Computer Science

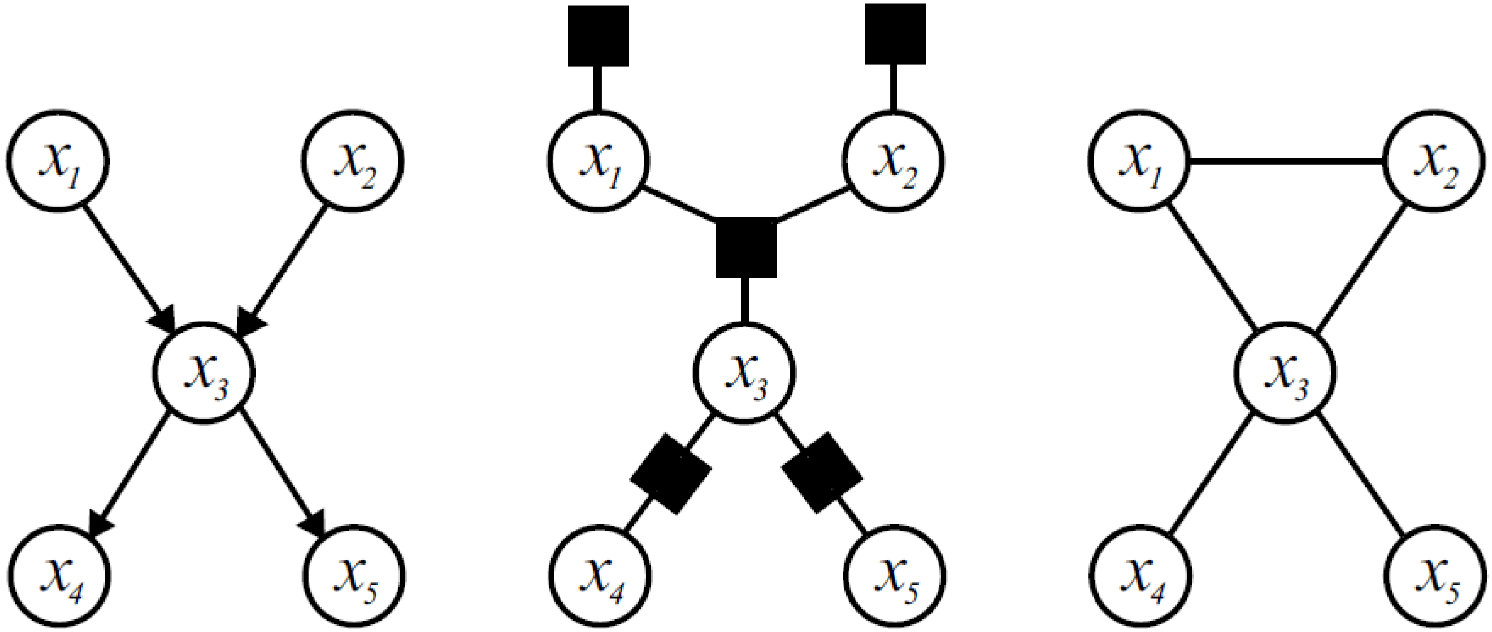
CSC696H: Advanced Topics in Probabilistic Graphical Models

Probabilistic Graphical Models

Prof. Jason Pacheco

Graphical Models

A variety of graphical models can represent the same probability distribution



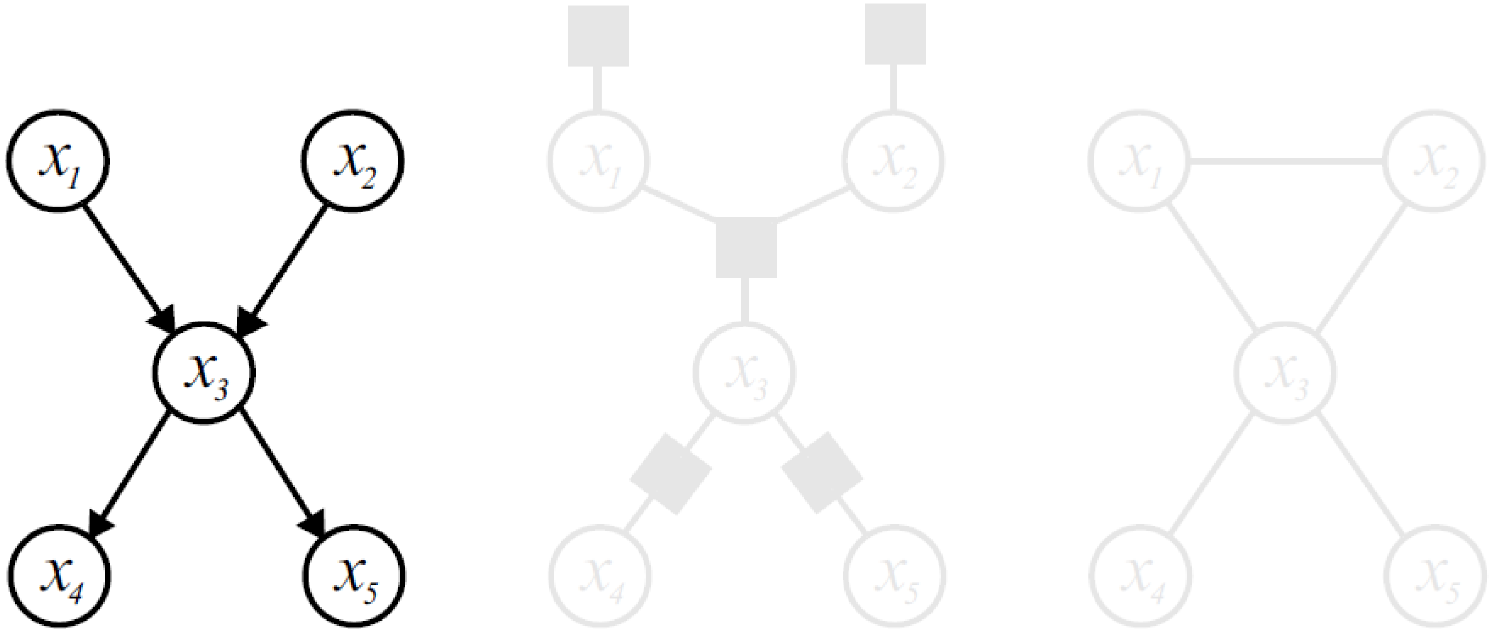
Bayes Network **Factor Graph** **Markov Random Field**

Directed Models **Undirected Models**

[Source: Erik Sudderth, PhD Thesis]

Graphical Models

A variety of graphical models can represent the same probability distribution



Bayes Network **Factor Graph** **Markov Random Field**

Directed Models **Undirected Models**

[Source: Erik Sudderth, PhD Thesis]

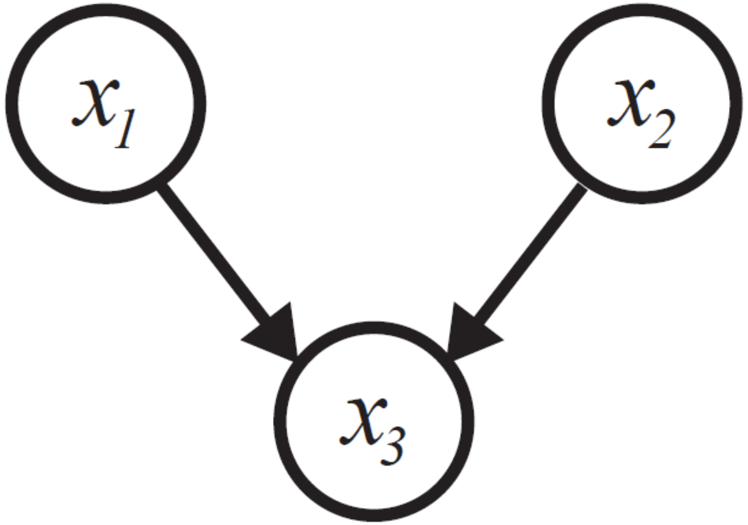
From Probabilities to Pictures

A probabilistic graphical model allows us to pictorially represent a probability distribution

Probability Model:

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$

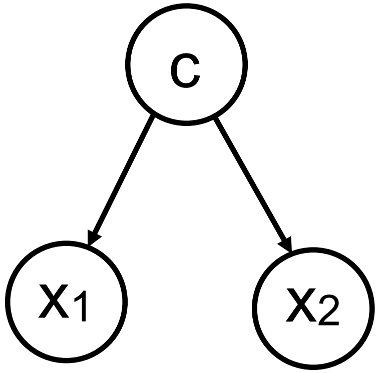
Graphical Model:



Conditional distribution on each RV is dependent on its parent nodes in the graph

Ancestral Sampling

Directed models describe data generation process…



$$p(C, X_1, X_2) = p(C)p(X_1 \mid C)p(X_2 \mid C)$$

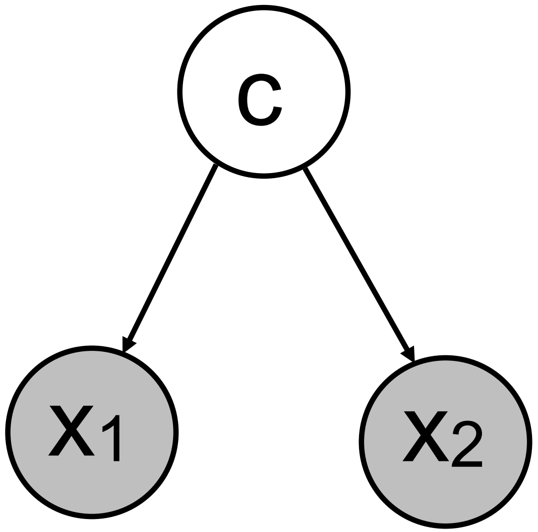
The graph and the formula say exactly the same thing.
(The graph has very specific semantics.)

