Machine learning: lecture 23

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Announcements

- Course evaluations ... on-going
- Project submission (Friday Dec 3):
  - only electronic submissions (pdf or ps); see the course website
  - if you need an extension (and have a reason), you need to ask. Late submissions are not possible otherwise.
- Final exam, in class (Wed Dec 8):
  - a part of the lecture on Monday Dec 6 will be review
  - comprehensive (covers all the course material) but the emphasis will be on the material since the midterm
  - as promised, EM and HMMs will be on the exam
  - open book, laptops fine if not connected

Exact inference

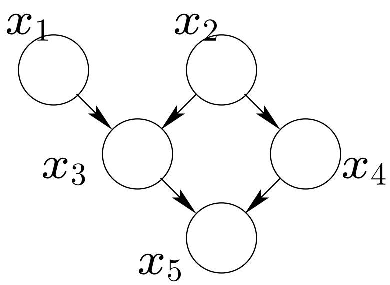
- All exact inference algorithms for Bayesian networks perform essentially the same calculations but operate on different representations
- The junction tree algorithm is a simple message passing algorithm over *clusters of variables*

  Preliminary steps:

1. transform the Bayesian network into an undirected model via moralization ("marry parents")
2. triangulate the resulting undirected graph (add edges)
3. identify the cliques (clusters) of the resulting triangulated graph
4. construct the junction tree from the cliques

