Name: _______________________

CS 536: Introduction to Graphical Models  
Midterm 2020

You have 50 minutes to complete this midterm. You are only allowed to use your textbook, your notes, your assignments and solutions to those assignments during this midterm. If you find that you are spending a large amount of time on a difficult question, skip it and return to it when you’ve finished some of the easier questions. Total marks for this midterm is 47.

<table>
<thead>
<tr>
<th>Section</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Networks</td>
<td>/ 22</td>
</tr>
<tr>
<td>Markov Networks</td>
<td>/ 10</td>
</tr>
<tr>
<td>Inference</td>
<td>/ 15</td>
</tr>
<tr>
<td>Total</td>
<td>/ 47</td>
</tr>
</tbody>
</table>
Section I: Bayesian networks [22 points]

1. A Tree-Augmented Network (TAN) is like a Naïve Bayes network but it allows a feature $X_i$ to have the class $C$ as a parent and another feature $X_j$ as a parent (see example with 4 features to the right).

a) How many independent parameters does a TAN with $M$ feature nodes require? Assume $C$ takes $k$ different values, each feature $X_i$ takes $l$ different values and each feature $X_i$ has the class $C$ and one other feature $X_j$ as parents. [4 points]

The node $C$ has $(k - 1)$ independent parameters
The node $X_1$ has $k(l - 1)$ independent parameters
Each of the $(M - 1)$ other nodes $X_i$ ($i \neq 1$) has $(M - 1) \times kl(l - 1)$ independent parameters

The total is: $(k - 1) + k(l - 1) + (M - 1) kl(l - 1)$ independent parameters

b) What is the advantage of using a TAN over a Naïve Bayes network? Give a real-world example in which this advantage is useful [4 points].

You can use it to model dependence between features. For instance, in spam classification, a TAN allows you to model the fact that having the two words “billion dollars” appear together is more indicative of spam than having the words “billion” and “dollars” appear separately.

c) What is the disadvantage of using a TAN over a Naïve Bayes network? [2 points]

It requires more data to estimate the parameters. In addition, you take up more memory and use up a bit more computation time (but not too much more).
2. Define $\text{OneHopNeighbor}(X)$ as the set of nodes that are one edge away from node $X$ (regardless of the direction of the edge). Similarly, define $\text{TwoHopNeighbor}(X)$ as the set of nodes that are exactly two edges away from node $X$, regardless of the edge direction. For instance, in the graph to the right, $\text{OneHopNeighbor}(X) = \{A, B, C, D\}$ and $\text{TwoHopNeighbor}(X) = \{E, F\}$.

\[ (X \perp Y \mid \text{OneHopNeighbor}(X) \cup \text{TwoHopNeighbor}(X)) \]

where $Y \notin \text{OneHopNeighbor}(X)$ and $Y \notin \text{TwoHopNeighbor}(X)$

Here is the general idea behind the proof.

1) Consider when the two hop neighbors are above X:

```
Y → W → Z → X
```

By the causal reasoning case of d-separation, $X \perp Y \mid \{W, Z\}$

When the two hop neighbors form a common cause:

```
Y → W → Z → X
```

By the common cause case of d-separation, $X \perp Y \mid \{W, Z\}$

2) Consider when the two hop neighbors are below X:

```
X → W → Z → Y
```

By the evidential reasoning case of d-separation, $X \perp Y \mid \{W, Z\}$

Now consider the case when a v-structure is involved e.g.
The two-hop neighbors include the parents of the children of $X$. Thus by d-separation, $X \perp Y|\{W, Z\}$

b) Note that $(X \perp Y|OneHopNeighbor(X)) \Rightarrow (Y \perp X|OneHopNeighbor(Y))$ is not true (where $Y \not\in OneHopNeighbor(X)$ and $X \not\in OneHopNeighbor(Y)$). Provide an example that illustrates how this implication is violated. [4 points]

Counterexample:

\[ X \perp Y|W \not\Rightarrow Y \perp X|Z \]
Section II: Markov networks [10 points]

1. Consider a distribution $P$ over four binary random variables $A, B, C, D$ which gives probability to each of the following configurations and 0 to all others (see table to the right).