Step 1 Sample root node (prior): $c \sim p(C)$

Step 2 Sample children, given sample of parent (likelihood):

$$x_1 \sim p(X_1 \mid C = c) \qquad x_2 \sim p(X_2 \mid C = c)$$

Inference



Denote observed data with shaded nodes,

$$X_1 = x_1 \qquad X_2 = x_2$$

Infer *latent* variable C via Bayes' rule:

$$p(c \mid x_1, x_2) = \frac{p(c)p(x_1 \mid c)p(x_2 \mid c)}{p(x_1, x_2)}$$

- This is (obviously) a simple example
- Models and inference task can get really complicated
- But the fundamental concepts and approach are the same

Chain Rule of Probability

Recall the **probability chain rule** says that we can decompose any joint distribution as a product of conditionals….

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)$$

Valid for *any ordering* of the random variables…

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4)$$

For a collection of N RVs and any permutation ρ:

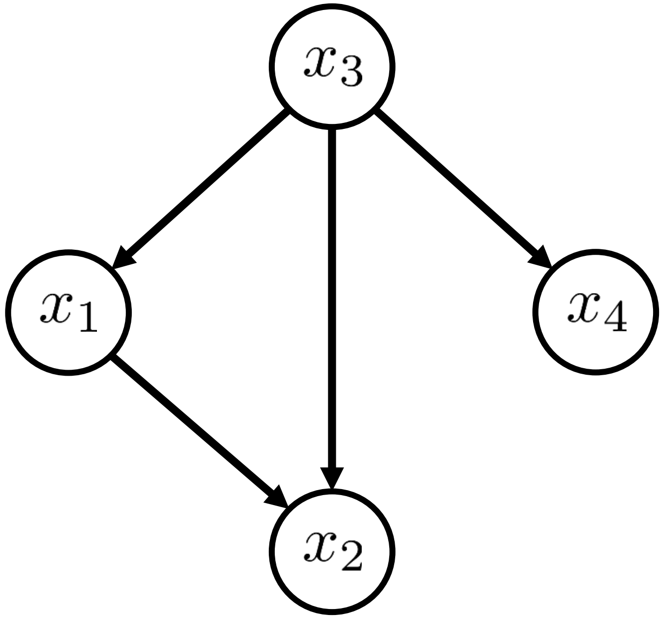
$$p(x_1, \ldots, x_N) = p(x_{\rho(1)}) \prod_{i=2}^{N} p(x_{\rho(i)} \mid x_{\rho(i-1)}, \ldots, x_{\rho(1)})$$

Conditional Independence

Recall two RVs X and Y are **conditionally independent** given Z (or $X \perp Y \mid Z$) iff:

$$p(X \mid Y, Z) = p(X \mid Z)$$

Idea Apply *chain rule* with ordering that exploits conditional independencies to simplify the terms



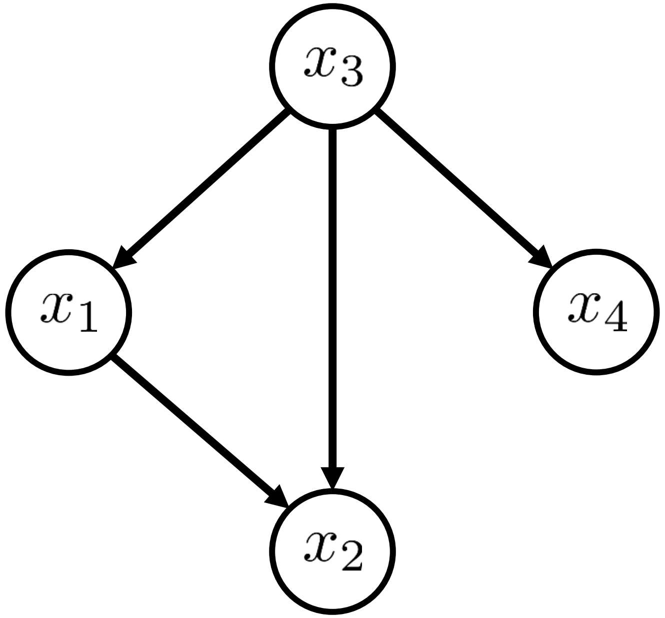
Ex. Suppose $x_4 \perp x_1 \mid x_3$ and $x_2 \perp x_4 \mid x_1$ then:

$$\begin{aligned} p(x) &= p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4) \\ &= p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3) \end{aligned}$$

Can visualize conditional dependencies using ***directed acyclic graph*** *(DAG)*

General Directed Graphs

Def. A directed graph is a graph with edges $(s,t) \in \mathcal{E}$ (arcs) connecting parent vertex $s \in \mathcal{V}$ to a child vertex $t \in \mathcal{V}$



Def. Parents of vertex $t \in \mathcal{V}$ are given by the set of nodes with arcs pointing to t ,

$$\mathrm{Pa}(t) = \{s : (s,t) \in \mathcal{E}\}$$

Children of $t \in \mathcal{V}$ are given by the set,

$$\mathrm{Ch}(t) = \{t : (t,k) \in \mathcal{E}\}$$

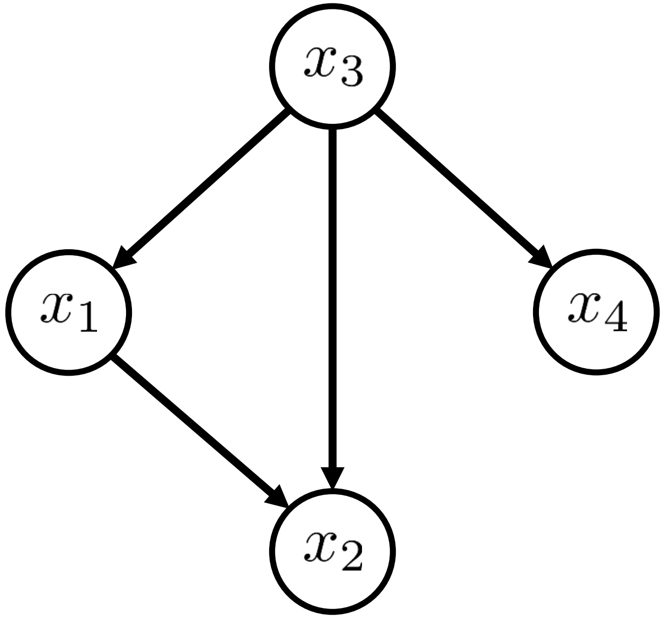
Ancestors are parents-of-parents.
Descendants are children-of-children.

Directed PGM = Bayes Network

Model factors are normalized conditional distributions:

$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\mathrm{Pa}(s)})$$

Parents of node *s*



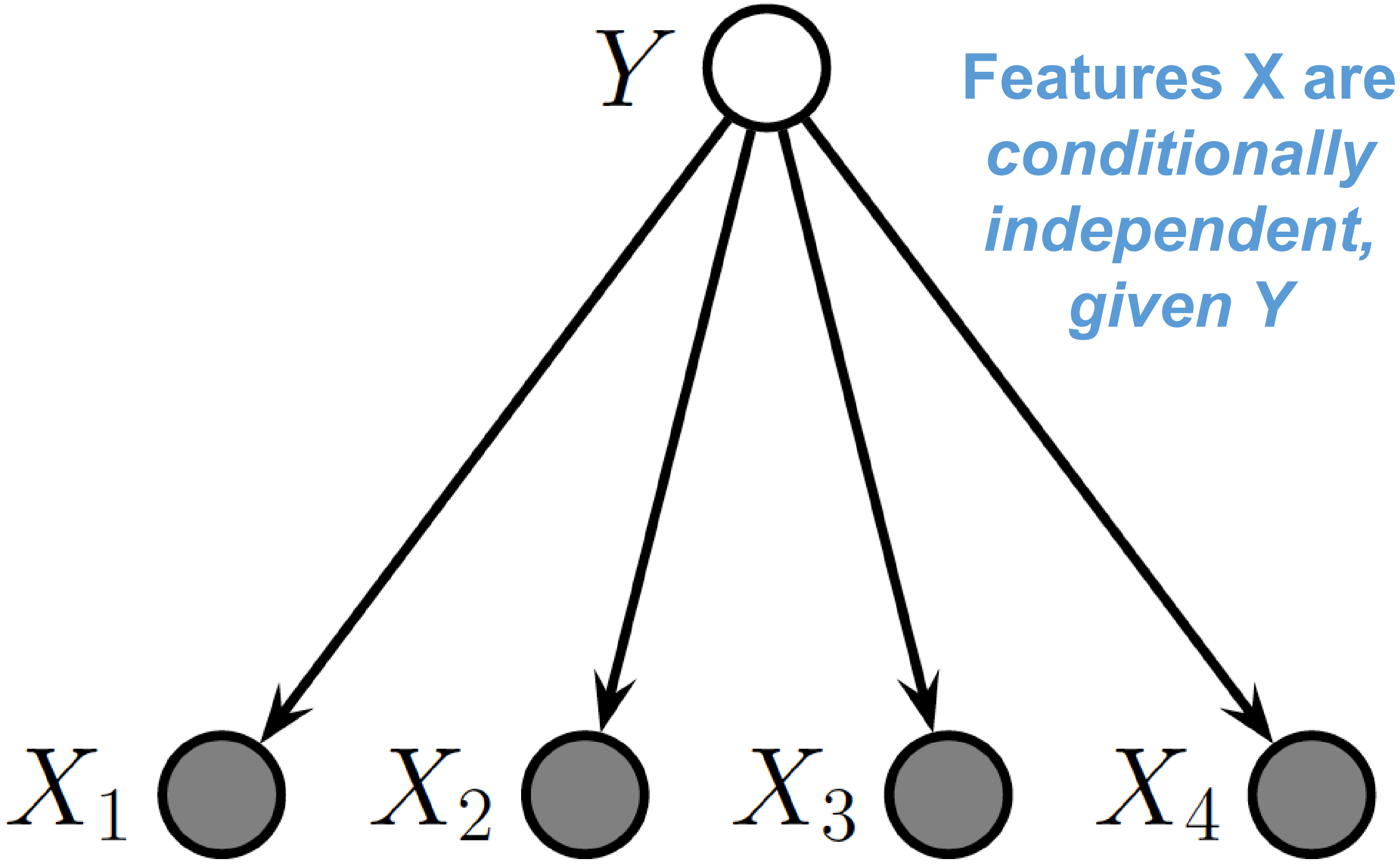
Directed acyclic graph (DAG) specifies factorized form of joint probability:

$$p(x) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3)$$

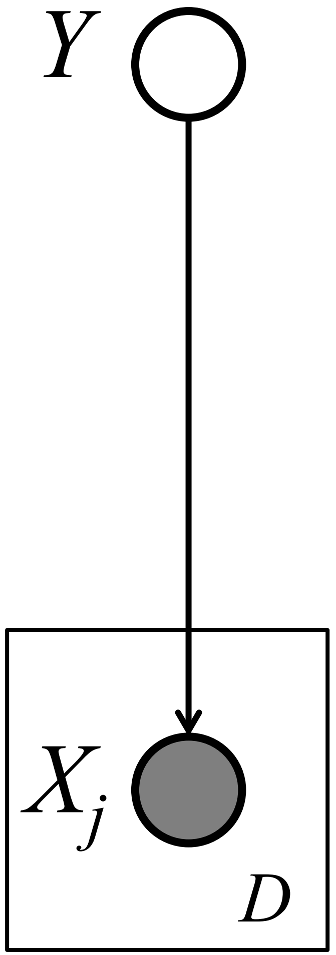
Locally normalized factors yield *globally normalized* joint probability

Shading & Plate Notation

Convention: *Shaded nodes are observed, open nodes are latent/hidden/unobserved*



Features X are *conditionally independent, given Y*



Plates denote replication of random variables

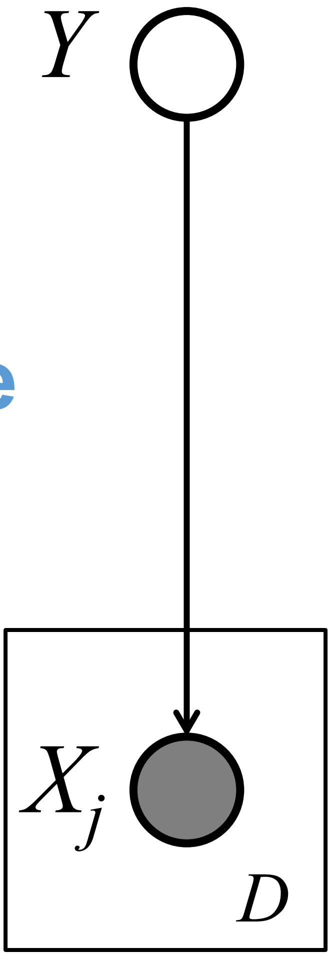
$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y)$$

Question Does anybody know the name for this model? **Naïve Bayes**

Inference

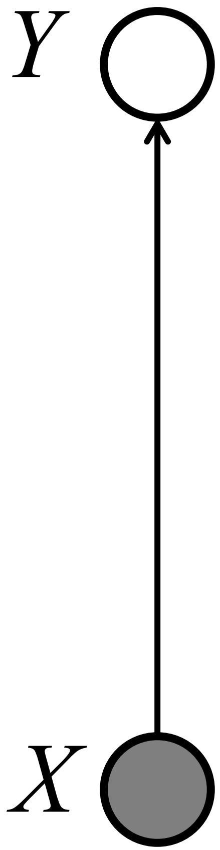
Interpret inference as <u>inverting arrows</u> in the graphical model

Naïve Bayes Generative Model



$$p(y,\mathbf{x}) = p(y)\prod_{j=1}^{D} p(x_j|y)$$

Posterior Model



$$p(y,\mathbf{x}) = \overbrace{p(y \mid \mathbf{x})}^{\text{Posterior}}\,\overbrace{p(\mathbf{x})}^{\text{Marginal Likelihood}}$$

Example: Gaussian Mixture Model

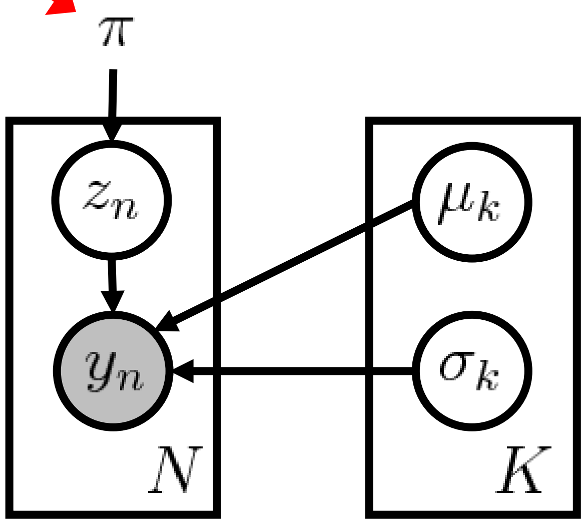
Bayes nets are easily simulated via <u>ancestral sampling</u>…

Fixed parameters
No Circle

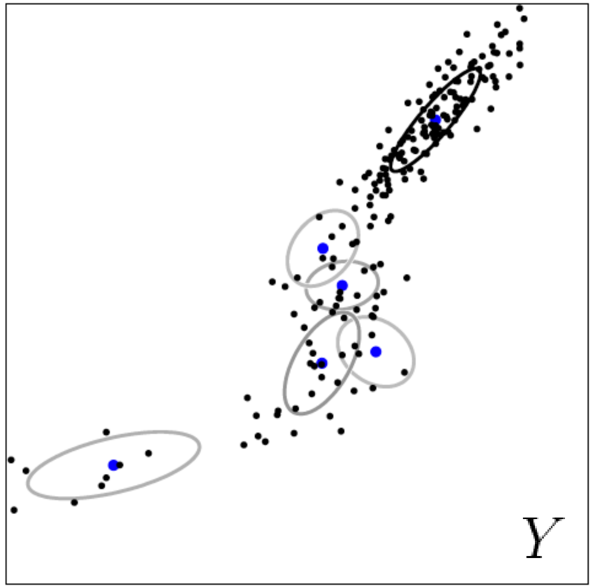
<u>Probability Model</u>

$$\mu_k \sim \mathcal{N}(\cdot)$$
$$\sigma_k \sim \text{Inv-Gamma}(\cdot)$$
$$z_n \mid \pi \sim \text{Cat}(\pi)$$
$$y_n \mid z_n, \mu_{z_n}, \sigma_{z_n} \sim \mathcal{N}(\mu_{z_n}, \sigma_{z_n})$$

