Probabilistic Graphical Models

Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 4:
Inference & Elimination Algorithms

Some figures courtesy Michael Jordan's draft textbook,
An Introduction to Probabilistic Graphical Models

Minimizing Expected Loss

$y \in \mathcal{Y} \longrightarrow$ unknown class or category, finite set of options

$x \in \mathcal{X} \longrightarrow$ observed data, can take values in any space

$\mathcal{A} = \mathcal{Y} \longrightarrow$ action is to choose one of the categories

$L(y, a) \longrightarrow$ table giving loss for all possible mistakes

- The *posterior expected loss* of taking action *a* is

$$\rho(a|\mathbf{x}) \triangleq \mathbb{E}_{p(y|\mathbf{x})}\left[L(y,a)\right] = \sum_{y} L(y,a)p(y|\mathbf{x})$$

- The optimal *Bayes decision rule* is then

$$\delta(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \rho(\mathbf{a}|\mathbf{x})$$

- Bayesian classification requires *both* model and loss

Minimizing Probability of Error

$$L(y,a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases}$$

- The *posterior expected loss* of taking action *a* is

$$\rho(a|\mathbf{x}) \triangleq \mathbb{E}_{p(y|\mathbf{x})}\left[L(y,a)\right] = \sum_y L(y,a)p(y|\mathbf{x})$$

$$\rho(a \mid x) = p(a \neq y \mid x) = 1 - p(a = y \mid x)$$

- Optimal decision is the *maximum a posteriori (MAP)* estimate:

$$\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} p(y \mid x)$$

- If classes are equally likely *a priori*, this becomes

$$\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} p(x \mid y) \quad \text{if} \quad p(y) = \frac{1}{C}$$

Inference in Graphical Models

$x_E \longrightarrow$ observed *evidence* variables (subset of nodes)

$x_F \longrightarrow$ unobserved *query* nodes we'd like to infer

$x_R \longrightarrow$ remaining variables, *extraneous* to this query but part of the given graphical representation

$$p(x_E, x_F) = \sum_{x_R} p(x_E, x_F, x_R) \qquad R = V \setminus \{E, F\}$$

Maximum a Posteriori (MAP) Estimates

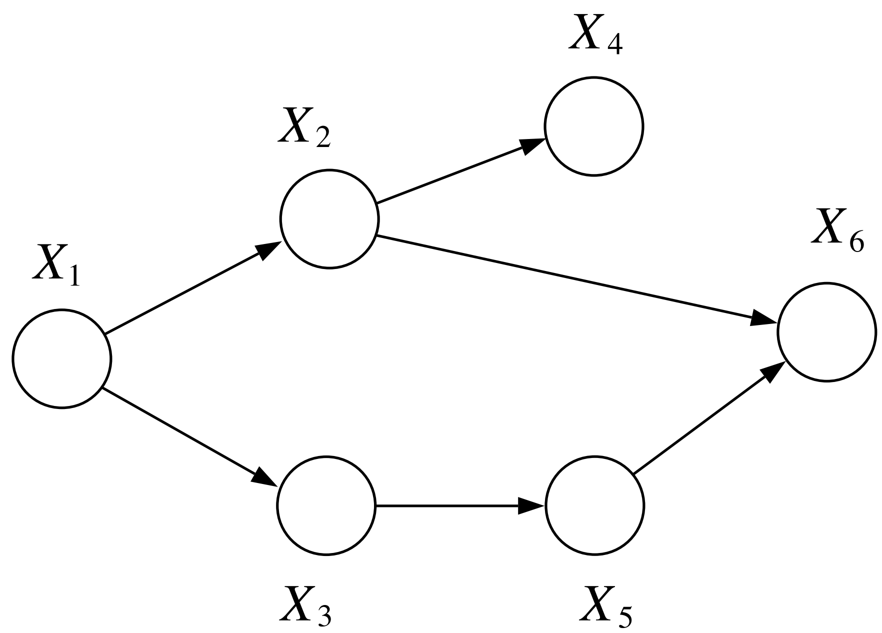
$$\hat{x}_F = \arg\max_{x_F} p(x_F \mid x_E) = \arg\max_{x_F} p(x_E, x_F)$$

Posterior Marginal Densities

$$p(x_F \mid x_E) = \frac{p(x_E, x_F)}{p(x_E)} \qquad p(x_E) = \sum_{x_F} p(x_E, x_F)$$

Provides Bayesian estimators, confidence measures, and sufficient statistics for iterative parameter estimation

Directed Graphical Models



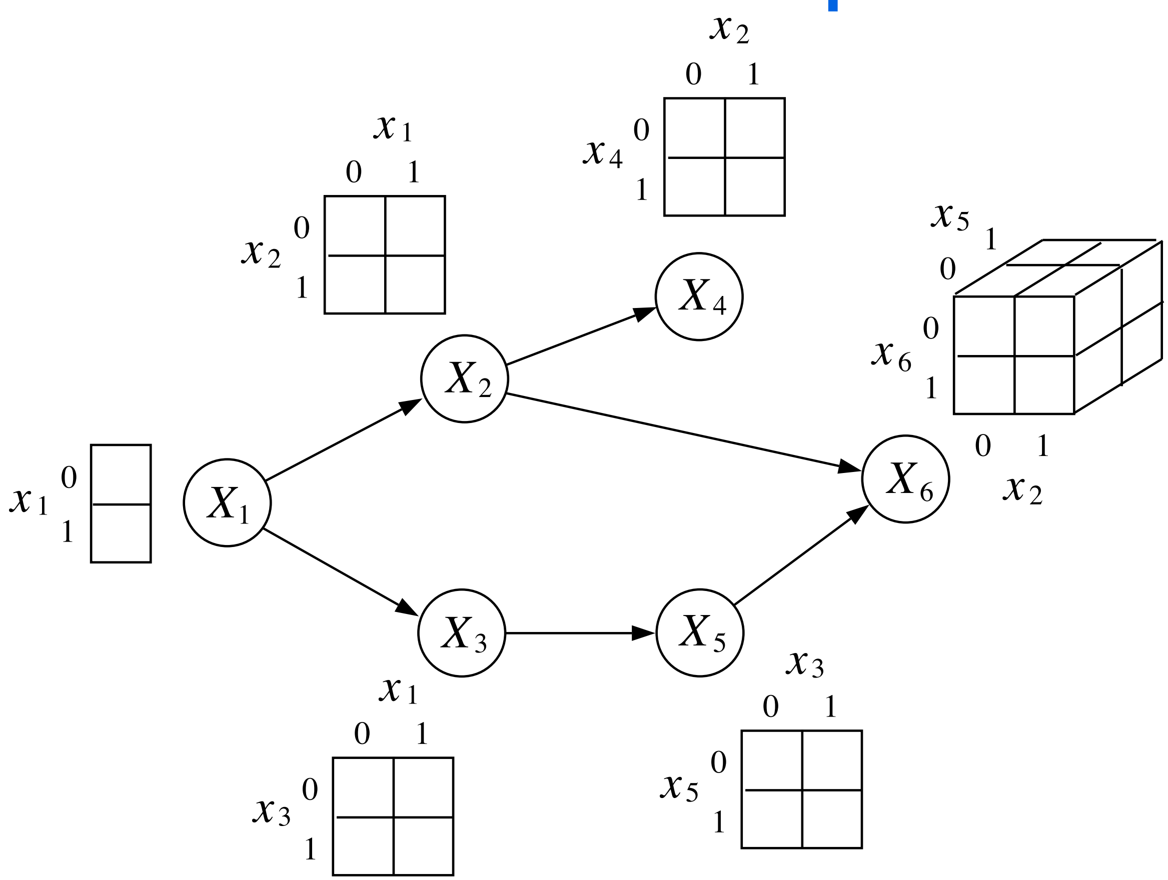
$\mathcal{V} \longrightarrow$ set of N nodes or vertices, $\{1, 2, \ldots, N\}$