Let $H$ be the graph below. Does the distribution $P$ factorize according to $H$? Explain your answer. [10 points]

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$D$</th>
<th>$P(A,B,C,D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

To show that the distribution $P$ factorizes according to $H$, you need to show that $\{A, B\} \perp D \mid C$ i.e. $P(A, B|C) = P(A, B|C, D)$. However, from the distribution above, it is clear that this is not the case. You can pick a number of different counterexamples but here is one:

\[
P(A = 0, B = 0|C = 1) = \frac{P(A = 0, B = 0, C = 1)}{P(C = 1)} = \frac{0.2 + 0.2}{0.2 + 0.2 + 0.4} = \frac{0.4}{0.8} = 0.5
\]

\[
P(A = 0, B = 0|C = 1, D = 1) = \frac{P(A = 0, B = 0, C = 1, D = 1)}{P(C = 1, D = 1)} = \frac{0.2}{0.2 + 0.4} = \frac{0.2}{0.6} = 0.333
\]

Clearly $P(A = 0, B = 0|C = 1) \neq P(A = 0, B = 0|C = 1, D = 1)$
Section III: Inference [12 points]

1. The following questions deal with the graph below.

![Graph Diagram]

a) Show all the intermediate factors produced by computing $P(A, D, G)$ using variable elimination using the ordering $F, C, E, B$ (where you eliminate $F$ first, then $C$, then $E$ and then $B$—meaning the rightmost summation is over $F$).[10 points]

\[
P(A, D, G) = \sum_B \sum_E \sum_C \sum_F P(A)P(B)P(D)P(E)P(C|A, B)P(F|C, D, E)P(G|C, F)
\]

\[
= P(A)P(D) \sum_B P(B) \sum_E P(E) \sum_C P(C|A, B) \sum_F P(F|C, D, E)P(G|C, F)
\]

1) Eliminating F

\[
P(A, D, G) = \frac{1}{Z} \phi_A(A)\phi_D(D) \sum_B \phi_B(B) \sum_E \phi_E(E) \sum_C \phi_C(A, B, C) \sum_F \phi_F(C, D, E, F)\phi_G(C, F, G)
\]

\[
\psi_1(C, D, E, F, G) = \phi_F(C, D, E, F)\phi_G(C, F, G)
\]

\[
\tau_1(C, D, E, G) = \sum_F \psi_1(C, D, E, F, G)
\]

2) Eliminating C

\[
P(A, D, G) = \frac{1}{Z} \phi_A(A)\phi_D(D) \sum_B \phi_B(B) \sum_E \phi_E(E) \sum_C \phi_C(A, B, C) \tau_1(C, D, E, G)
\]

\[
\psi_2(A, B, C, D, E, G) = \phi_C(A, B, C)\psi_1(C, D, E, F, G)
\]

\[
\tau_2(A, B, D, E, G) = \sum_C \psi_2(A, B, C, D, E, G)
\]

3) Eliminating E

\[
P(A, D, G) = \frac{1}{Z} \phi_A(A)\phi_D(D) \sum_B \phi_B(B) \sum_E \phi_E(E) \tau_2(A, B, D, E, G)
\]

\[
\psi_3(A, B, D, E, G) = \phi_E(E)\tau_2(A, B, D, E, G)
\]

\[
\tau_3(A, B, D, G) = \sum_E \psi_3(A, B, D, E, G)
\]

4) Eliminating B

\[
P(A, D, G) = \frac{1}{Z} \phi_A(A)\phi_D(D) \sum_B \phi_B(B)\tau_3(A, B, D, G)
\]
\[ \psi_4(A, B, D, G) = \phi_B(B)\tau_3(A, B, D, G) \]
\[ \tau_4(A, D, G) = \sum_B \psi_4(A, B, D, G) \]

At this point:
\[ P(A, D, G) = \frac{1}{Z} \phi_A(A)\phi_E(E)\tau_4(A, D, G) \]

b) What is the induced width from (a)? [2 points]

The induced width is the size of the largest clique -1. In this example, the largest clique is \( \psi_2(A, B, C, D, E, G) \), which is of size 6.
The induced width is 6-1 = 5.

c) Draw the induced graph after the first step of variable elimination i.e. after you eliminate F. [3 points]