assumption, X_i blocks some information flow to X_c, therefore these surrogate variables must block the same set of paths while themselves can not be d-separated from X_c when X_i is not observed. Moreover, these variables must appear in the topological order before X_c. For each parent of X_i in G, X_p, we know that $\neg(X_p \perp X_c)$ when X_i is not observed due to the active trail $X_p \rightarrow X_i \rightarrow X_c$. Therefore, $X_p \in \text{Pa}^{g'}(X_c)$. Since all such parents must appear before X_C in any topological order of G, all of them must be in $\text{Pa}^{g'}(X_c)$. Moreover, for each child of X_i, $X_{c'}$ that appears before X_c in the topological order, we also know that $\neg(X_{c'} \perp X_c)$ when X_i is not observed due to the active trail $X_{c'} \leftarrow X_i \rightarrow X_c$, therefore $X_{c'} \in \text{Pa}^{g'}(X_c), c' < c$. Are these variables sufficient? Do they create any independence relationship that is not in G? From part 4.1, we know that we need to consider the co-parents of X_i as well. Lets consider a co-parent of X_i, $X_{c'p}$, that appears in the topological order before X_c. As the name implies,$X_{c'p}$ is a parent of $X_{c'}$ which is a sibling of X_c. Ing', we have the following independency $X_{c'p} \perp X_c|\mathrm{Pa}^{g'}(X_c)$, and $X_{c'}$ is a parent of X_c in G'. However, we are now creating an independency that is NOT in G, since observing $X_{c'}$ will connect $X_{c'p}$ and X_c via an active trail in G that passes though X_i, therefore, $X_{c'p}$ can not be d-separated from X_c using the current parent set of X_c. Since each element of this parent set is necessary, we must add all such $X_{c'p}$ as parents of X_c to avoid creating these extra independencies. These variables are now in fact necessary and sufficient replacement of X_i in *parenting* X_c. In fact, we know that these variables are subset of the Markov Blanket of X_i, namely, they are the variables in the $\mathrm{MB}(X_i)$ that appear in the topological order before X_c. These variables shield X_i from other variables in the network, thus using them to replace X_i no other active trail can reach X_i and continue to X_c when these variables are observed. Therefore, using this set of new parents, the local Markov assumption is satisfied for X_c in G'.

We can explicitly create G' as follows:

$$\begin{aligned}\mathrm{Pa}^{g'}(X_j) &= \mathrm{Pa}^{g}(X_j), \qquad X_i \notin \mathrm{Pa}^{g}(X_j),\\ &= \mathrm{Pa}^{g}(X_j) \cup \mathrm{Pa}^{g}(X_i) \cup \{X_{c'}, \mathrm{Pa}^{g}(X_{c'})|c' < c, X_i \in \mathrm{Pa}^{g}(X_{c'})\} \text{ otherwise}\end{aligned}$$

Problem 3.17.

1. Consider first the triangle (A,B,C) in Figure 2, and consider one of its bases B. Now if the edges connecting B to the other two nodes in the triangle pointing to it in the same directions, like $B \leftarrow A$ and $B \leftarrow C$, then regardless of the direction of the edge $D{-}B$, both the trails $D{-}B \leftarrow A$ and $D{-}B \leftarrow C$ are either enabled or blocked at the same time, thus we can always choose to avoid going through A if they are both enabled. Now if the same holds for the other base C, we can definitely short-cut the center node A. Therefore, a necessary condition to not being able to short-cut the center of a triangle is that at least one of its bases must have the edges coming to it form the other two nodes in the triangle in different directions. Now we reason by cases over the local configuration around x_i and without loss of generality, we always direct the edge between x_{i-1} and x_{i+1} as shown (the other case follows by symmetry). (A) will result in a cycle. In (B) neither of x_{i-1} nor x_{i+1} satisfy the necessary condition. In (C) only x_{i+1} satisfies the condition and to leverage it we must connect it to x_{i+2} as shown (otherwise this node would be a don't care node as well), but now the trail $x_{i-2} - x_{i-1} \rightarrow x_{i+1} \leftarrow x_{i+2}$ is active if the trail $x_{i-2} - x_{i-1} \rightarrow x_i \leftarrow x_{i+1} \leftarrow x_{i+2}$ is active because x_i which is a decedent of the v-structure center x_{i+1} is observed, thus we can short-cut x_i. We ended with (D) and indeed node x_{i-1} satisfies our condition, and to leverage that we must connect it to x_{i-2} as shown (otherwise this node would be a don't care node as well). Now if x_{i-1} is observed, $x_{i-2} \rightarrow x_{i-1} \rightarrow x_{i+1}$ is blocked and we can not short-cut x_i in the active trail $x_{i-2} \rightarrow x_{i-1} \leftarrow x_i \rightarrow x_{i+1} - x_{i+2}$.