$\mathcal{E} \longrightarrow$ set of oriented edges (s,t) linking parents s to children t, so that the set of parents of a node is

$$\mathrm{pa}(t) = \Gamma(t) = \{s \in \mathcal{V} \mid (s,t) \in \mathcal{E}\}$$

$X_s = x_s \longrightarrow$ random variable associated with node s

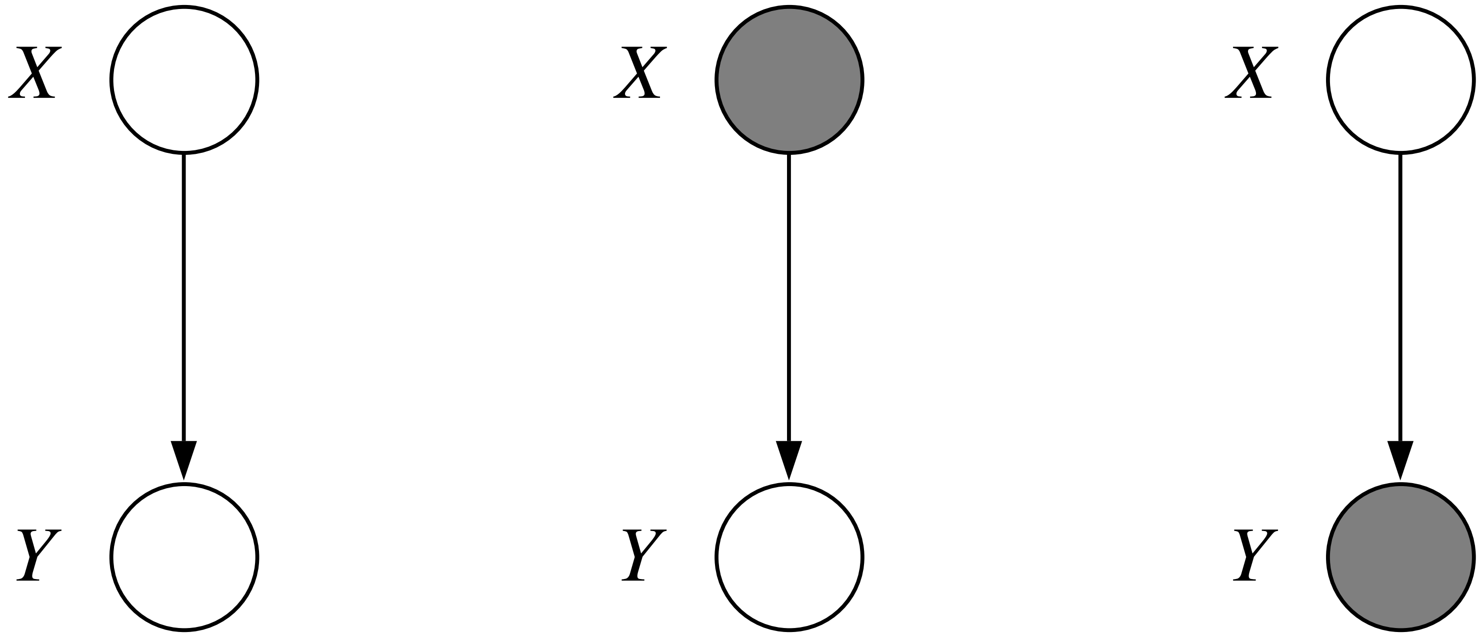
Parameterization & Representation



Representational (storage, learning, computation) Complexity

- *Joint distribution:* Exponential in number of variables
- *Directed graphical model:* Exponential in number of parents ("fan-in") of each node, linear in number of nodes

Inference with Two Variables



$$p(x, y) = p(x)p(y \mid x)$$

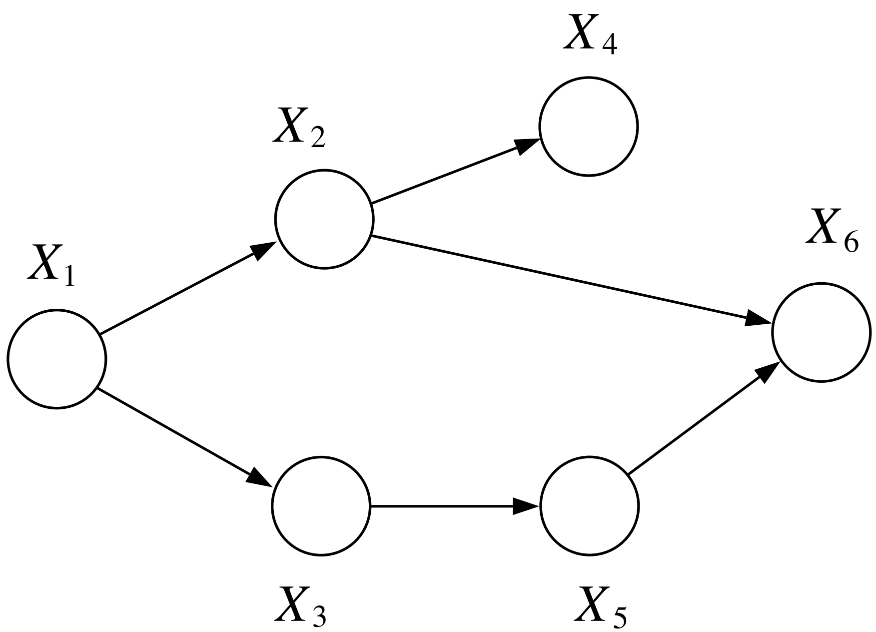
Table Lookup:

$$p(y \mid x = \bar{x})$$

Bayes Rule:

$$p(x \mid y = \bar{y}) = \frac{p(\bar{y} \mid x)p(x)}{p(\bar{y})}$$

Naïve Inference is Intractable



- Suppose each variable takes one of *k* discrete states:

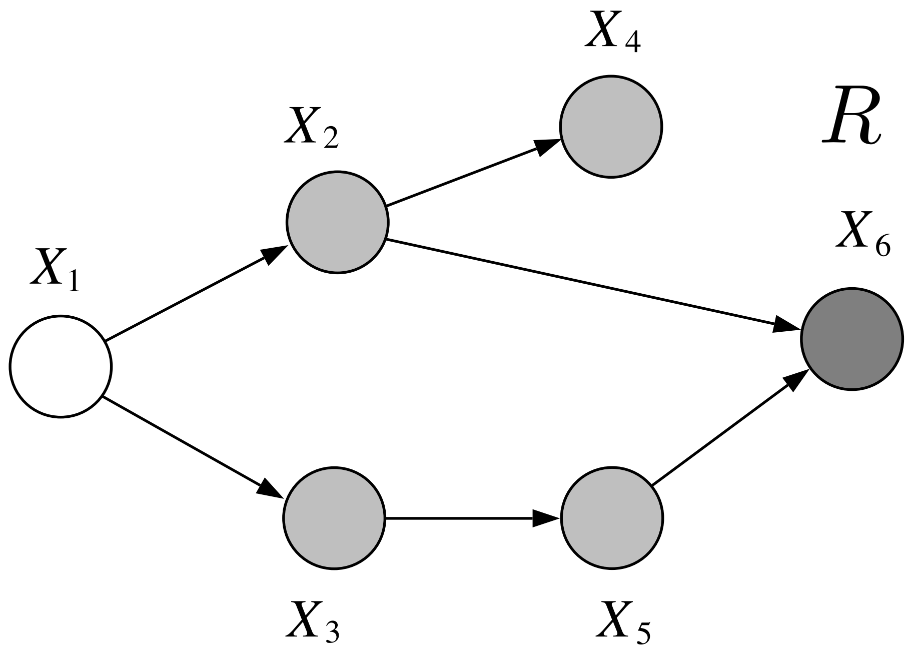
$$p(x_1, x_2, \ldots, x_5) = \sum_{x_6} p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1)p(x_4 \,|\, x_2)p(x_5 \,|\, x_3)p(x_6 \,|\, x_2, x_5)$$

Costs $\mathcal{O}(k)$ operations to update each of $\mathcal{O}(k^5)$ table entries

- Use *factorization* and *distributive law* to reduce complexity:

$$= p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1)p(x_4 \,|\, x_2)p(x_5 \,|\, x_3) \sum_{x_6} p(x_6 \,|\, x_2, x_5)$$

Inference in Directed Graphs



$$F = \{1\}$$

$$p(x_1 \mid \bar{x}_6) = ?$$

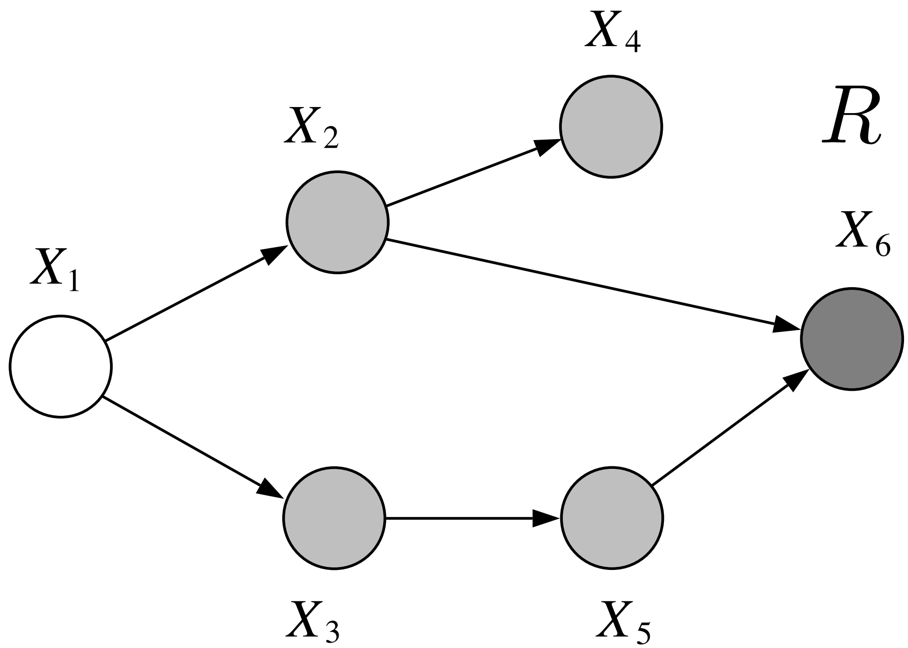
$$R = \{2, 3, 4, 5\}$$

$$E = \{6\}$$

$$X_6 = \bar{x}_6$$

$$\begin{aligned}
p(x_1, \bar{x}_6) &= \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2) p(x_5 \mid x_3) p(\bar{x}_6 \mid x_2, x_5) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) \sum_{x_4} p(x_4 \mid x_2) \sum_{x_5} p(x_5 \mid x_3) p(\bar{x}_6 \mid x_2, x_5) \\
&= p(x_1) \sum_{x_2} p(x_2 \mid x_1) \sum_{x_3} p(x_3 \mid x_1) \sum_{x_4} p(x_4 \mid x_2) m_5(x_2, x_3)
\end{aligned}$$