Exact inference: preliminary steps

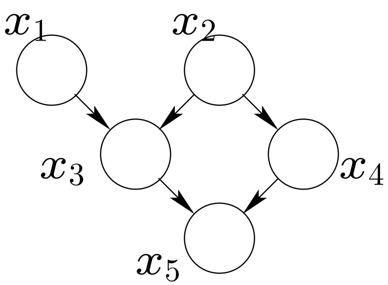
- Moralization



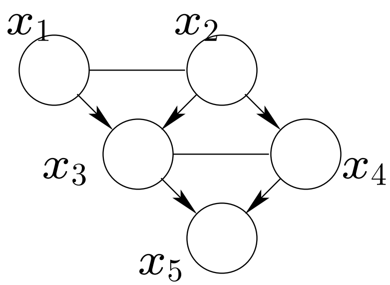
original graph

Exact inference: preliminary steps

- Moralization



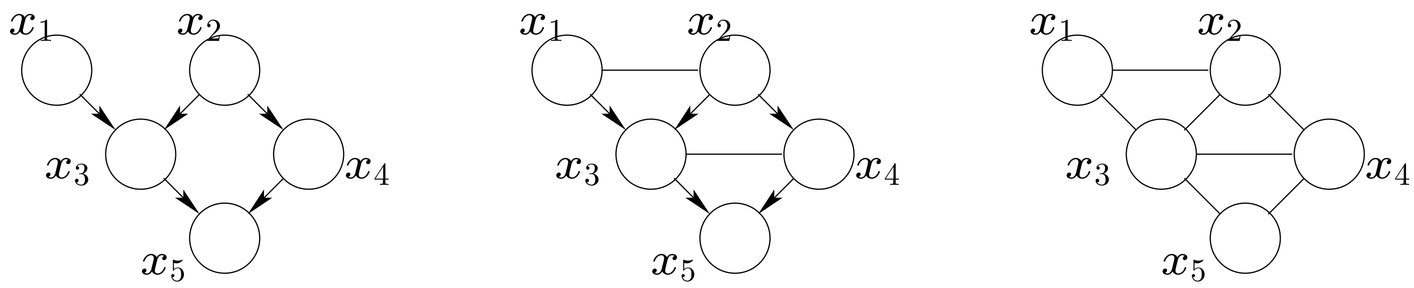
original graph



"marry" parents

Exact inference: preliminary steps

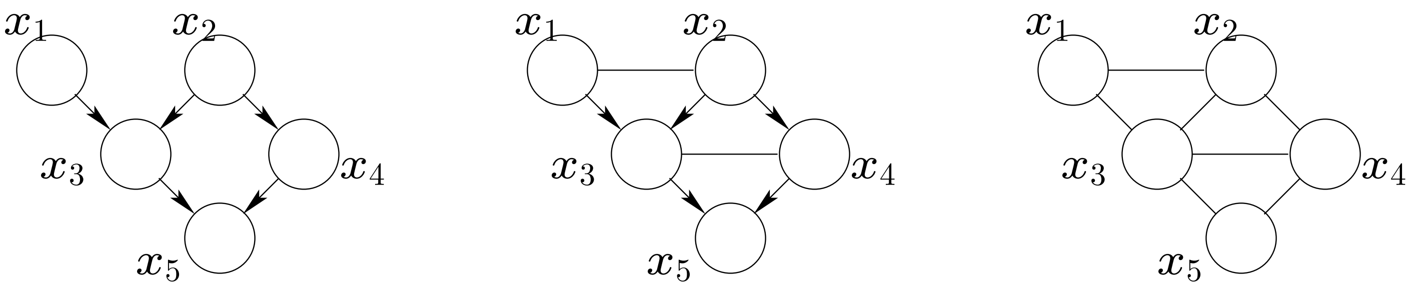
- Moralization



original graph "marry" parents moral graph

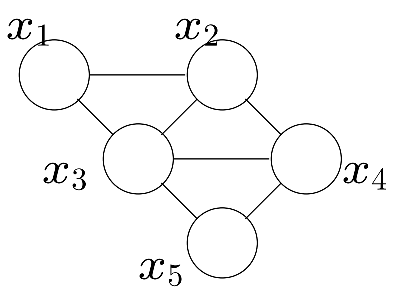
Exact inference: preliminary steps

- Moralization



original graph   “marry” parents   moral graph

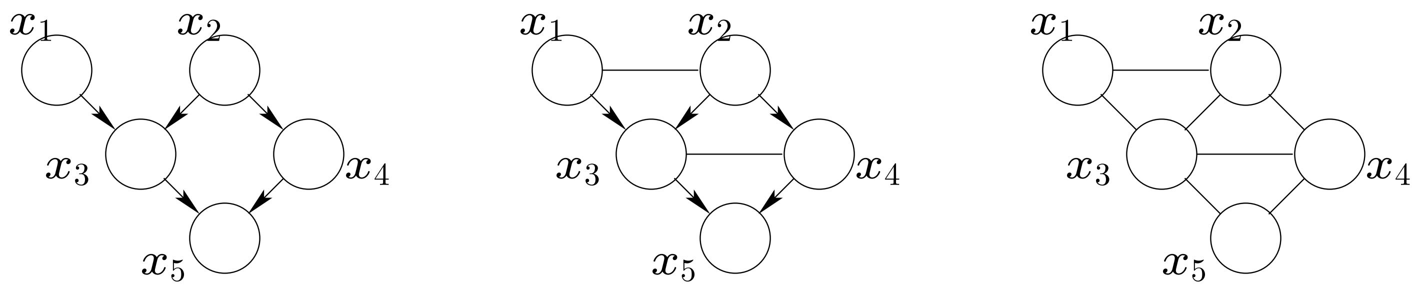
- Triangulation (add edges so that any cycle of four or more nodes has a “chord”)



already triangulated

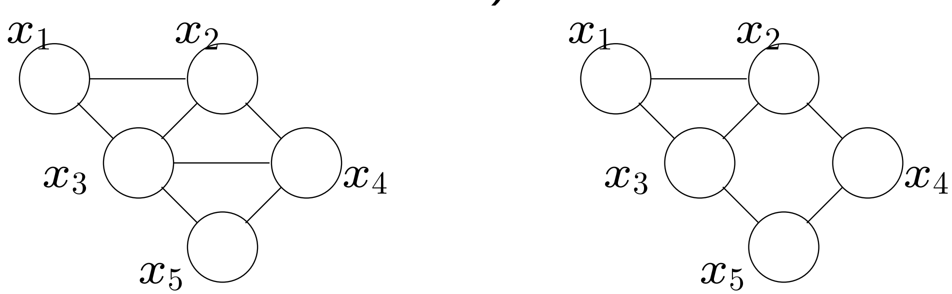
Exact inference: preliminary steps

- Moralization



original graph "marry" parents moral graph

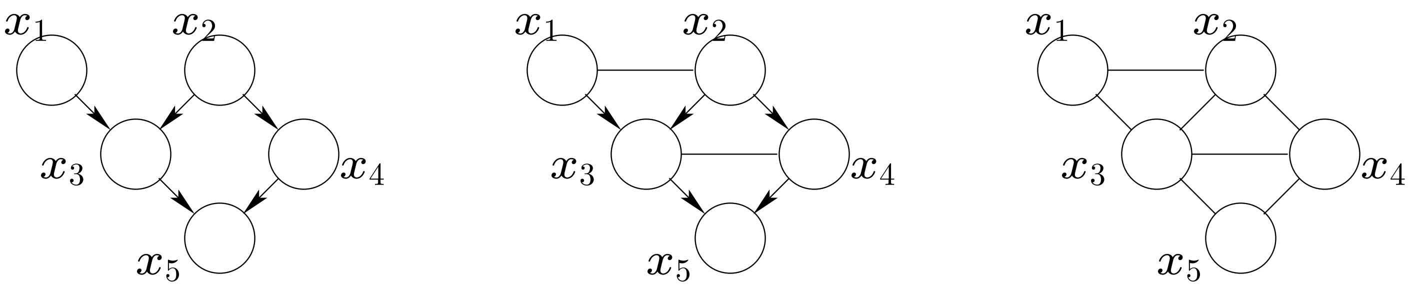
- Triangulation (add edges so that any cycle of four or more nodes has a "chord")