<u>Bayes Net</u>



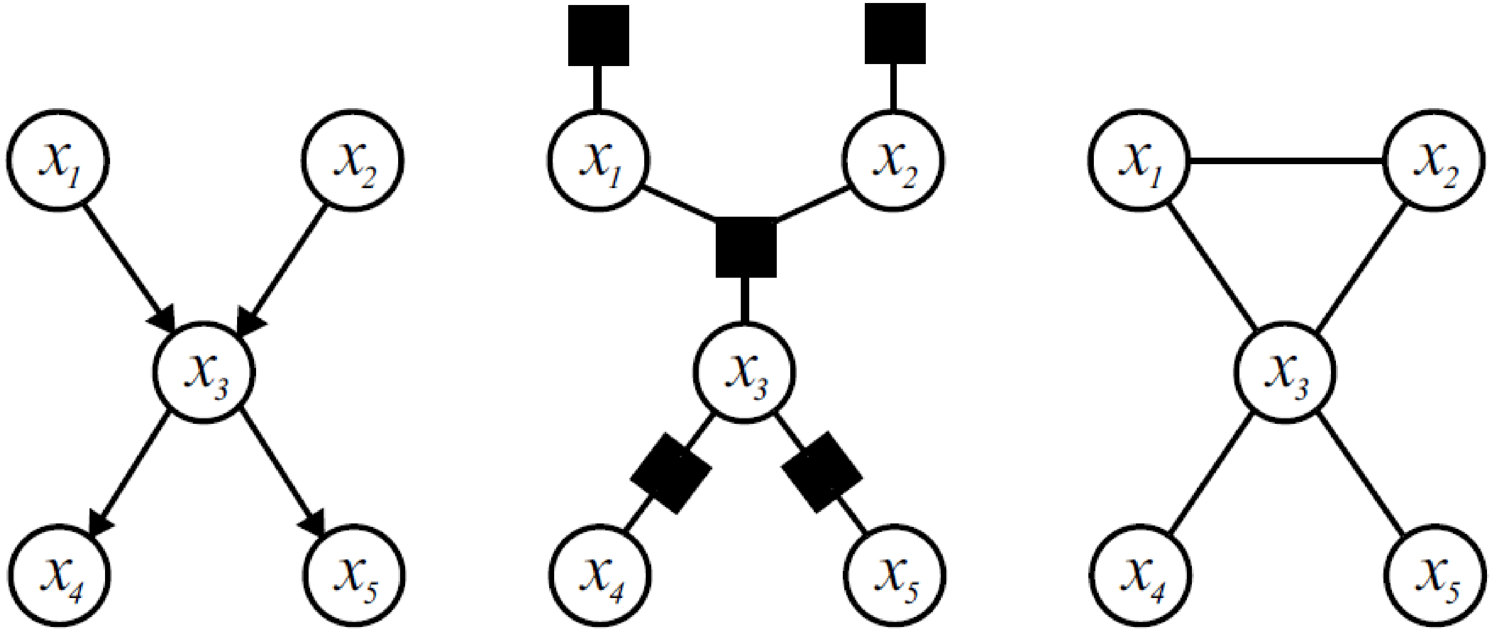
<u>Joint Sample</u>



Sample all nodes with no parents, then children, etc., to terminals. Can sample nodes at same level in parallel.

Graphical Models

A variety of graphical models can represent the same probability distribution



Bayes Network **Factor Graph** **Markov Random Field**

Directed Models

Undirected Models

[Source: Erik Sudderth, PhD Thesis]

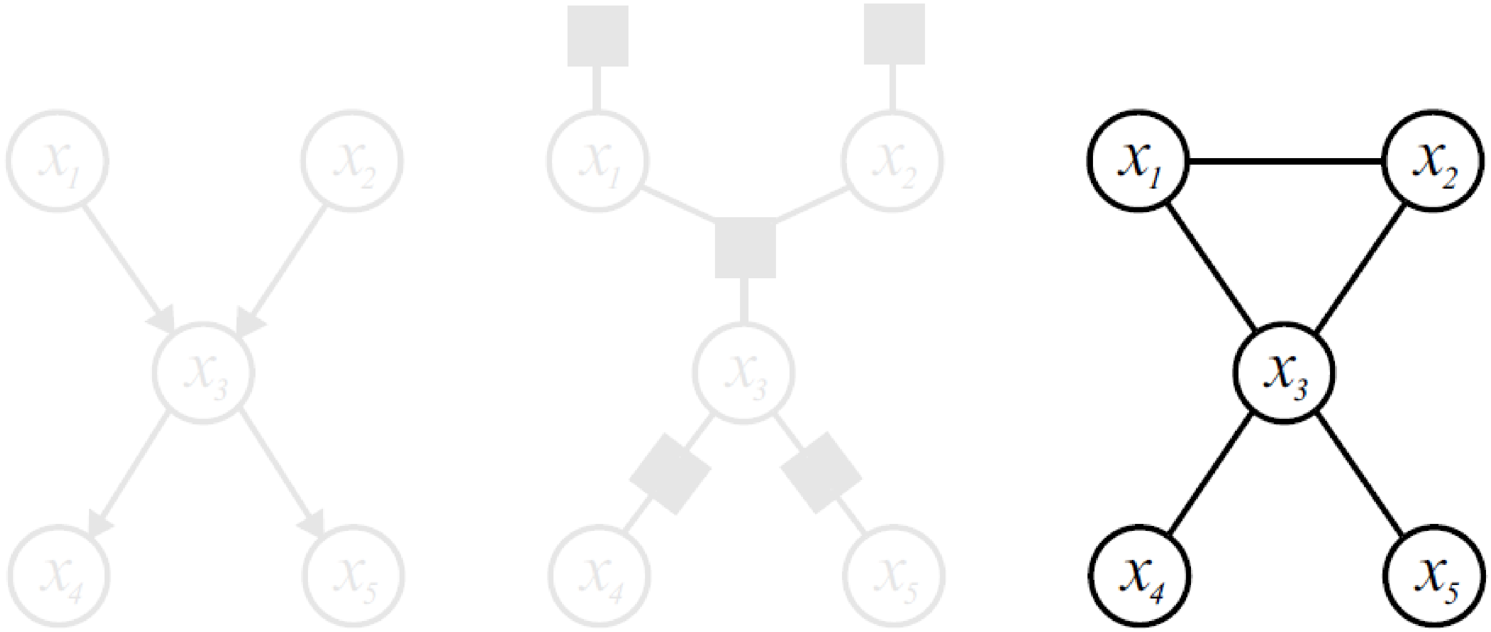
Administrative Items

- Sign up for paper presentation before Wed 9/7
  - Reply to thread on Piazza
  - Don’t wait… otherwise you will be assigned by default

- Create Github repository
  - Title “CSC969H Fall 2022 – <Name>”
  - Add Markdown document “critical_summary.md”
  - Add me as collaborator “pachecoj”
  - Set repository as Private
  - I will add this to D2L as a grade item

Graphical Models

A variety of graphical models can represent the same probability distribution



Bayes Network

Factor Graph

Markov Random Field

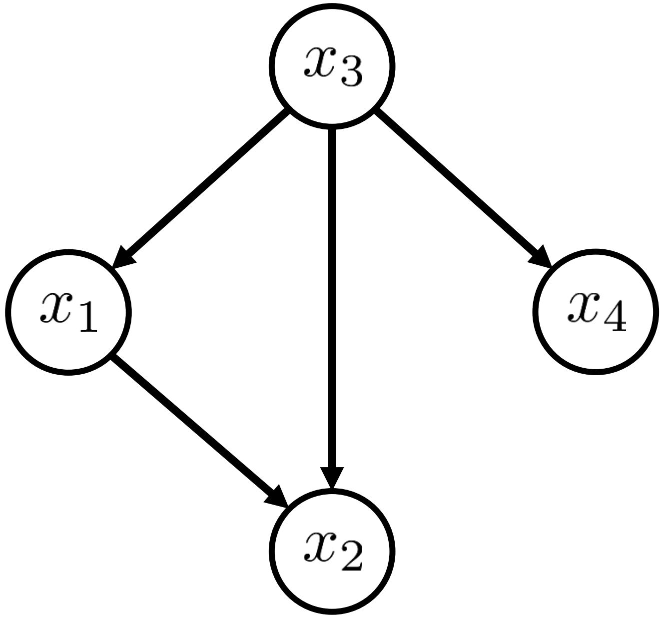
Directed Models

Undirected Models

[Source: Erik Sudderth, PhD Thesis]

Directed PGM = Bayes Network

Model factors are normalized conditional distributions:



$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\mathrm{Pa}(s)})$$

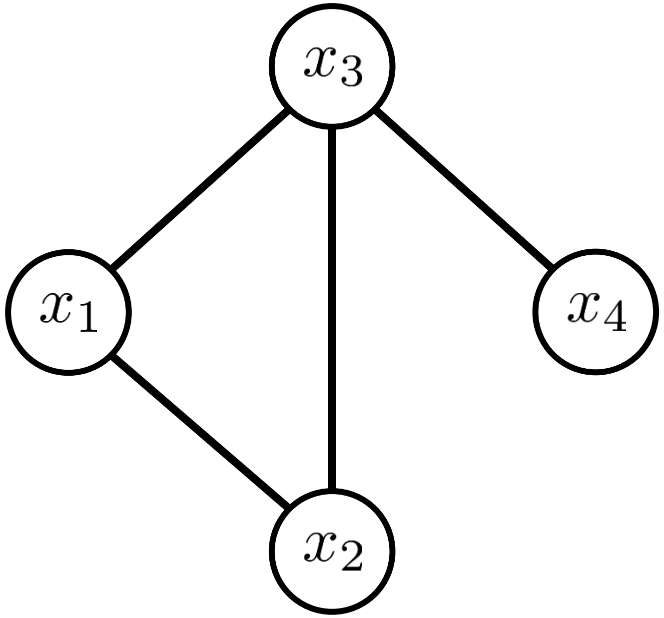
Parents of node *s*

Locally normalized factors yield *globally normalized* joint probability

Often difficult to specify joint in terms of product of normalized probabilities…

Markov Random Field

Specify joint as product of unnormalized functions…



Functions model how variables interact

$$p(x) = \frac{1}{Z}\psi_a(x_1, x_2, x_3)\psi_b(x_3, x_4)$$

Global normalization constant

Potential functions ψ and are non-negative and their product is normalizable…**they are not unnormalized probabilities!**

- More general class of models than Bayes Nets
- Any Bayes Net easily converts to MRF by dropping local normalizers
- MRF→Bayes Net not as straightforward