Inference in Directed Graphs



$$F = \{1\}$$

$$p(x_1 \mid \bar{x}_6) = ?$$

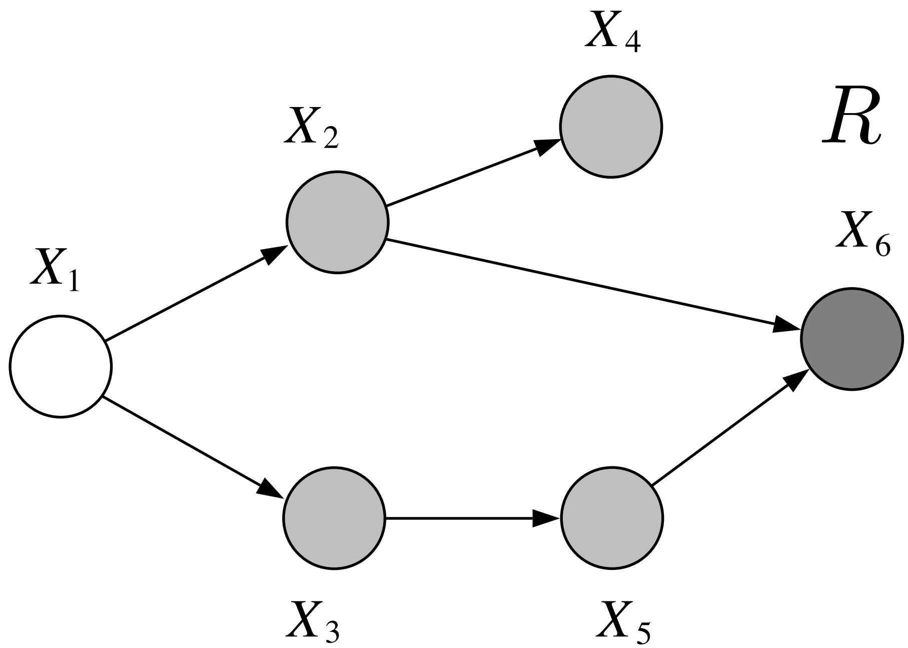
$$R = \{2, 3, 4, 5\}$$

$$E = \{6\}$$

$$X_6 = \bar{x}_6$$

$$\begin{aligned}
p(x_1, \bar{x}_6) &= \sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_5} p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1)p(x_4 \mid x_2)p(x_5 \mid x_3)p(\bar{x}_6 \mid x_2, x_5) \\
&= p(x_1)\sum_{x_2} p(x_2 \mid x_1)\sum_{x_3} p(x_3 \mid x_1)m_5(x_2, x_3)\sum_{x_4} p(x_4 \mid x_2) \\
&= p(x_1)\sum_{x_2} p(x_2 \mid x_1)m_4(x_2)\sum_{x_3} p(x_3 \mid x_1)m_5(x_2, x_3).
\end{aligned}$$

Inference in Directed Graphs



$$F = \{1\}$$

$$p(x_1 \mid \bar{x}_6) = ?$$

$$R = \{2, 3, 4, 5\}$$

$$E = \{6\}$$

$$X_6 = \bar{x}_6$$

$$\begin{aligned} p(x_1, \bar{x}_6) &= \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2) p(x_5 \mid x_3) p(\bar{x}_6 \mid x_2, x_5) \\ &= p(x_1) \sum_{x_2} p(x_2 \mid x_1) m_4(x_2) m_3(x_1, x_2) \\ &= p(x_1) m_2(x_1). \end{aligned}$$

$$p(x_1 \mid \bar{x}_6) = \frac{p(x_1) m_2(x_1)}{\sum_{x_1} p(x_1) m_2(x_1)}$$

$$p(\bar{x}_6) = \sum_{x_1} p(x_1) m_2(x_1)$$

Evidence Potentials

$$g(\bar{x}_i) = \sum_{x_i} g(x_i)\delta(x_i, \bar{x}_i)$$

$$\delta(x_i, \bar{x}_i) = 1 \text{ if } x_i = \bar{x}_i$$
$$\delta(x_i, \bar{x}_i) = 0 \text{ if } x_i \neq \bar{x}_i$$

$$m_6(x_2, x_5) = \sum_{x_6} p(x_6 \,|\, x_2, x_5)\delta(x_6, \bar{x}_6) = p(\bar{x}_6 \mid x_2, x_5)$$

- Encoding observations via evidence potentials:

$$\delta(x_E, \bar{x}_E) \triangleq \prod_{i \in E} \delta(x_i, \bar{x}_i)$$

$$p^E(x) \triangleq p(x)\delta(x_E, \bar{x}_E)$$

$$p(x_F, \bar{x}_E) = \sum_{x_E} p(x_F, x_E)\delta(x_E, \bar{x}_E)$$

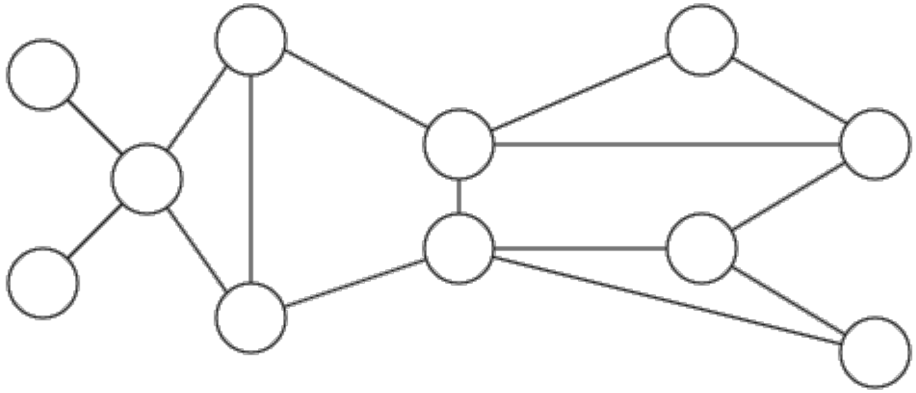
$$p(\bar{x}_E) = \sum_{x_F} \sum_{x_E} p(x_F, x_E)\delta(x_E, \bar{x}_E).$$