2. If G_1 and G_2 have the same skeleton then they must have the same set of trails. Consider an active trail in G_1: $x_1 - x2 - -x_n$, and assume for the sake of contradiction it is not active in G_2. Let x_i be the first node that blocks the trail in G_2. There are 2 cases.

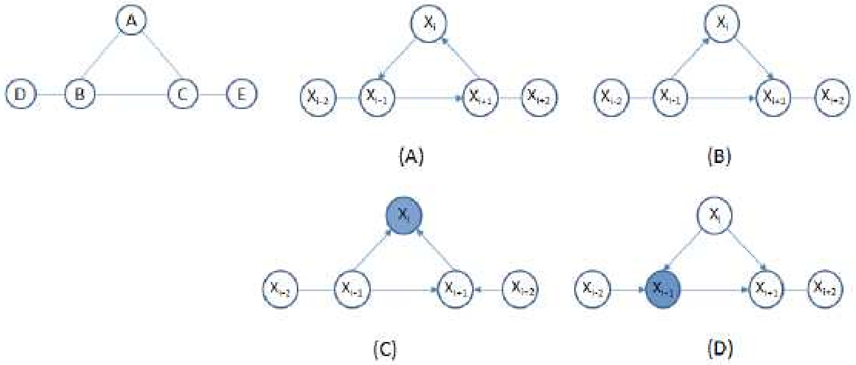

FIGURE 2. In (A-D), shaded nodes are observed, undirected edges are treated as don't care, i.e., they can be directed either way.

1. If x_i is the center of a triangle in G_1 then above we know that it participates with either of x_{i-1} or x_{i+1} in an immorality, which must be shared with G_2 as well. Thus at least $x_i \to x_{i-1}$ or $x_i \to x_{i+1}$ must exists in G_2, and thus x_i can NOT be the center of a v-structure in G2. Since xi is not observed, otherwise the trail would be blocked in G1, then xi can not block the trail in G2.

2. If xi is not the center of a triangle in G1, then we have 3 cases.

(a) xi is the center of an immorality in G1 and xi is observed, then the same must hold in G2 and xi can not block the trail.

(b) xi is NOT the center of an immorality in G1 then xi is NOT observed. Since the same must hold in G2, xi can not block the trail.

(c) xi is the center of an immorality in G1 and xi is not observed but one of its decedent E is observed. We know that xi must be the center of an immorality in G2, but we need to do more work here to show that either E is still a decedent of xi in G2 or there exists a surrogate of xi in G2. The proof here is involved but follows the same structure as the similar case in above. We will utilize figure . Let the minimal trail between xi and E be $x_i \to z1 \to z2 \to\to E$, and as we argued above, if E is not a decedent of xi in G2, then one of the edges along this trail is reversed. Let the first of these edges be $z_m \leftarrow z_{m+1}$, to avoid creating an immorality in G2, we must have $z_{m-1} - z_{m+1}$ in the skeleton of both graphs. This edge must be directed as $z_m \to z_{m+1}$ in G1 to avoid creating cycles, but this violates the fact that our path from xi to E is minimal. Therefore, the first reversed edge is $x_i \leftarrow z_1$, and again to avoid creating extra immoralities in G2, we must have $x_{i-1} - z_1$ and $x_{i+1} - z_1$ in the skeleton of both graphs. These edges must be directed as shown in figure (B) to avoid creating cycles in G1. Now we have a new immorality in G1: $x_{i-1} \to z_1 \leftarrow x_{i+1}$ which must exists as well in G2 as shown in figure (C). Now we ended up where we started with xi replaced by z1. We can recursively follow our reasoning until one of two conditions hold: either we find a node zm which is an immorality in both graphs and has E as a decedent in both graphs, or we move all the way down to E as show in the figure. In either case, there exists a substitute immorality in G2 for xi, and thus the trail can not be blocked in this case as well. Therefore, the trail must be active in G2 as well. Similarly, we can show the converse. Therefore, a trail is active in G1 iff it is active in G2. And since we already showed that G1 and G2 must have the same set of trails, the required result follows from part 1.1.