Factorized Probability Distributions

A probability distribution over RVs $x = (x_1, \ldots, x_d)$ can be written as a product of factors,

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

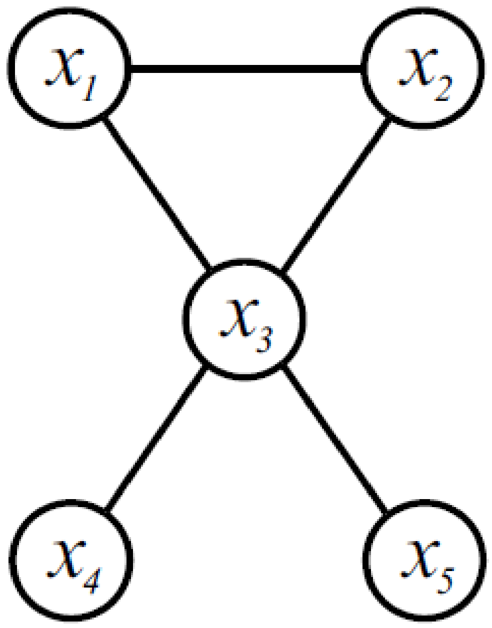
<u>Where:</u>

- $\mathcal{C}$ a collection of subsets of indices $\{1, \ldots, d\}$
- $\psi(\cdot)$ are nonnegative *factors* (or *potential functions*)
- Z the normalizing constant (or *partition function*)

$$Z = \int \prod_{c \in \mathcal{C}} \psi_c(x_c) \, dx_c$$

Undirected Graphical Models

A **graph** $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a set of vertices $\mathcal{V}$ and edges $\mathcal{E}$. An edge $(s,t) \in \mathcal{E}$ connects two vertices $s, t \in \mathcal{V}$.



In **undirected models** edges are specified irrespective of node ordering so that,

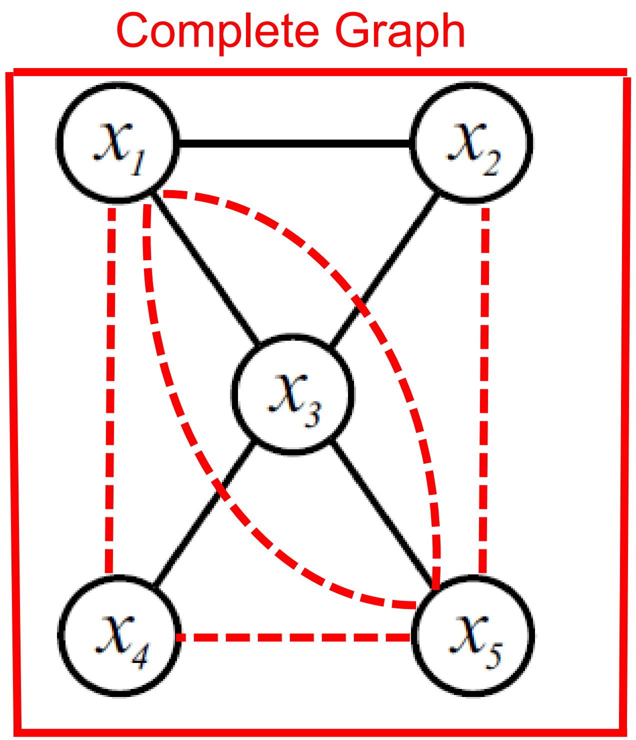
$$(s,t) \in \mathcal{E} \Leftrightarrow (t,s) \in \mathcal{E}$$

Distributions are typically specified with unknown normalization (easier to specify),

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

Markov Random Fields (MRFs)

A factor $\psi_c(x_c)$ corresponds to a clique $c \in \mathcal{C}$ (fully connected subgraph) in the MRF

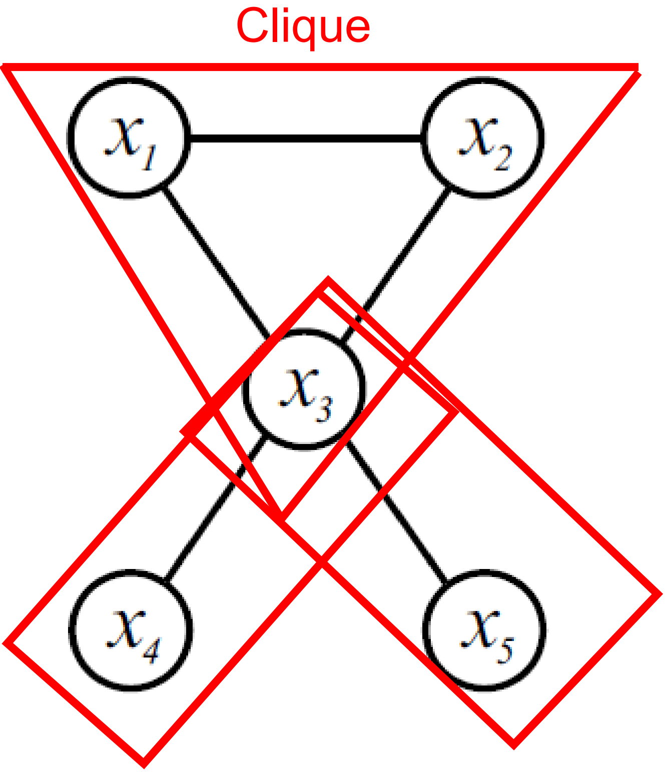


An MRF does not imply a unique factorization, for example all the following are "*valid*":

$$\psi(x_1, x_2, x_3, x_4, x_5)$$

Markov Random Fields (MRFs)

A factor $\psi_c(x_c)$ corresponds to a clique $c \in \mathcal{C}$ (fully connected subgraph) in the MRF



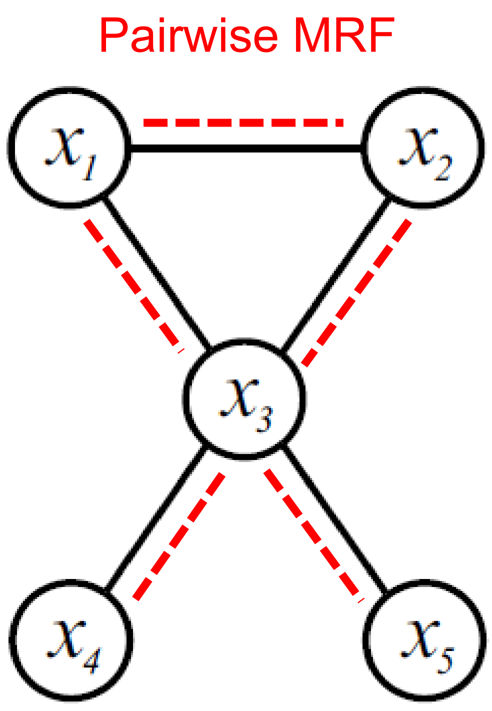
An MRF does not imply a unique factorization, for example all the following are "*valid*":

$$\psi(x_1, x_2, x_3, x_4, x_5)$$

$$\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

Markov Random Fields (MRFs)

A factor $\psi_c(x_c)$ corresponds to a clique $c \in \mathcal{C}$ (fully connected subgraph) in the MRF



An MRF does not imply a unique factorization, for example all the following are "*valid*":

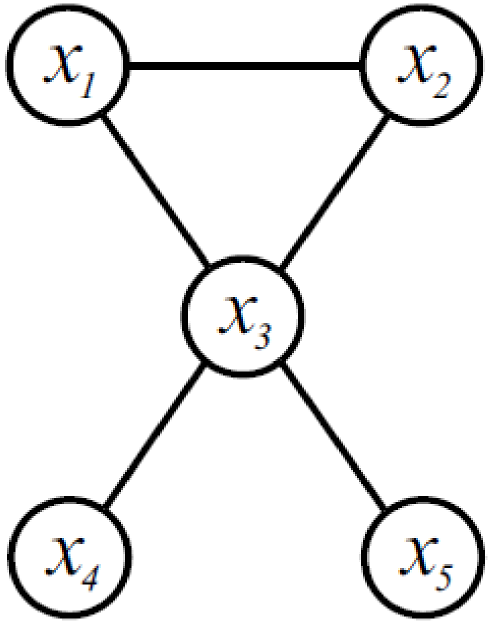
$$\psi(x_1, x_2, x_3, x_4, x_5)$$

$$\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

$$\psi(x_1, x_2)\psi(x_2, x_3)\psi(x_1, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

A **minimal factorization** is one where all factors are **maximal cliques** (not a strict subset of any other clique) in the MRF

Example: Gaussian MRF



Interaction potential between each pair of nodes $(i, j) \in \mathcal{E}$ is exponentiated quadratic,

$$\psi_{ij}(x_i, x_j) = \exp\left(-\frac{1}{2}(x_i - x_j)^2\right)$$

Joint probability is proportional to product,

$$p(x) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{23}(x_2, x_3)\psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