already triangulated not triangulated

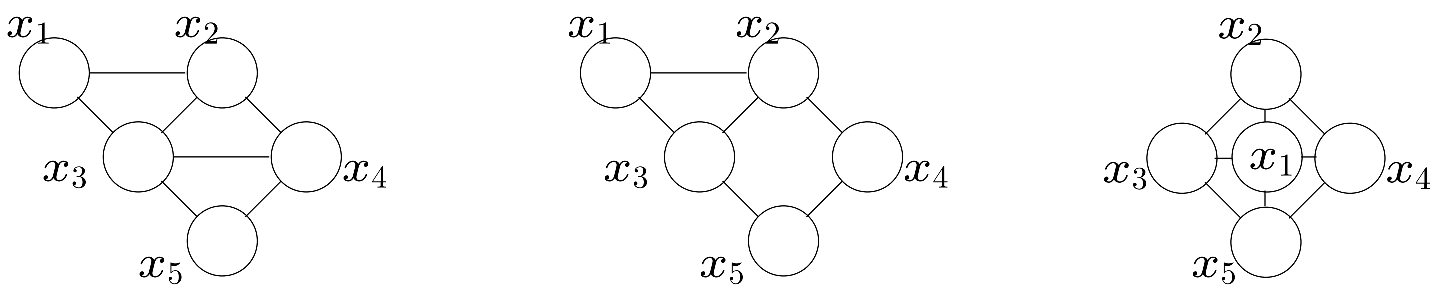
Exact inference: preliminary steps

- Moralization



original graph   “marry” parents   moral graph

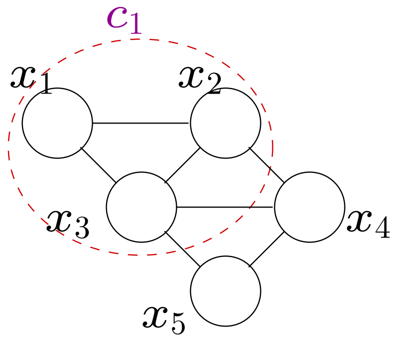
- Triangulation (add edges so that any cycle of four or more nodes has a “chord”)



already triangulated   not triangulated   not triangulated

Exact inference: preliminary steps cont'd

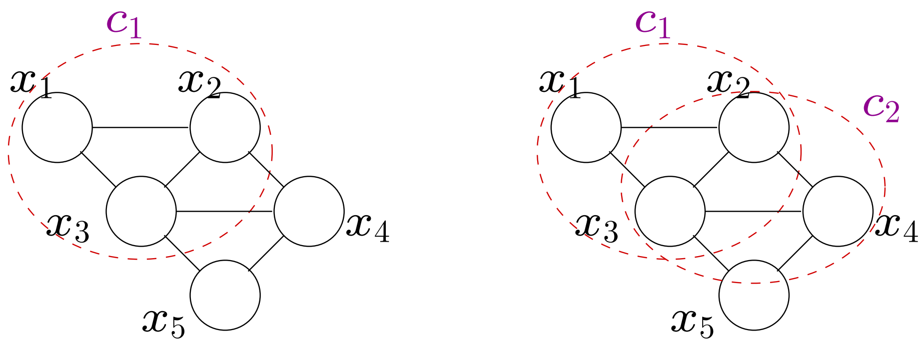
- Find the maximal cliques of the triangulated graph



$$c_1 = \{x_1, x_2, x_3\}$$

Exact inference: preliminary steps cont'd

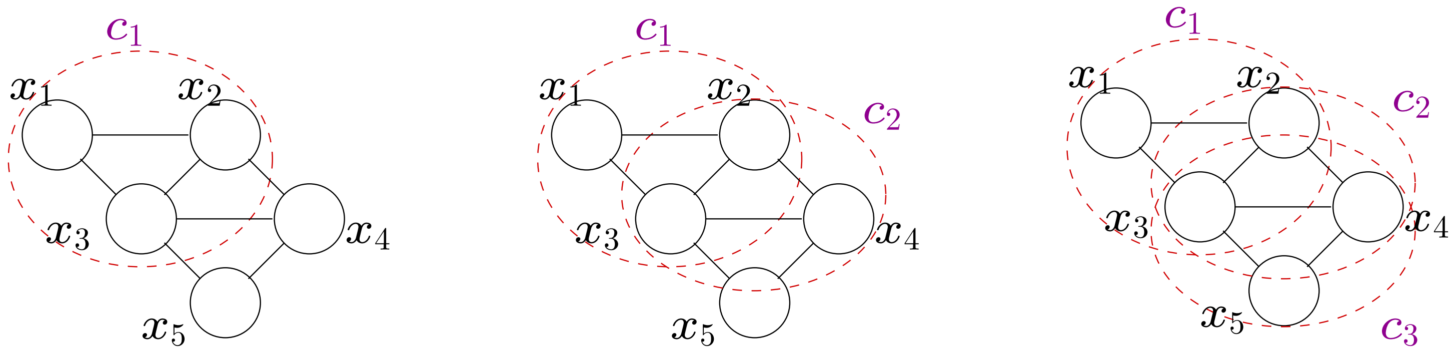
- Find the maximal cliques of the triangulated graph



$$c_1 = \{x_1, x_2, x_3\} \quad c_2 = \{x_2, x_3, x_4\}$$

Exact inference: preliminary steps cont'd

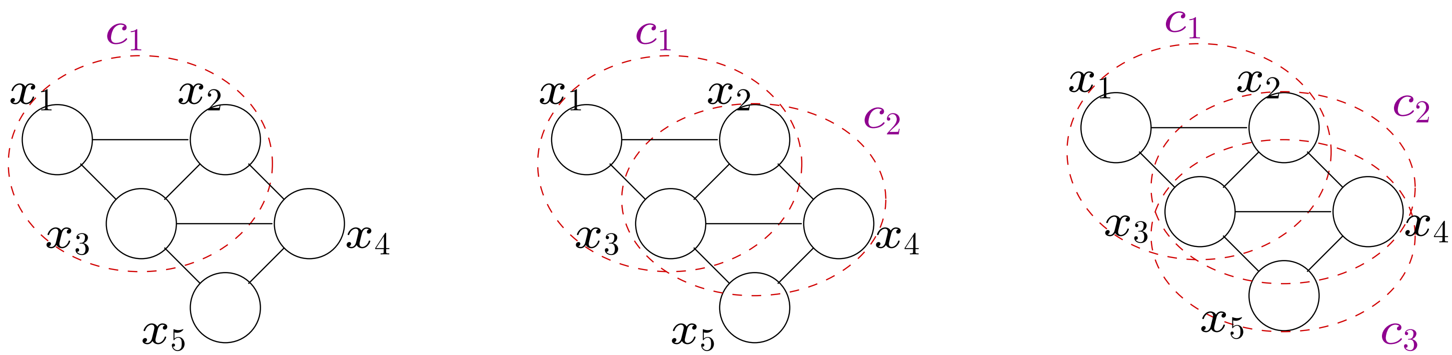
- Find the maximal cliques of the triangulated graph