- For undirected graphical models:

$$p^E(x) \triangleq \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi^E_{X_C}(x_C)$$

$$\psi^E_i(x_i) \triangleq \psi_i(x_i)\delta(x_i, \bar{x}_i)$$

Undirected Graphical Models



- A *clique* is a fully connected subset of nodes

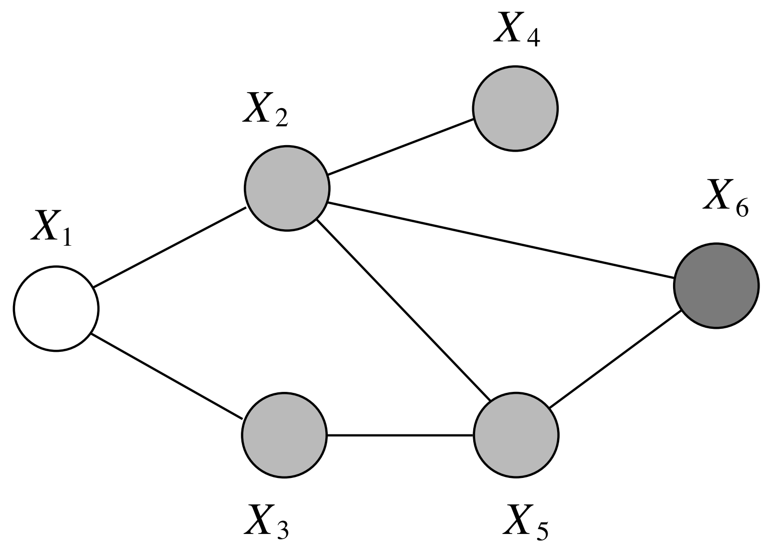
Theorem 2.2.1 (Hammersley-Clifford). *Let $\mathcal{C}$ denote the set of cliques of an undirected graph $\mathcal{G}$. A probability distribution defined as a normalized product of non-negative potential functions on those cliques is then always Markov with respect to $\mathcal{G}$:*

$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c) \tag{2.71}$$

Conversely, any strictly positive density ($p(x) > 0$ for all x) which is Markov with respect to $\mathcal{G}$ can be represented in this factored form.

- It is possible, but not necessary, to restrict factorization only to the *maximal cliques* (not strict subsets of other cliques)

Inference in Undirected Graphs



$$
\begin{aligned}
p(x_1, \bar{x}_6) &= \frac{1}{Z} \sum_{x_2} \sum_{x_3} \sum_{x_4} \sum_{x_5} \sum_{x_6} \psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)\psi(x_3, x_5)\psi(x_2, x_5, x_6)\delta(x_6, \bar{x}_6) \\
&= \frac{1}{Z} \sum_{x_2} \psi(x_1, x_2) \sum_{x_3} \psi(x_1, x_3) \sum_{x_4} \psi(x_2, x_4) \sum_{x_5} \psi(x_3, x_5) \sum_{x_6} \psi(x_2, x_5, x_6)\delta(x_6, \bar{x}_6) \\
&= \frac{1}{Z} \sum_{x_2} \psi(x_1, x_2) \sum_{x_3} \psi(x_1, x_3) \sum_{x_4} \psi(x_2, x_4) \sum_{x_5} \psi(x_3, x_5) m_6(x_2, x_5) \\
&= \frac{1}{Z} \sum_{x_2} \psi(x_1, x_2) \sum_{x_3} \psi(x_1, x_3) m_5(x_2, x_3) \sum_{x_4} \psi(x_2, x_4) \\
&= \frac{1}{Z} \sum_{x_2} \psi(x_1, x_2) m_4(x_2) \sum_{x_3} \psi(x_1, x_3) m_5(x_2, x_3) \\
&= \frac{1}{Z} \sum_{x_2} \psi(x_1, x_2) m_4(x_2) m_3(x_1, x_2) \;=\; \frac{1}{Z} m_2(x_1)
\end{aligned}
$$

A Graph Elimination Algorithm

Algebraic Marginalization Operations

- Marginalize out the variable associated with sum node
- Compute a new potential table involving all other variables which depend on the just-marginalized variable

Graph Manipulation Operations

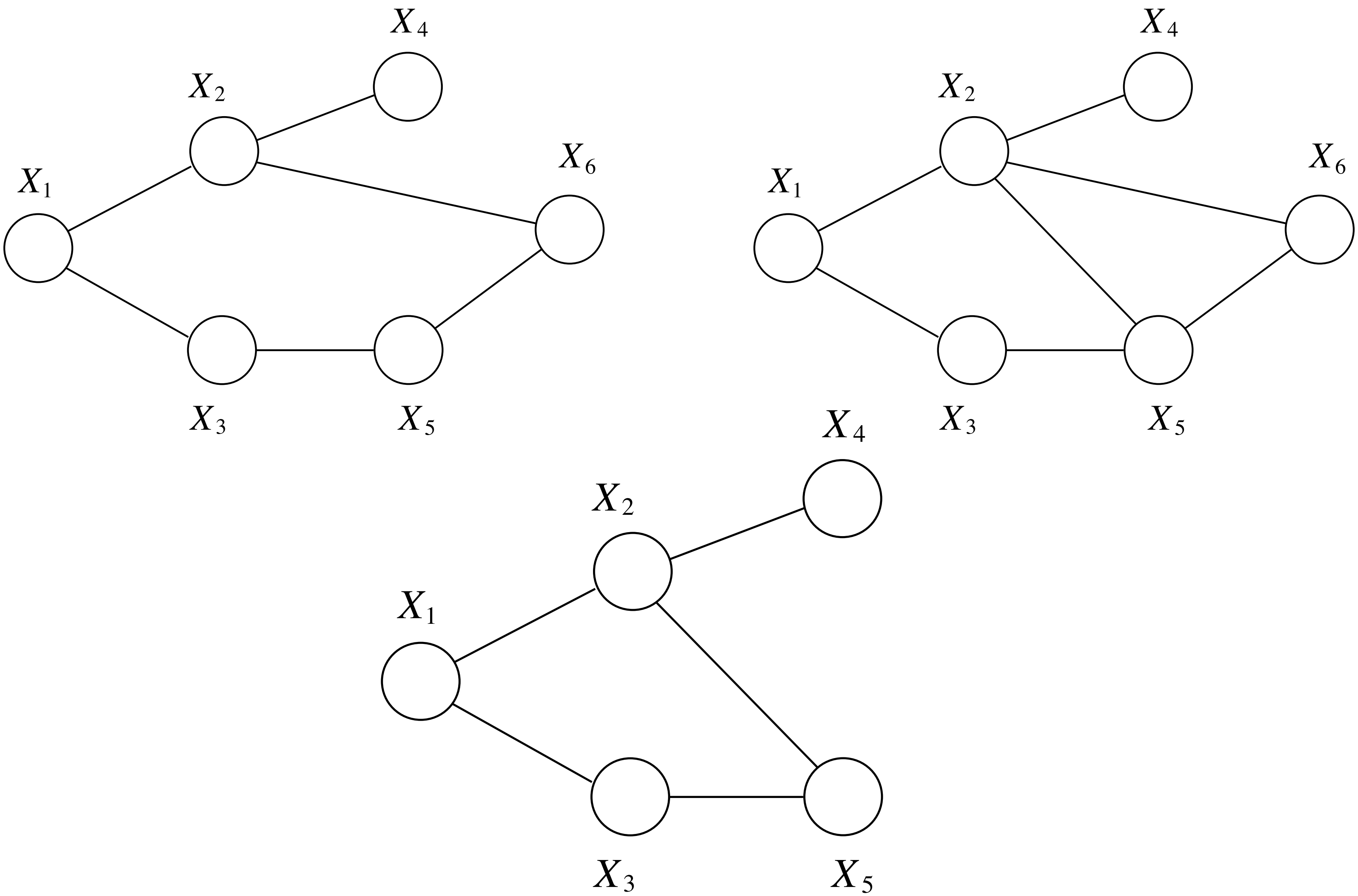
- Remove, or *eliminate*, a single node from the graph
- Add edges (if they don’t already exist) between all pairs of nodes who were neighbors of the just-removed node