Multivariate Gaussian distribution

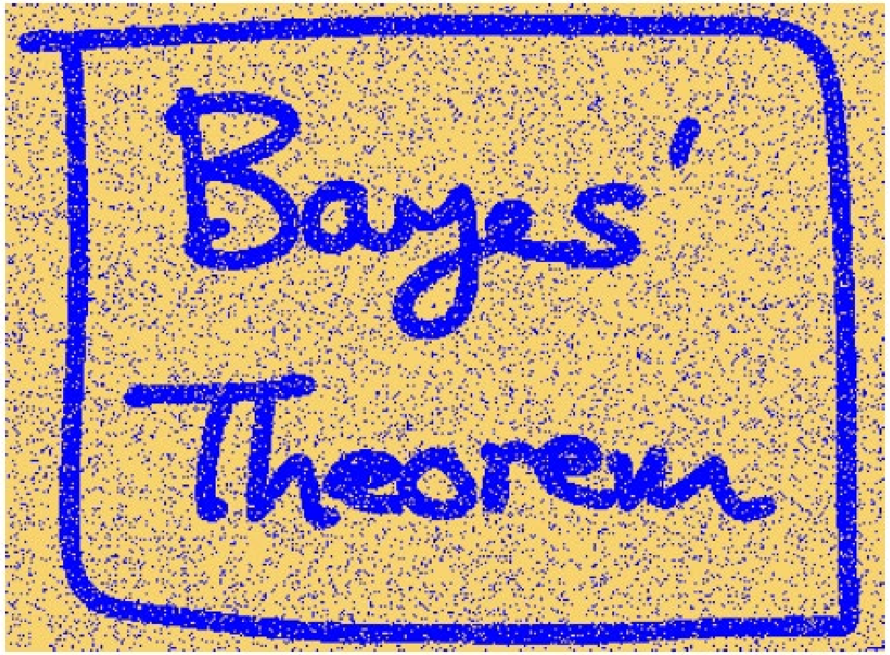
$$p(x) = \mathcal{N}(x \mid \mu, \Sigma) \qquad Z = (2\pi)^{5/2}|\Sigma|^{1/2}$$

Can easily read off inverse covariance…

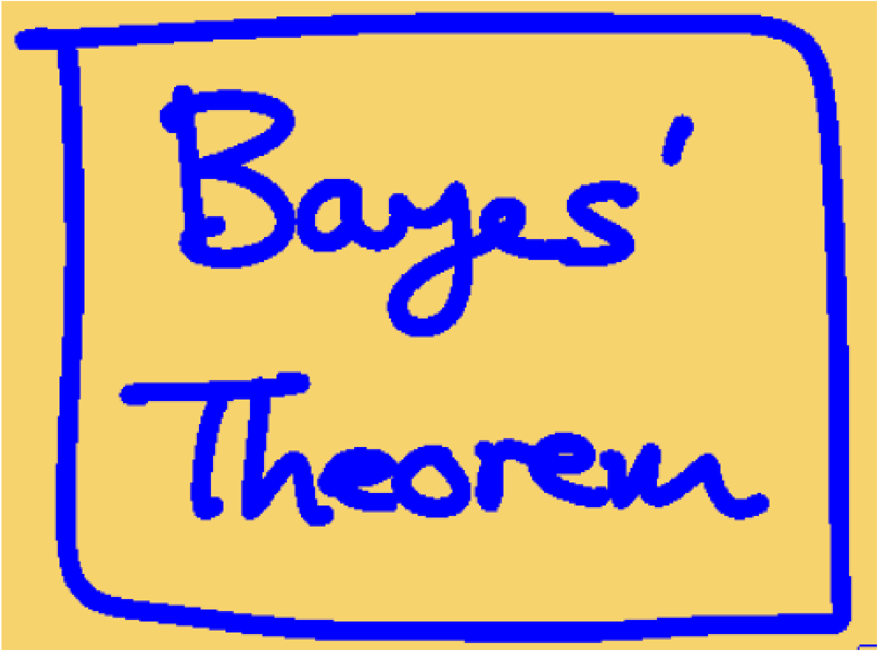
$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 4 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Example: Image Denoising

Noisy Image



Latent Image

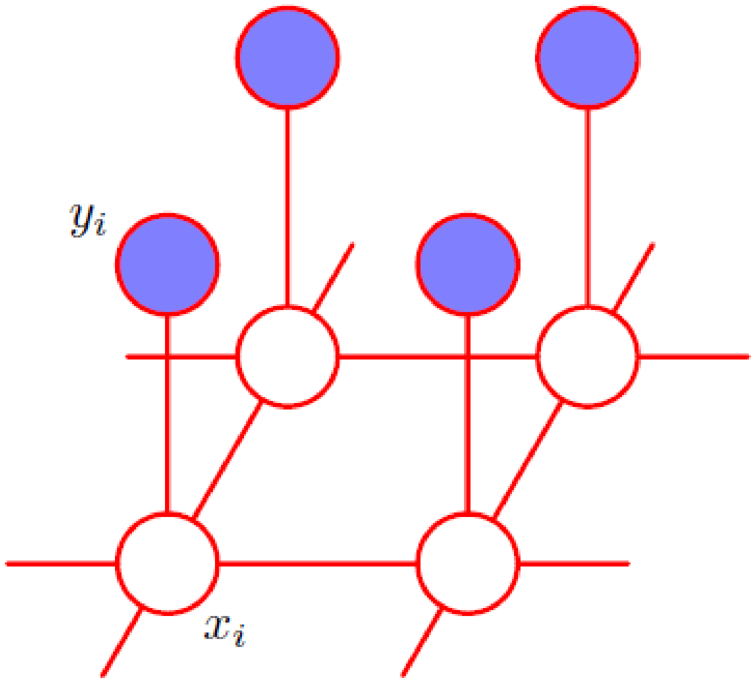


Problem Given observed image corrupted by i.i.d. noise, infer "clean" denoised image.

[Source: Bishop, C. PRML]

Example: Image Denoising

Model Assume binary image with latent pixels $x_i \in \{-1, +1\}$ and observed $y_i \in \{-1, +1\}$



Observed pixels randomly flipped i.i.d.,

$$\log \phi_i(x_i) = \eta x_i y_i$$ Eta parameter controls noise

Neighboring pixels should appear similar,

$$\log \phi_{ij}(x_i, x_j) = \beta x_i x_j$$ Beta parameter controls *smoothness*

Full MRF (in "energy" form):

$$E(x, y) = -\sum_i \log \phi_i(x_i) - \sum_{(i,j)} \log \phi_{ij}(x_i, x_j)$$

Often specify MRF in "energy" or negative log-probability form (minimize energy → maximize probability)

[Source: Bishop, C. PRML]

Normalizing MRFs

Joint probability of *image denoising* model,

$$p(x, y) = \frac{1}{Z} \exp\{-E(x, y)\}$$

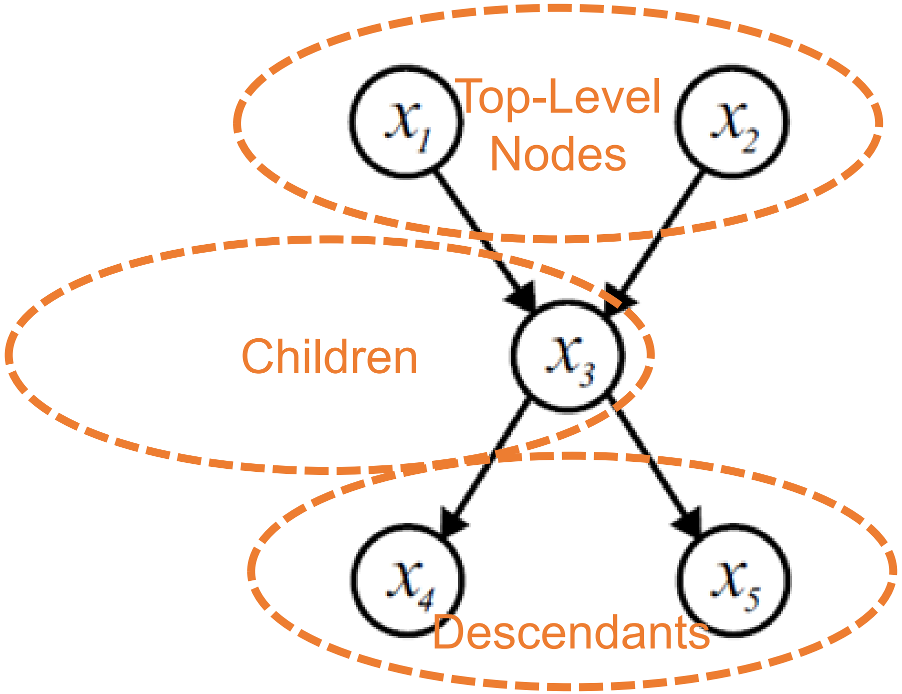
Normalization (a.k.a. partition function) for N pixel image:

$$Z = \underbrace{\sum_{x_1} \sum_{x_2} \cdots \sum_{x_N}}_{\mathbf{O(2^N)\ terms}} \exp\{-E(x, y)\}$$

Normalization not always possible in closed-form : i.e. need to sum over *all possible N-pixel images*