$$c_1 = \{x_1, x_2, x_3\} \quad c_2 = \{x_2, x_3, x_4\} \quad c_3 = \{x_3, x_4, x_5\}$$

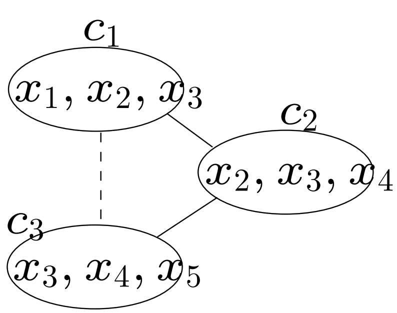
Exact inference: preliminary steps cont'd

- Find the maximal cliques of the triangulated graph



$$c_1 = \{x_1, x_2, x_3\} \quad c_2 = \{x_2, x_3, x_4\} \quad c_3 = \{x_3, x_4, x_5\}$$

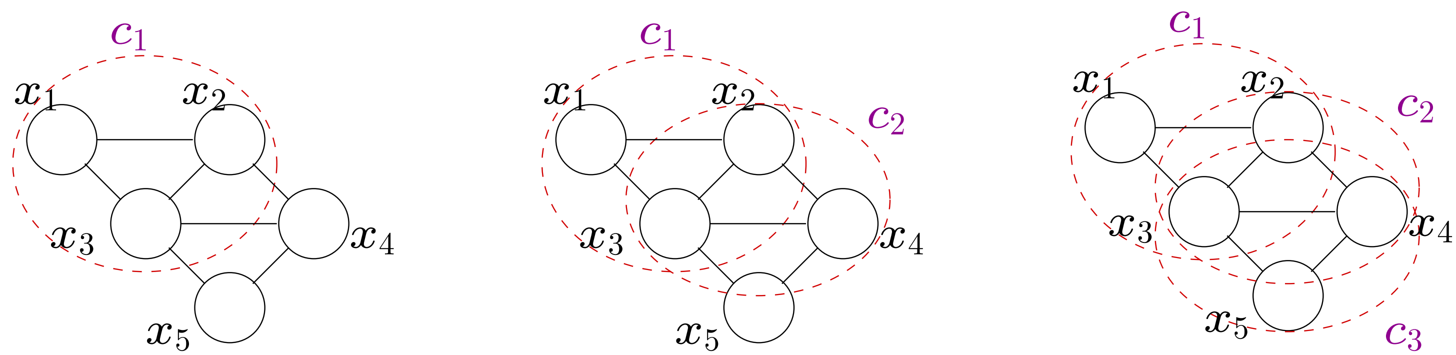
- Clique trees and junction trees



clique tree

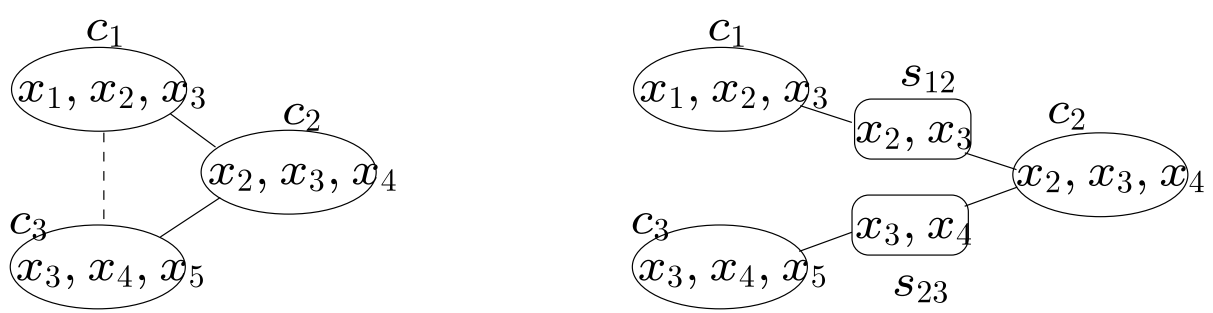
Exact inference: preliminary steps cont'd

- Find the maximal cliques of the triangulated graph



$$c_1 = \{x_1, x_2, x_3\} \quad c_2 = \{x_2, x_3, x_4\} \quad c_3 = \{x_3, x_4, x_5\}$$