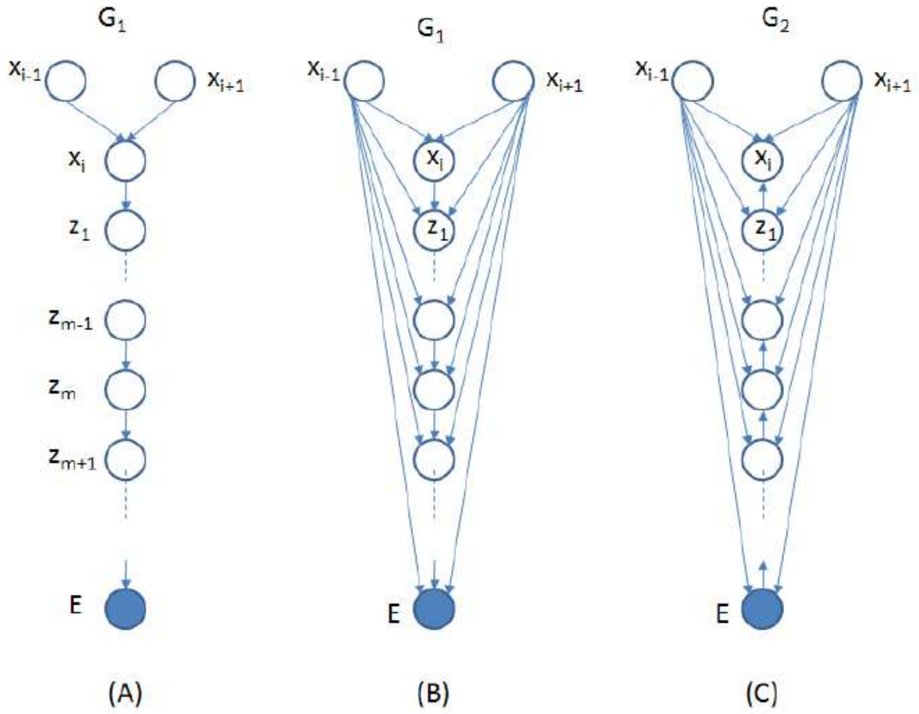

FIGURE 3. In (A-D), shaded nodes are observed, undirected edges are treated as don't care, i.e., they can be directed either way.

3. If $X - Y \in Skeleton(G1)$ then $\forall U \subseteq \mathcal{X} - X, Y; (X \perp Y|U) \notin I(G1)$, since I(G1) = I(G2), the same must hold in G2, and thus $X - Y \in Skeleton(G2)$ as well. Therefore the two graphs must have the same skeleton. Similarly, consider an immorality in G1 of the form $X \rightarrow Z \leftarrow Y$ that does not hold in G2. Since we showed that the two graphs must have the same skeleton, this can come about ONLY if Z is the parent of at least one of X or Y in G2. However now we have that $(X \perp Y|pa_{G2}(X) \cup pa_{G2}(Y)) \in I(G2)$, and this conditioning set contains Z. This independence must hold in I(G1) as well since I(G1) = I(G2), but this is a contradiction since we know that X can not be d-separated from Y given Z, that is: $\nexists U : Z \in U$ and $(X \perp Y|U) \notin I(G1)$. Therefore, the same immorality must hold in G2 as well.

HOMEWORK 3

Problem 4.4. The class of bipartite graphs $H_n = K_{n/2,n/2}$ has cliques of size 2, but any Bayes Net for H_n has $O(n)$ parents.

Also, the class of undirected $\sqrt{n}$ by $\sqrt{n}$ planar grids where each variable $X_{i,j}$ is connected to its four immediate neighbors has clique size 2 and degree 4. Pick any ordering for constructing an I-map. There is always a node whose parent-set will be order $\sqrt{n}$.

Problem 4.9. Done in class.

HOMEWORK 4

Problem 9.4. 1. Given $W = \cup_{i=1}^{k}(\{Y_i\} \cup \mathbf{Pa}_{Y_i})$, we let $Y = \cup_{i=1}^{k} Y_i$ and $Z = W - Y$. We construct the network B' over W by throwing away those edges of the original network B which end in Z. Thus, all nodes in Z have no parent. We also carry over the CPTs for each Y_i to B' (since all the parents of Y_i in B also exist in B'). The nodes in Z are set to have discrete uniform probabilities. Sort the nodes in Y topologically, so that there are no descendants of Y_i in $\{Y_1, \ldots, Y_{i-1}\}$

$$P_{B'}(Y|Z) = \frac{P_{B'}(Y,Z)}{P_{B'}(Z)} \tag{0.1}$$

$$= \frac{\prod_{i=1}^{k} P_{B'}(Y_i|Y_1, \ldots, Y_{i-1}, Z) \prod_{w \in Z} P_{B'}(w)}{\prod_{w \in Z} P_{B'}(w)} \tag{0.2}$$

$$= \prod_{i=1}^{k} P_B(Y_i|\mathbf{Pa}Y_i) = \tau(W) \tag{0.3}$$

2. If the elimination ordering is $\{X_1, X_2, \cdots\}$, then from the procedural definition of variable elimination, it follows that the intermediate factor produced at elimination step i is of the form

$$g_i = \sum_{x_{j_1}, \ldots, x_{j_i}, x_i} \prod_{i=1}^{k} P(Y_i|Pa(Y_i)) \tag{0.4}$$

