# Lecture 18:

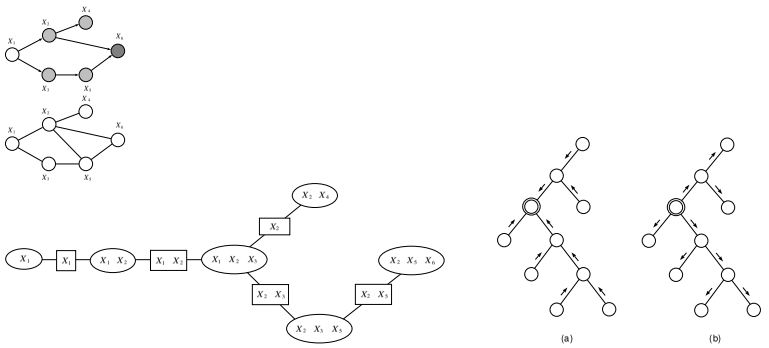
# Junction Trees

Sam Roweis

March 18, 2002

## The Main Setup

- Three main steps:
  1. Pre-processing (compiling) the graphical model to prepare for inference: building the clique junction tree.
  2. Conditioning on the evidence.
  3. Marginalizing out the non-query nodes efficiently.



## Recap: General Inference Problem

- We want to be able to condition on some "evidence" $\mathbf{x}_E$ (observed nodes) and compute the posterior probabilities of some "query" nodes $\mathbf{x}_F$ while marginalizing out "nuisance" nodes $\mathbf{x}_H$.
- We will focus on sets of nodes corresponding to *maximal cliques* of the original undirected or moralized graphical model.
- Q: Why maximal cliques?
  A: Because they are the sufficient statistics for learning with EM if the joint distribution factors according to our graphical model semantics.

## Pre-Processing

The pre-processing step has several parts:

1. If the original graph is directed, *moralize* it.
2. Add edges until the graph is *triangulated*. (HARD!)
3. Identify the maximal cliques. (easy if triangulated).
4. Arrange the cliques into a *junction tree*. (MST)
5. Associate a potential function with each clique in this hypergraph, and define the joint probability as the product of these potentials:

$$p(\mathbf{X}) = \frac{1}{Z} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c) \qquad Z = \sum_{\mathbf{X}} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c)$$

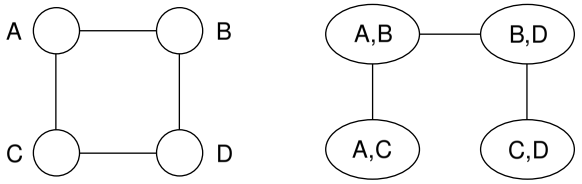
   very similar to setup for undirected models, except that the cliques here are maximal, whereas in undirected graphs they might not be.
6. Initialize the clique potentials based on the parameters of the underlying graphical model.

## Manipulations on Clique Potentials

- Initialization: the clique potentials must be initialized so that they represent the correct joint distribution.
  We use the potentials or conditionals from the underlying graph, assigning ambiguous potentials to one and only one clique.
- Conditioning on evidence:
- Marginalizing:

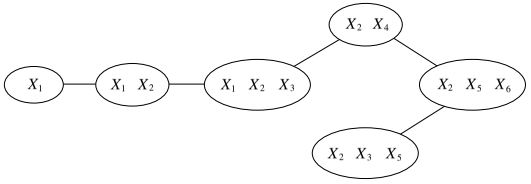
## Not All Graphs Have Junction Trees

- Consider the graph below, with 4 maximal cliques.
- No clique tree has the junction tree-property.
- Hint: elimination algorithm, with any ordering, would have created at least one extra link.
- Who cares? We do! Our inference algorithms will be based on maintaining local pariwise consistency between cliques in the clique tree. If the same variables appear in disconnected cliques their representations will not be guaranteed to be correct.