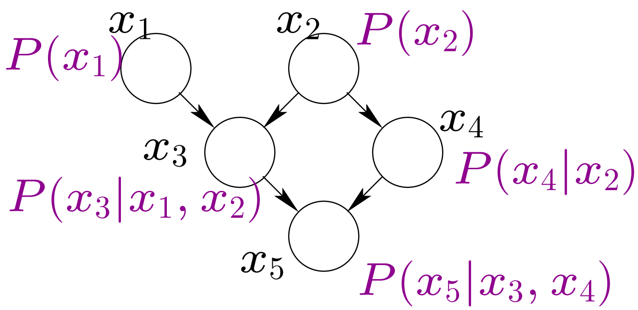
- Clique trees and junction trees



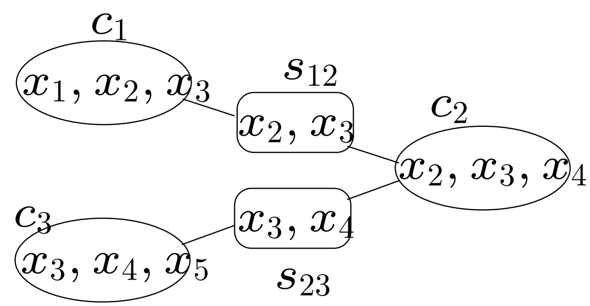
clique tree   junction tree (with separators)

Exact inference: potentials

- Associating graphs and potentials



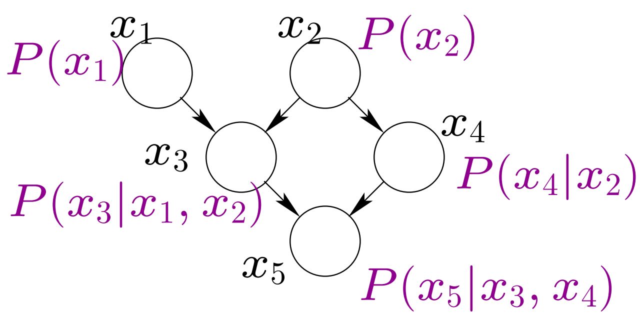
original graph w/ probs



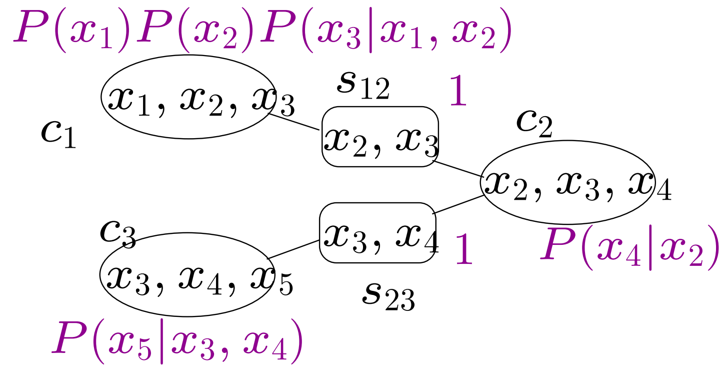
junction tree

Exact inference: potentials

- Associating graphs and potentials



original graph w/ probabilities



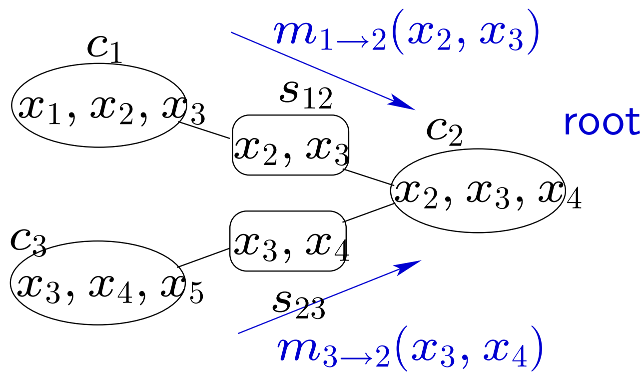
junction tree w/ probs

$$
\begin{aligned}
\psi_{c_1}(x_1,x_2,x_3) &= P(x_1)P(x_2)P(x_3|x_1,x_2) \\
\psi_{c_2}(x_2,x_3,x_4) &= P(x_4|x_2) \\
\psi_{c_3}(x_3,x_4,x_5) &= P(x_5|x_3,x_4) \\
\psi_{s_{12}}(x_2,x_3) &= 1 \quad \text{(separator)} \\
\psi_{s_{23}}(x_3,x_4) &= 1 \quad \text{(separator)}
\end{aligned}
$$