Often ignore Z and specify MRFs up to proportionality…

Simulation



Bayes Nets Straightforward simulation via ancestral sampling successively samples from conditionals:

$$p(\mathbf{x}) = \prod_{i \in \mathcal{V}} p(x_i \mid x_{\mathrm{Pa}(i)})$$

so

$$x_i \sim p(x_i \mid x_{\mathrm{Pa}(i)})$$

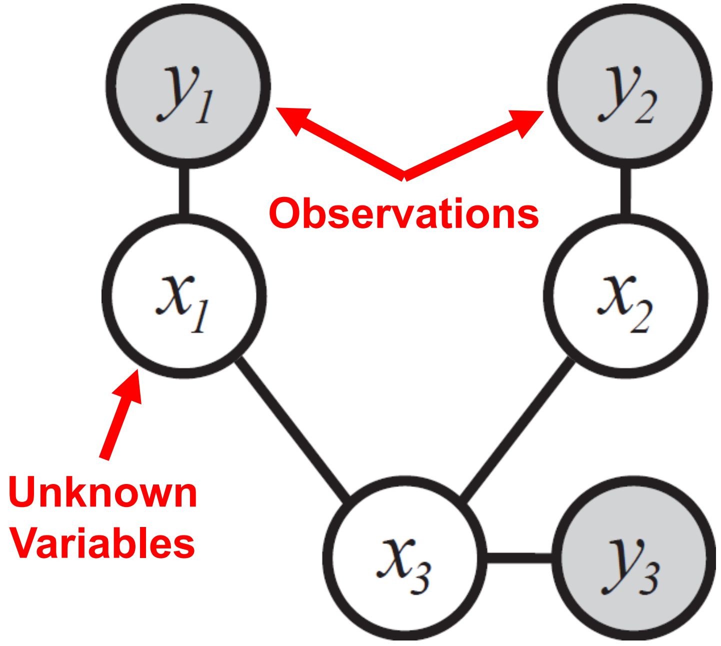
Undirected Graphs Sampling not as straightforward…

- Lack locally normalized conditionals to sample from
- Lack partial ordering of nodes

We will return to this when we discuss Markov chain Monte Carlo

Pairwise Markov Random Field

Often easier to specify and do inference on pairwise model



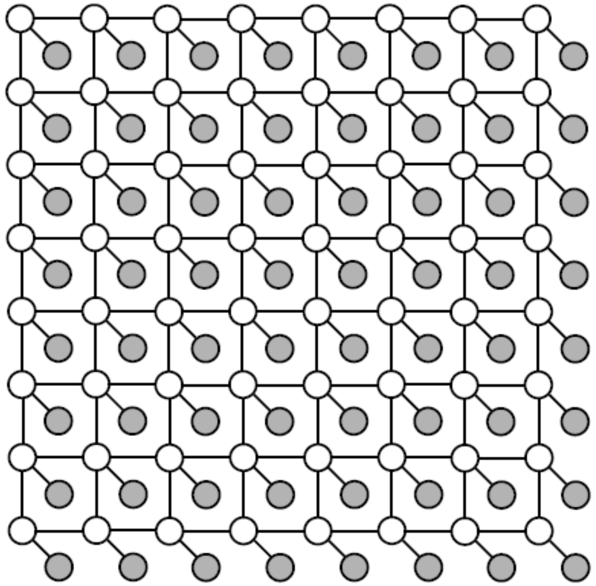
$$p(x, y) \propto \prod_{s \in \mathcal{V}} \psi_s(x_s, y) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t)$$

Likelihood　　Prior

Restricted class of MRFs

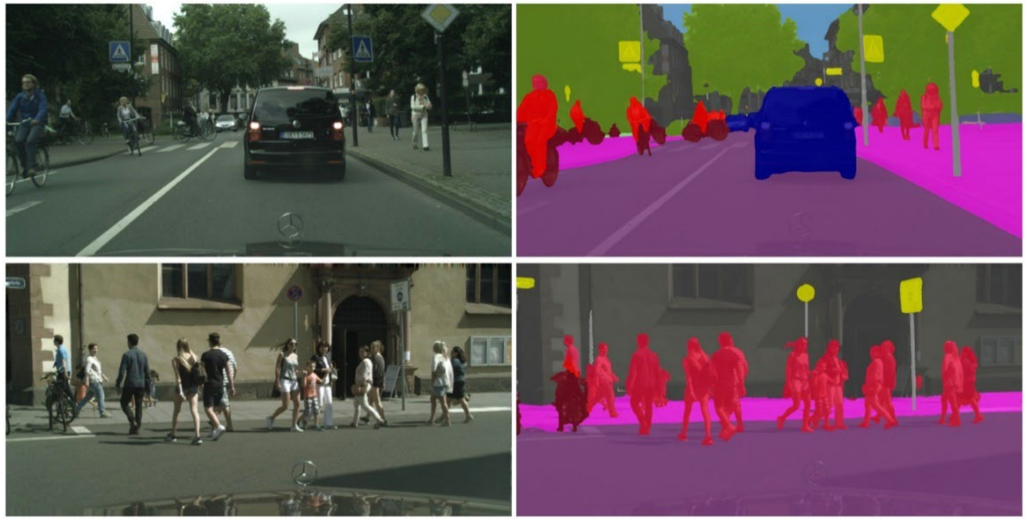
- 2-node factor exists for every edge
- Explicit factorization of joint distribution
- High-order factors not always easily decomposed into pairwise terms

Example: Image Segmentation



Don't need to know log-partition to specify model

[Source: Kundu, A. et al., CVPR16]



Pairwise MRF *energy*: $-\log p(x,y) = \log Z + \sum_i \psi_i(x_i, y_i) + \sum_{(i,j)} \psi_{i,j}(x_i, x_j)$

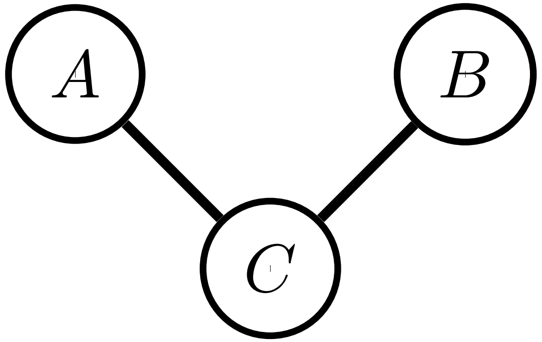
Don't need to specify normalized conditionals as in Bayes Nets

Low energy configurations = High probability

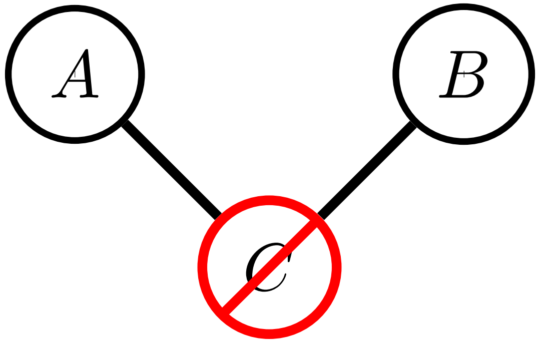
L2 Likelihood: $\psi_i(x_i, y_i) = \|x_i - y_i\|^2$ **Potts model:** $\psi_{i,j}(x_i, x_j) = \mathbb{I}(x_i = x_j)$

MAP (minimum energy) configuration = Piecewise constant regions

Transformations of Undirected Models



$$p(A,B,C) = \psi_{AC}(A,C)\psi_{BC}(B,C)/Z$$

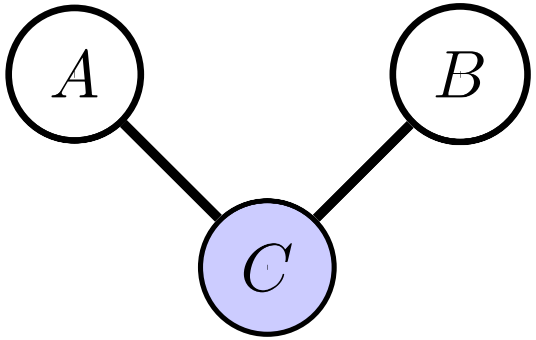


$\rightarrow$

$$p(A,B) \neq p(A)p(B)$$

Marginalization: Join all nodes that have path through C

Marginalising over C makes A and B (graphically) dependent.