## Reminder: Junction Tree Property

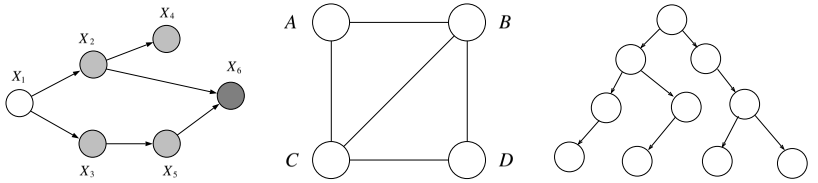
- We will arrange the maximal cliques into a clique tree
  Q: Which structure should we pick for efficient calculations?
  A: One in which local information transfer is sufficient to update all necessary quantities consistently.
- KEY IDEA: not all clique trees are equal! If, for every variable, all occurances of the variable appear in a connected subtree, then the clique tree has the *junction tree* property, and this is very desirable.
- Formally: for every pair of cliques $V, W$ all cliques on the unique path b/w $V$ and $W$ contain $V \cap W$.




e.g. this clique tree does not have the j-tree property ($x_3$)

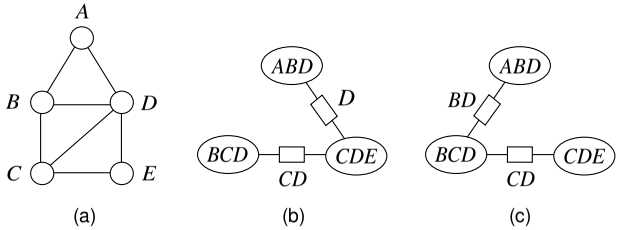
## Which Graphs Have Junction Trees?

- It turns out that having a junction tree is equivalent to another, well know graph property: *triangulation*.
- Triangulated $\equiv$ no undirected cycles of length $> 3$ containing only edges between successor nodes.
- Examples:



## Constructing the Junction Tree

- Even for triangulated graphs, not every clique tree has the junction tree property.




- A junction tree is a maximum weight spanning tree where the weights are the number of variables shared by two cliques.
- We can construct a junction tree using Kruskal's or Prim's algorithms: at each step add an edge that has the most number of variables in common, unless this would create a cycle.

## Example: two node binary model

$$p(A,B) = p(A)p(B|A)$$
$$p(A=1) = 0.8$$
$$p(B=1|A=1) = 0.7$$
$$p(B=1|A=0) = 0.4$$
$$\psi(A,B) = \begin{bmatrix} .12 & .08 \\ .24 & .56 \end{bmatrix}$$

Now, if we observe $B = 1$:

$$p(B=1) = 1$$
$$\bar{\psi}(A) = [.08; .56]$$
$$\bar{Z} = .64 = p(B=1)$$
$$\frac{1}{\bar{Z}}\bar{\psi}(A) = [.125; .875] = p(A|B=1)$$

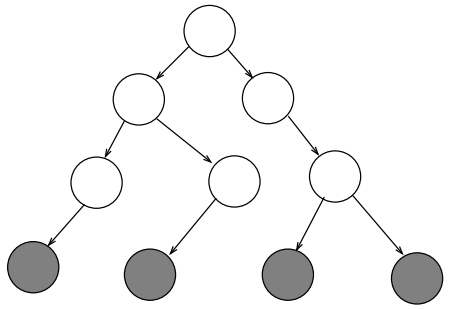
## Conditioning (introducing evidence)

- Conditioning is like "generalized indexing" or "slicing".
- Assume $\mathbf{x}_F$ empty for now: evidence $\mathbf{x}_E$; non-evidnce $\mathbf{x}_H$.
- For every clique potential, consider the subset of its variables on which we are conditioning and take a "slice" of the potential function by fixing those variables to their observed values.
- Thus, each clique becomes a function of the variables in the intersection of its original variables and the non-evidence variables.
- Now take the product of these "reduced" clique potentials and we get an unnormalized quantity proportional to $p(\mathbf{x}_H, \bar{\mathbf{x}}_E)$:

$$p(\mathbf{x}_H, \bar{\mathbf{x}}_E) = \frac{1}{Z} \prod_{\text{cliques}} \psi_C(\mathbf{x}_{C \cap H} | \bar{\mathbf{x}}_{C \cap E})$$

- The normalizer $\bar{Z}$ for just these sliced potentials is now the evidence probability $p(\bar{\mathbf{x}}_E)$. The conditional we want is $p(\mathbf{x}_H|\bar{\mathbf{x}}_E) = p(\mathbf{x}_H, \bar{\mathbf{x}}_E)/p(\bar{\mathbf{x}}_E) = p(\mathbf{x}_H, \bar{\mathbf{x}}_E)/\bar{Z}$

## Example: Trees with Hidden Nodes



- No moralization or triagulation necessary.
- Cliques are pairs of nodes.
- Potentials $\psi_C$ initialized to parent-conditionals.
- Evidence will flow from leaves to root and then marginals will be corrected on the way back down.