A Graph Elimination Algorithm

- Choose an elimination ordering (query nodes should be last)
- Eliminate a node, remove its incoming edges, add edges between all pairs of its neighbors
- Iterate until all non-query nodes are eliminated

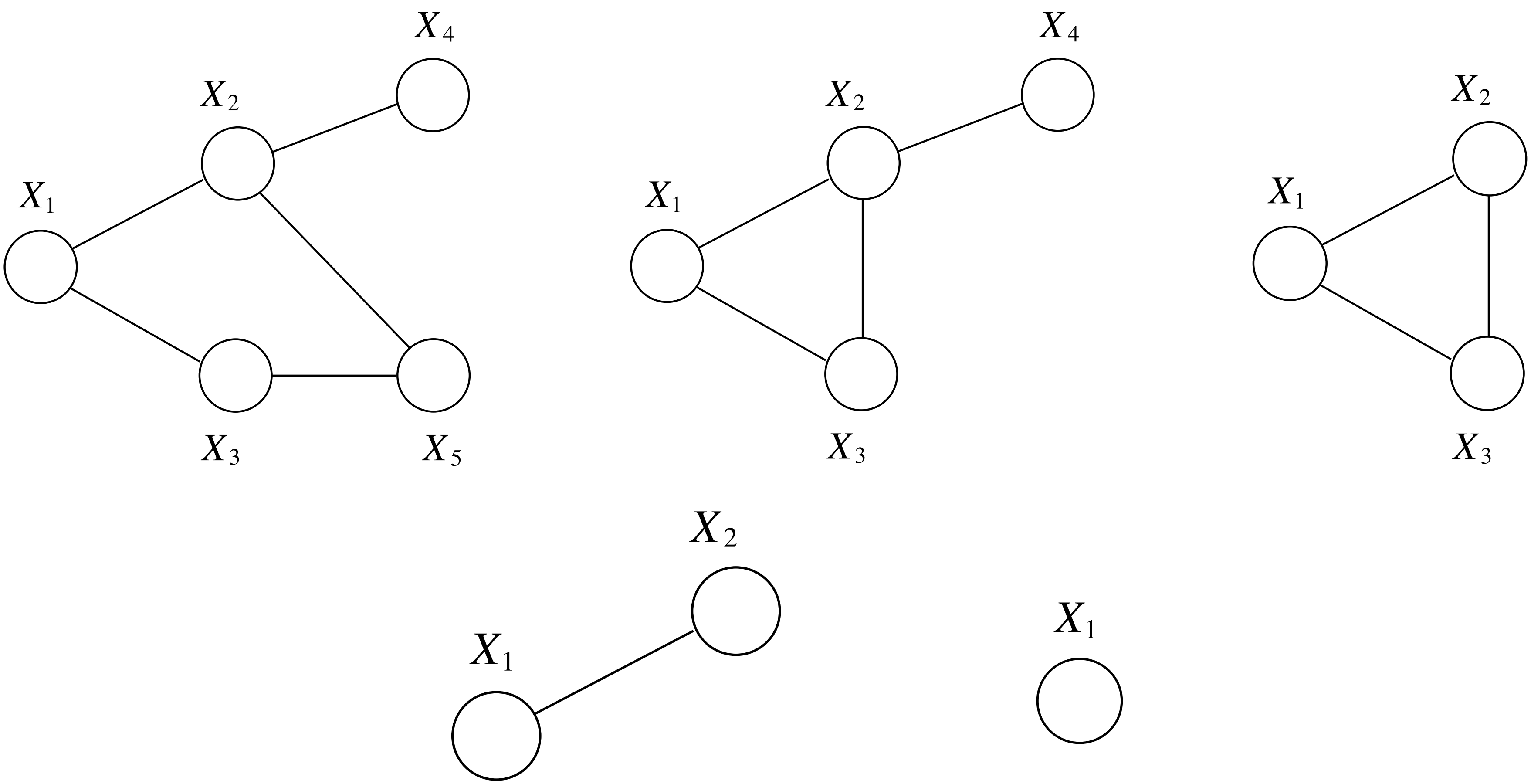
Graph Elimination Example

Elimination Order: (6,5,4,3,2,1)