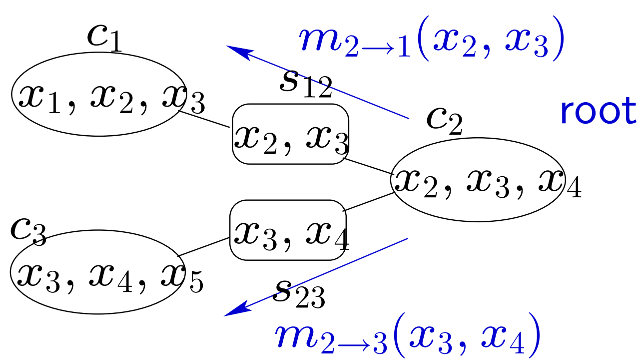
Exact inference: message passing

- Select a root clique

- Collect evidence

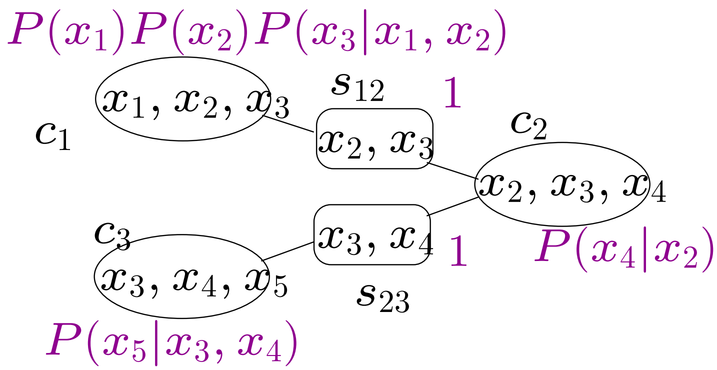


- Distribute evidence



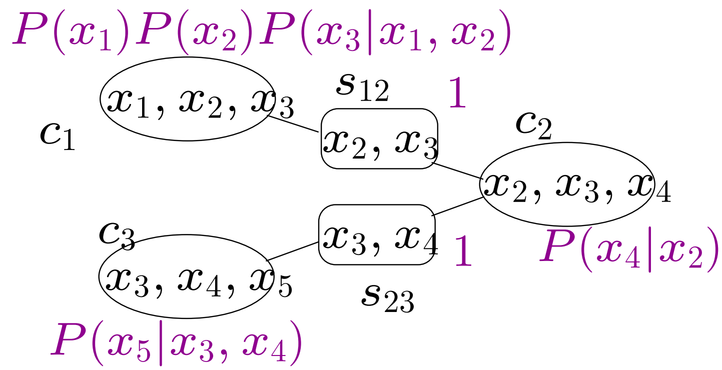
Exact inference: message passing

- Collect evidence



Exact inference: message passing

- Collect evidence



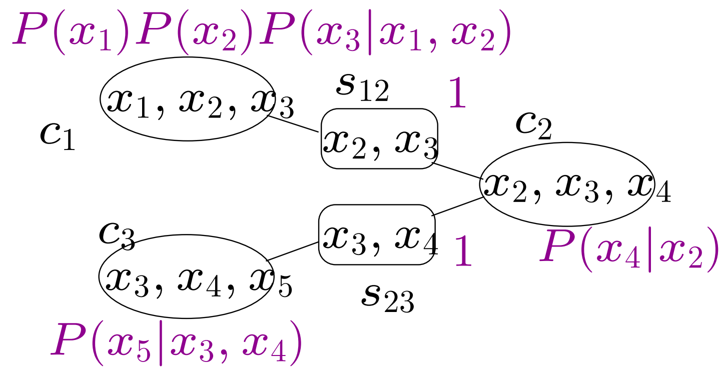
Evaluate new separators:

$$\psi'_{s_{12}}(x_2, x_3) = \sum_{x_1} \psi_{c_1}(x_1, x_2, x_3) \quad = P(x_2, x_3)$$

$$\psi'_{s_{23}}(x_3, x_4) = \sum_{x_5} \psi_{c_3}(x_3, x_4, x_5) \quad = 1$$

Exact inference: message passing

- Collect evidence



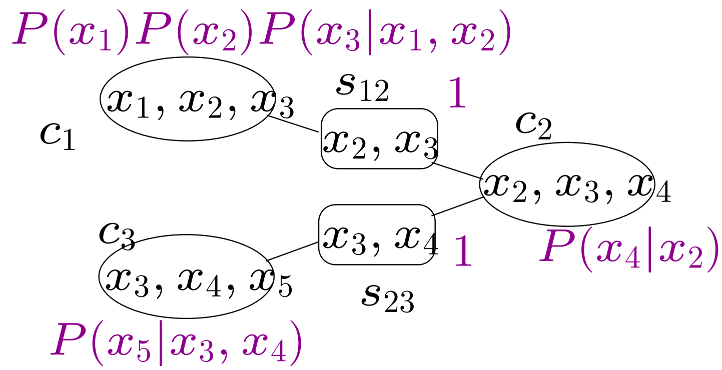
Messages (not explicitly used in the algorithm):

$$m_{1\to 2}(x_2,x_3) = \frac{\psi'_{s_{12}}(x_2,x_3)}{\psi_{s_{12}}(x_2,x_3)} = \frac{P(x_2,x_3)}{1}$$

$$m_{3\to 2}(x_3,x_4) = \frac{\psi'_{s_{23}}(x_3,x_4)}{\psi_{s_{23}}(x_3,x_4)} = \frac{1}{1}$$

Exact inference: message passing

- Collect evidence



Update clique potentials (based on messages):

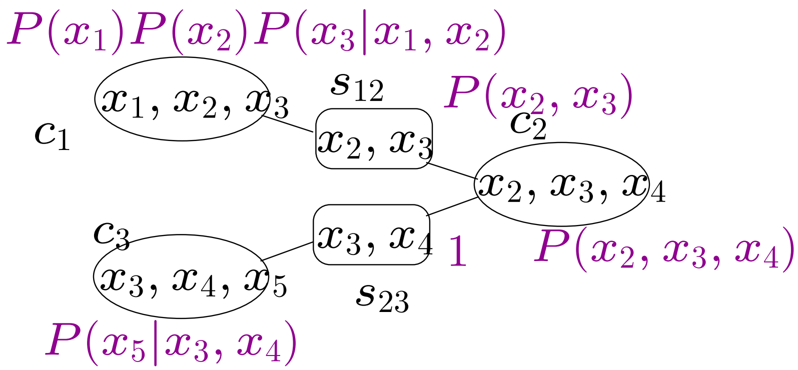
$$\psi_{c_2}(x_2,x_3,x_4) \;\leftarrow\; \underbrace{\frac{\psi'_{s_{12}}(x_2,x_3)}{\psi_{s_{12}}(x_2,x_3)}}_{m_{1\to 2}(x_2,x_3)} \cdot \underbrace{\frac{\psi'_{s_{23}}(x_3,x_4)}{\psi_{s_{23}}(x_3,x_4)}}_{m_{3\to 2}(x_3,x_4)} \cdot \psi_{c_2}(x_2,x_3,x_4)$$