Where $V = \{x_{j_1}, \ldots, x_{j_i}\} \subseteq \{X_1 \cdots X_{i-1}\}$. Variable elimination just computes this expression in an efficient way by pushing the sums to the right whenever possible and chasing computations[1], however since here we care about the semantic of this expression, we can work with the above form as it is easier to interpret. Note that all the multiplied factors in (0.4) have one of the variables in $V \cup \{X_i\}$ in either the child or the parent position. Moreover, each variable in $V \cup \{X_i\}$ must appear exactly once in one of these CPTs in a child position, even with empty parents (this follows from the semantic of Bayes network).

We can now use the result from section 2.1 that the product of the set of BN basic CPTs in (0.4) is a conditional probability in some network of the form $P(Y|Z)$. And as argued above $V \cup \{x_i\} \subseteq Y$, then marginalizing them from $P(Y|Z)$ still gives a conditional probability in some network of the form $P(Y - \{X_{j_1}, \ldots, X_{j_i}, X_i\}|Z)$

---

[1] If you are not convinced with this argument, start by writing $g_i = \sum_{x_i} \prod_{k=1} f_k$. Then for each factor f_k that is an intermediate factor produced when variable x_{j_i} was eliminated, replace it with $g_{j_i} = \sum_{x_{j_i}} \prod_{k'=1} f_{k'}$, then push the sum over x_{j_i} to the left (this is always correct). Since each intermediate factor is exactly used once, each of these operations when applied recursively will reduce the non-basic factors in the product and we will end with the expression mentioned in (0.4)

9.11. If: All Z_i are eliminated before X, Y. Every time Z_i is eliminated, either neighboring Z_{i-1} and Z_{i+1} are connected since they are in some factor with Z_i, or one of the neighbors is X or Y, and then X or Y get connected to neighboring Z. The last Z_i to get eliminated is connected to X and Y and fills in the edge.

Only if: The edge is filled in because there is variable Z before X,Y that is directly connected to X and Y when it is being eliminated, which means that at some point, a path with no variables after X,Y, namely $X - Z - Y$, was created. Since every elimination only short-cuts existing paths by eliminating the intermediate variable, every path that contains a variable after X,Y will still contain it when X or Y's turn comes up. Hence no such Z is created by elimination.

HOMEWORK 5

10.6. The clique tree nodes grow by at most 1. Simply add the new parent X to every clique in T. Clearly, the family preserving property is restored since the clique that used to hold the CPT for Y can now accommodate the new CPT. Running intersection also holds since every clique between containing X is connected. Now this is overkill: you can eliminate X from many cliques as long as still RIP holds.

10.16. There is a nice proof that illustrates an interesting property of junction trees rather than through the Max-Cardinality search. Let T be a clique tree of m maximal cliques of a chordal graph over variables $X_1, \ldots, X_n$. Consider the number of times some variable X_i appears in the nodes $C_1, \ldots, C_m$ and the edges $S_1, \ldots, S_{m-1}$ of the clique tree T. If T has running intersection property for X_i, then the cliques containing X_i form a subtree, so

$$\sum_{j=1}^{m-1} 1(X_i \in S_j) = \sum_{k=1}^{m} 1(X_i \in C_k) - 1.$$

Now if T is not a junction tree, the cliques containing X_i form a (sub) forest of T, so in general we have an inequality:

$$\sum_{j=1}^{m-1} 1(X_i \in S_j) \leq \sum_{k=1}^{m} 1(X_i \in C_k) - 1,$$

which is tight if and only if T has RIP for X_i.

Now the weight of a clique tree in our max-weight spanning tree algorithm is the sum of edge cardinalities, so we can sum the above over all nodes i to get:

$$\sum_{i=1}^{n} \sum_{j=1}^{m-1} 1(X_i \in S_j) = \sum_{j=1}^{m-1} |S_j| \leq \sum_{i=1}^{n} \left(\sum_{k=1}^{m} 1(X_i \in C_k) - 1 \right) = \sum_{k=1}^{m} |C_k| - n.$$

Notice that the right hand side is independent of T and is an upper bound on the weight. This upper bound is achieved when T has RIP for every variable, which is precisely the maximum weight tree.