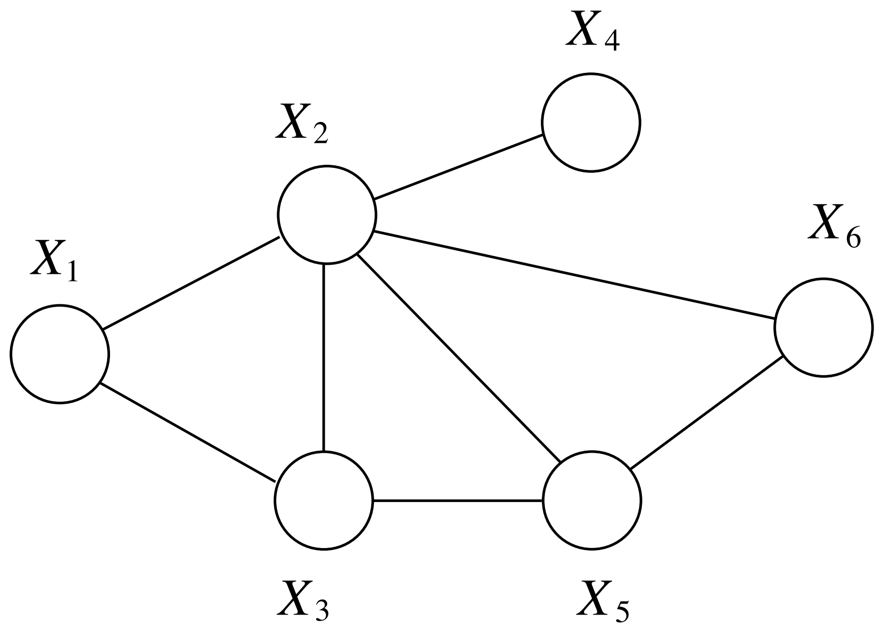


Graph Elimination Example

Elimination Order: (6,5,4,3,2,1)

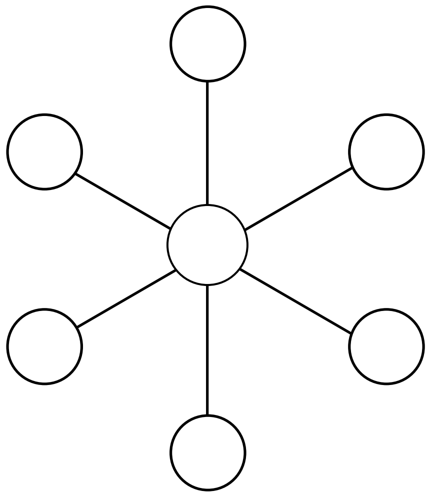


Elimination Algorithm Complexity

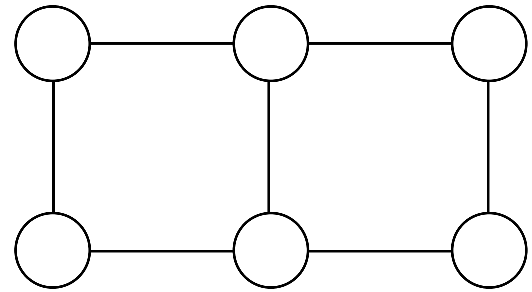


- *Elimination cliques:* Sets of neighbors of eliminated nodes
- *Marginalization cost:* Exponential in number of variables in each elimination clique (dominated by largest clique)
- *Treewidth of graph*: Over all possible elimination orderings, the smallest possible max-elimination-clique size, minus one
- *NP-Hard:* Finding the best elimination ordering for an arbitrary input graph (but heuristic algorithms often effective)