$$= P(x_2,x_3) \cdot 1 \cdot P(x_4|x_2) = P(x_2,x_3,x_4)$$

followed by $\psi_{s_{12}} \leftarrow \psi'_{s_{12}}$ and $\psi_{s_{23}} \leftarrow \psi'_{s_{23}}$

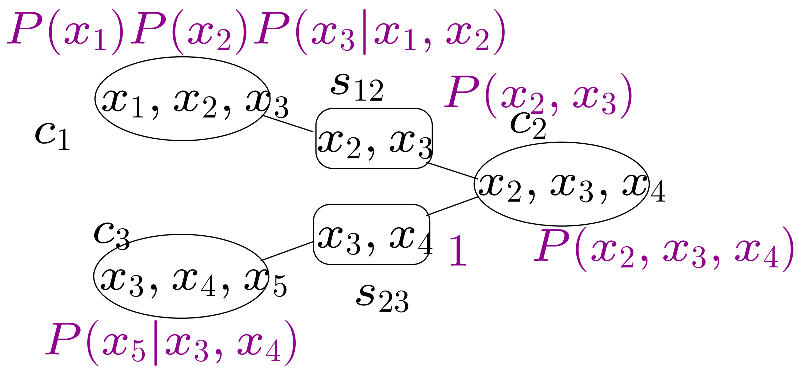
Exact inference: message passing

- Distribute evidence



Exact inference: message passing

- Distribute evidence



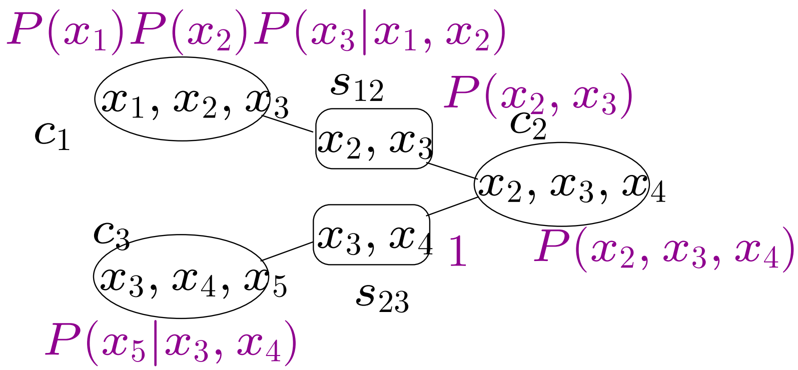
Evaluate new separators:

$$\psi'_{s_{12}}(x_2,x_3) = \sum_{x_4} \psi_{c_2}(x_2,x_3,x_4) = P(x_2,x_3)$$

$$\psi'_{s_{23}}(x_3,x_4) = \sum_{x_2} \psi_{c_2}(x_2,x_3,x_4) = P(x_3,x_4)$$

Exact inference: message passing

- Distribute evidence



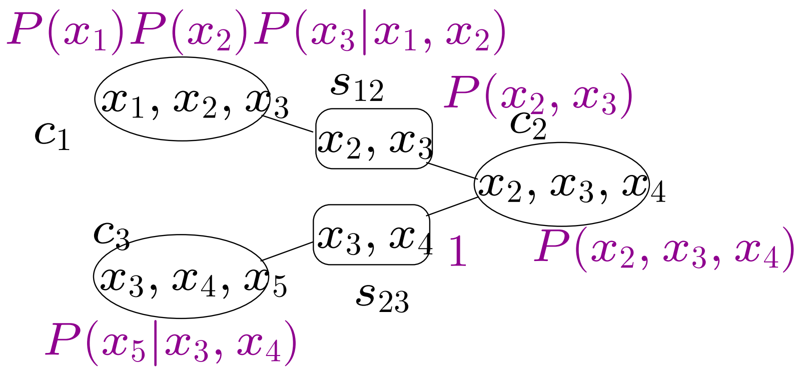
Messages (not explicitly used in the algorithm):

$$m_{2\to1}(x_2,x_3) = \frac{\psi'_{s_{12}}(x_2,x_3)}{\psi_{s_{12}}(x_2,x_3)} = \frac{P(x_2,x_3)}{P(x_2,x_3)} = 1$$

$$m_{2\to3}(x_3,x_4) = \frac{\psi'_{s_{23}}(x_3,x_4)}{\psi_{s_{23}}(x_3,x_4)} = \frac{P(x_3,x_4)}{1}$$

Exact inference: message passing

- Distribute evidence



Update clique potentials (based on messages):

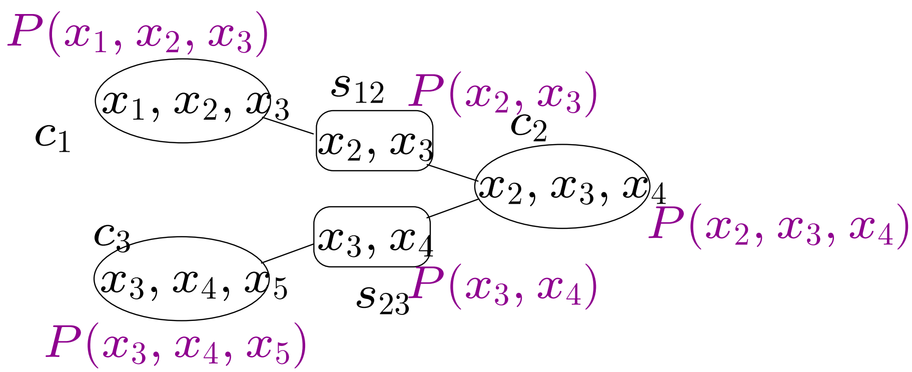
$$\psi_{c_1}(x_1, x_2, x_3) \;\leftarrow\; \frac{\psi'_{s_{12}}(x_2, x_3)}{\psi_{s_{12}}(x_2, x_3)}\psi_{c_1}(x_1, x_2, x_3) \;\; = P(x_1, x_2, x_3)$$

$$\psi_{c_3}(x_3, x_4, x_5) \;\leftarrow\; \frac{\psi'_{s_{23}}(x_3, x_4)}{\psi_{s_{23}}(x_3, x_4)} \cdot \psi_{c_3}(x_3, x_4, x_5) \;\; = P(x_3, x_4, x_5)$$

followed by $\psi_{s_{12}} \leftarrow \psi'_{s_{12}}$ and $\psi_{s_{23}} \leftarrow \psi'_{s_{23}}$

Exact inference

- After the collect and distribute steps the marginal probabilities are stored *locally* at the clique potentials (and the separators)



- The algorithm maintains the joint distribution as a product of clique potentials over separators

$$P(x_1, \ldots, x_5) = \frac{\prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)}{\prod_{s \in \mathcal{S}} \psi_s(\mathbf{x}_s)}$$

(cf. H-C theorem)

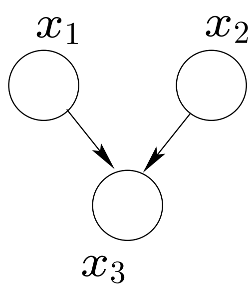
Outline

- Learning Bayesian networks: complete data
  - estimating the parameters with fixed structure
  - learning the graph structure
- Learning Bayesian networks: incomplete data
  - EM and structural EM

Probabilities and conditional tables

- For simplicity we will consider only Bayesian network models with discrete variables
- A fully parameterized model is one where there are no restrictions on the probability tables describing the conditional (marginal) probabilities

$$P(x_1): \begin{array}{c} 0.5 \\ 0.5 \end{array} \qquad P(x_2): \begin{array}{c} 0.5 \\ 0.5 \end{array}$$



$$P(x_3|x_1, x_2): \begin{array}{c|cccc} & hh & ht & th & tt \\ h & 0.9 & 0.2 & 0.3 & 0.6 \\ t & 0.1 & 0.8 & 0.7 & 0.4 \end{array}$$

(there are $1 + 1 + 4 = 6$ adjustable parameters in this model)

Likelihood and complete data

- When the observed data points are complete, the likelihood can be decomposed into a product of terms involving only each conditional table:

$$
\begin{aligned}
P(D|G,\theta) &= \prod_{t=1}^{n} P(x_1^t)P(x_2^t)P(x_3^t|x_1^t,x_2^t) \\
&= \prod_{x_1} P(x_1)^{N(x_1)} \times \prod_{x_2} P(x_2)^{N(x_2)} \\
&\quad \times \prod_{x_1,x_2,x_3} P(x_3|x_1,x_2)^{N(x_1,x_2,x_3)}
\end{aligned}
$$

Likelihood and complete data

- When the observed data points are complete, the likelihood can be decomposed into a product of terms involving only each conditional table:

$$
\begin{aligned}
P(D|G,\theta) &= \prod_{t=1}^{n} P(x_1^t)P(x_2^t)P(x_3^t|x_1^t,x_2^t) \\
&= \prod_{x_1} P(x_1)^{N(x_1)} \times \prod_{x_2} P(x_2)^{N(x_2)} \\
&\quad \times \prod_{x_1,x_2} \prod_{x_3} P(x_3|x_1,x_2)^{N(x_1,x_2,x_3)}
\end{aligned}
$$

Each conditional table such as $P(x_3|x_1,x_2)$ for a fixed x_1 and x_2, can be estimated separately based on the observed counts such as $N(x_1,x_2,x_3)$.

ML parameter estimates

- Let $\theta_{\cdot|x_1,x_2} = \{\theta_{1|x_1,x_2}, \ldots, \theta_{m|x_1,x_2}\}$ be the parameters defining the conditional table so that

$$P(x_3|x_1,x_2) = \theta_{x_3|x_1,x_2}$$

where $\sum_{x_3=1}^{m} \theta_{x_3|x_1,x_2} = 1$ for all values of x_1 and x_2.

- The ML estimates of these parameters are simply normalized counts (cf. Markov models):

$$\hat{\theta}_{x_3|x_1,x_2} = \frac{N(x_1,x_2,x_3)}{N(x_1,x_2)}$$

where $N(x_1,x_2) = \sum_{x_3} N(x_1,x_2,x_3)$.