Elimination Order Matters

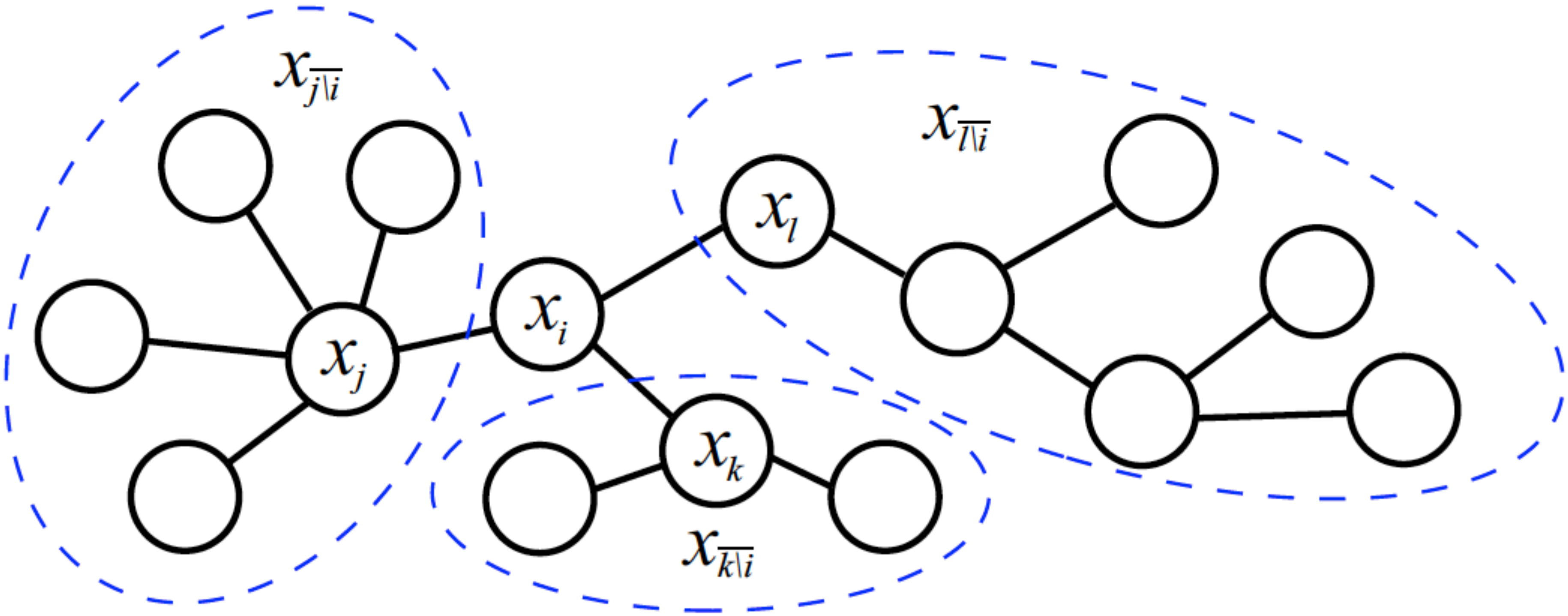


Treewidth = 1



Treewidth = 2

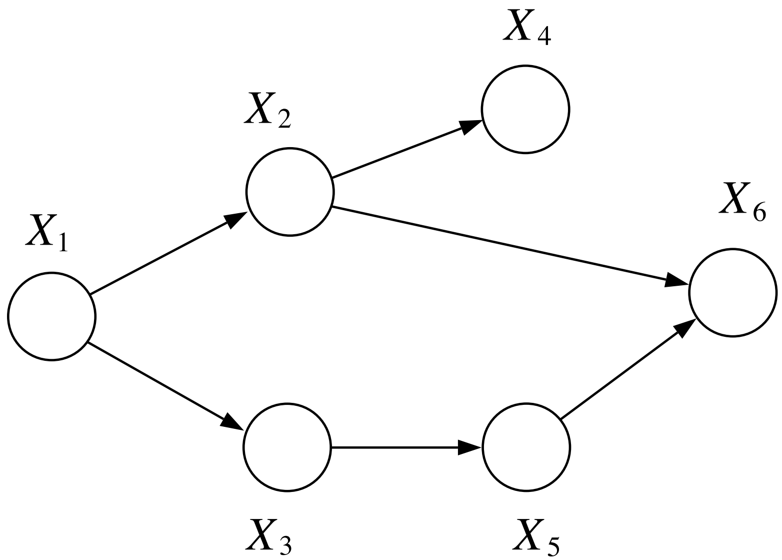
Elimination in Undirected Trees



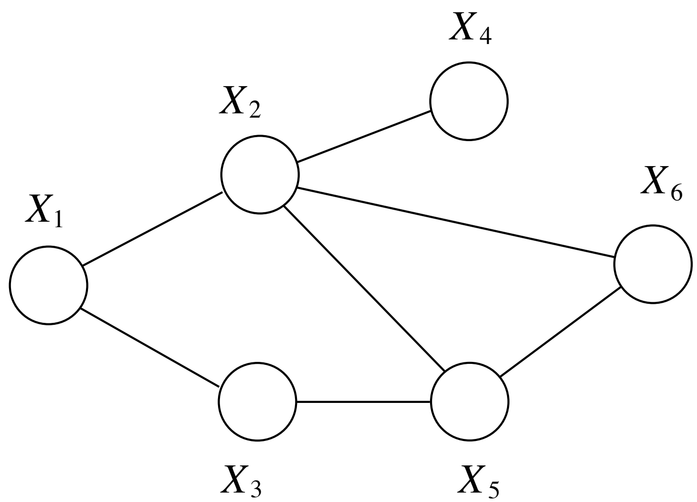
$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s)$$

Cost linear in number of nodes, quadratic in number of states

Directed to Undirected Graphs



Directed Graph

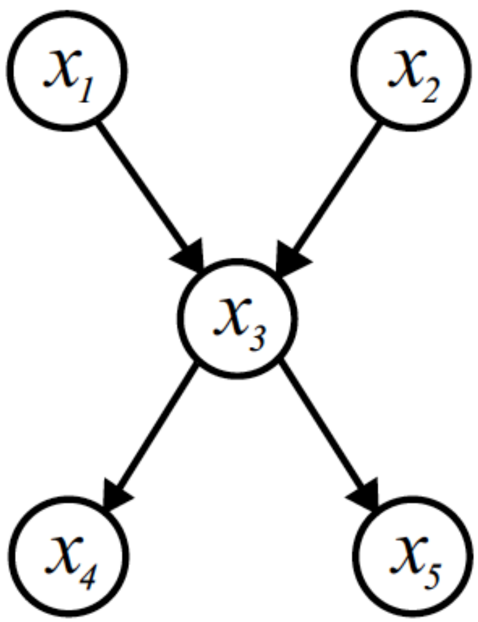


Moral Graph

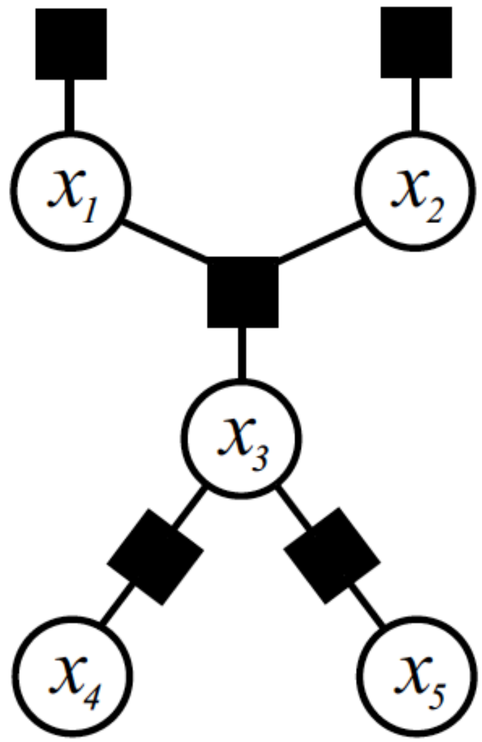
MORALIZE(G)
  for each node X_i in I
    connect all of the parents of X_i
  end
  drop the orientation of all edges
  return G

- Moral graph links ("marries") all parents with a common child
- Any directed graphical model factorizes according to the cliques of the resulting undirected graph, and is thus Markov

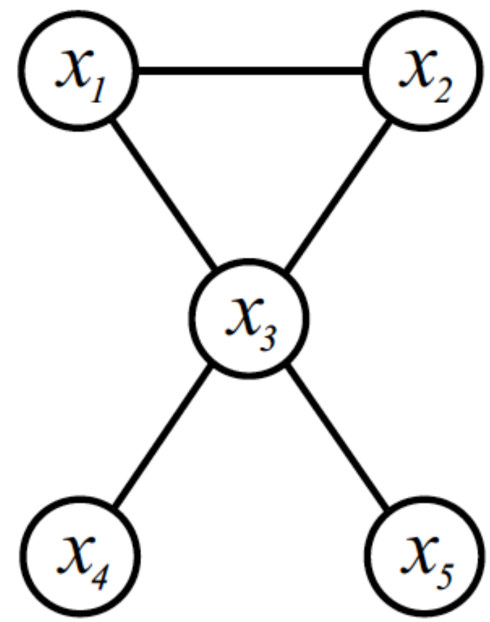
Types of Graphical Models



Directed



Factor



Undirected

Factor Graphs Allow Fine-grained Factorization

$$p(x) = \frac{1}{Z} \prod_{f \in \mathcal{F}} \psi_f(x_f)$$

- Each potential, or *factor*, depends on a subset of nodes *f*
- Create factor nodes (black squares) linked to dependent variable nodes, resulting in bipartite *factor graph*