$\rightarrow$

$$p(A,B|C) = p(A|C)p(B|C)$$

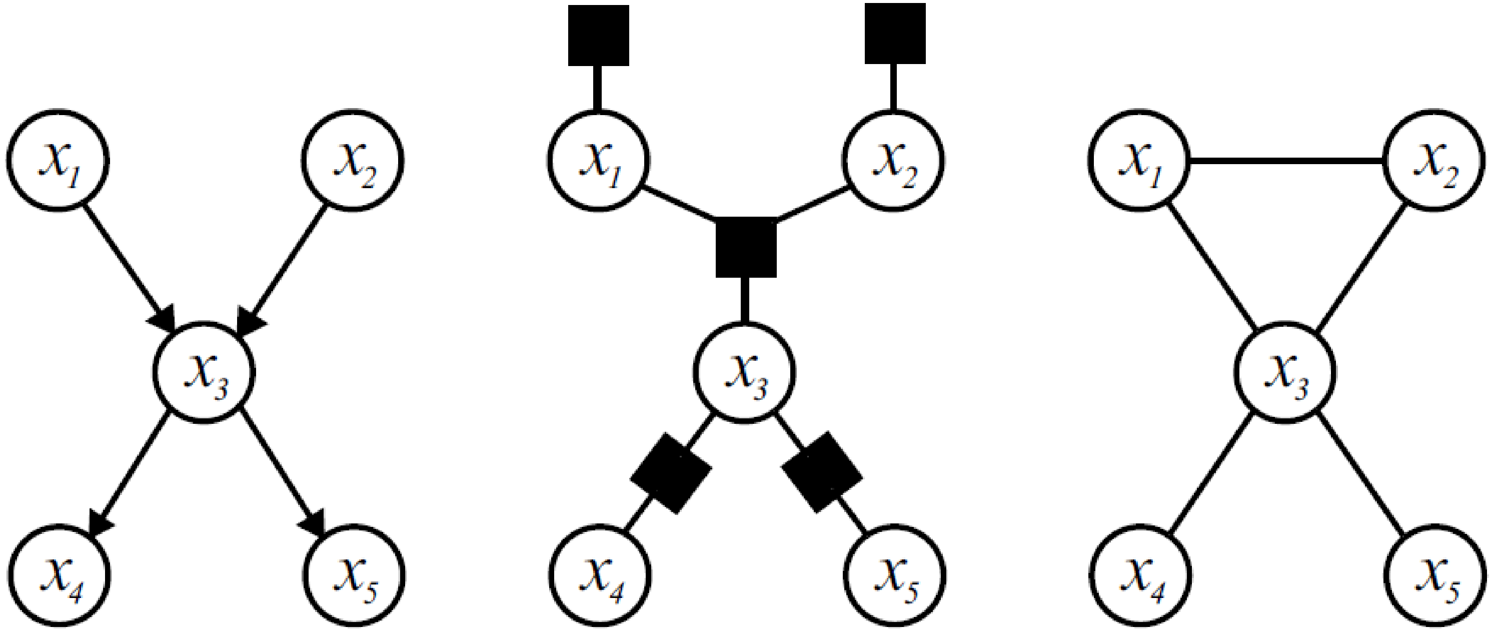
Conditioning: Drop all edges on path through C

Conditioning on C makes A and B independent:

[Source: Erik Sudderth]

Graphical Models

A variety of graphical models can represent the same probability distribution



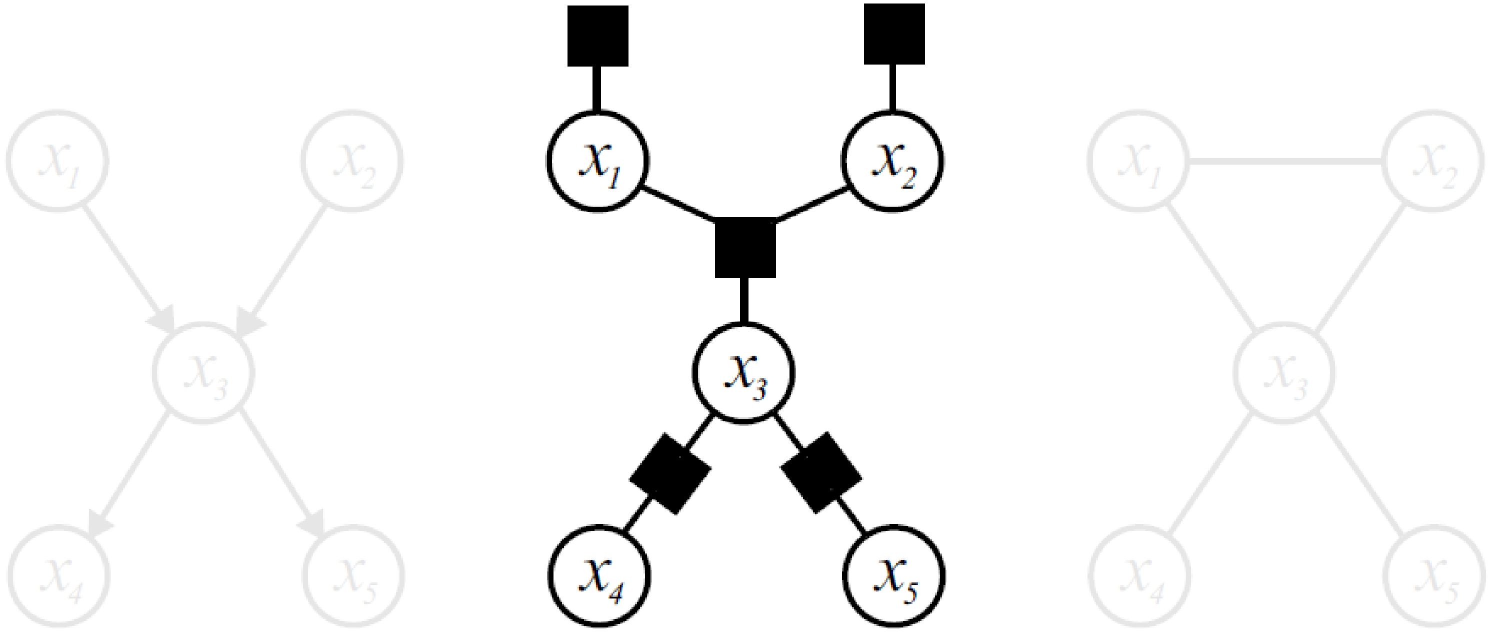
Bayes Network **Factor Graph** **Markov Random Field**

Directed Models **Undirected Models**

[Source: Erik Sudderth, PhD Thesis]

Graphical Models

A variety of graphical models can represent the same probability distribution



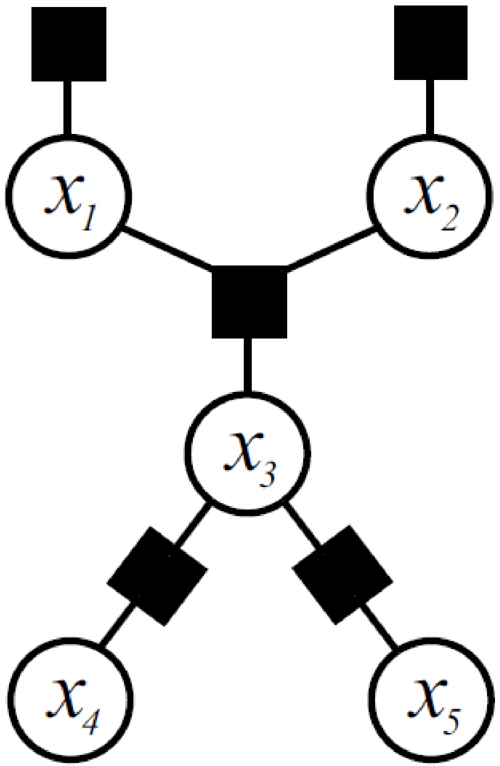
Bayes Network **Factor Graph** Markov Random Field

Directed Models **Undirected Models**

[Source: Erik Sudderth, PhD Thesis]

Factor Graphs

A *hypergraph* $\mathcal{H} = (\mathcal{V}, \mathcal{F})$ where a *hyperedge* $f \in \mathcal{F}$ is a subset of vertices $f \subset \mathcal{V}$.



Factor node for each product term in the joint factorization:

Graphical model makes factorization explicit

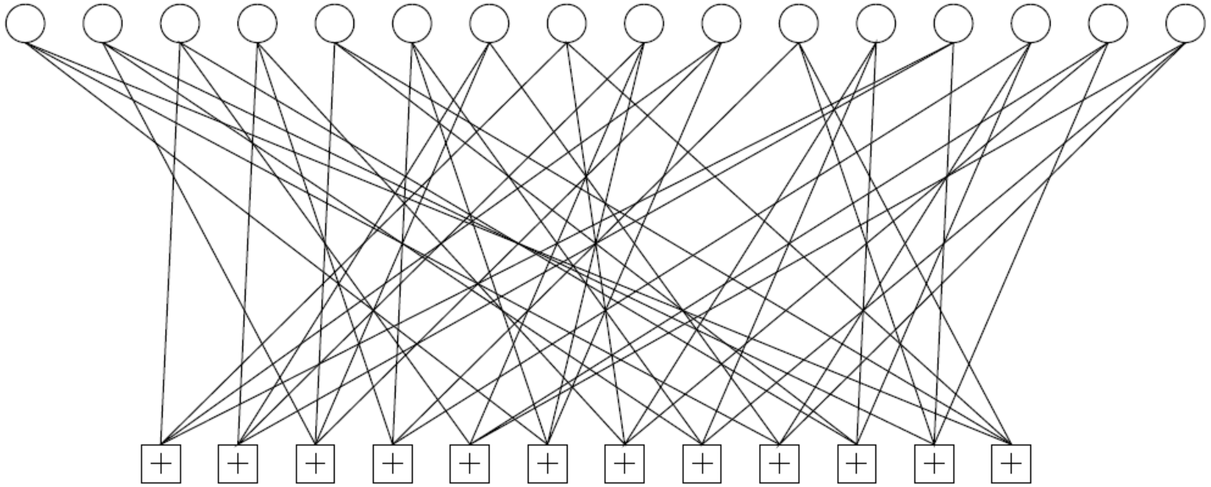
$$p(x) \propto \prod_{f \in \mathcal{F}} \psi_f(x_f)$$

where $x_f = \{x_i : i \in f\}$ the set of variables in factor *f*. For example:

$$\psi(x_1)\psi(x_2)\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

Example: Low Density Parity Check Codes

Factor Graph Representation



Problem Setup

- A code t is transmitted over a noisy
- Received code r is corrupted by noise
- Estimate the most probable code that was sent $t*$ (*maximum a posteriori)*

Transmitted Code

$t \sim p(t)$

Noisy Channel

Received Code

$r \mid t \sim p(r \mid t)$

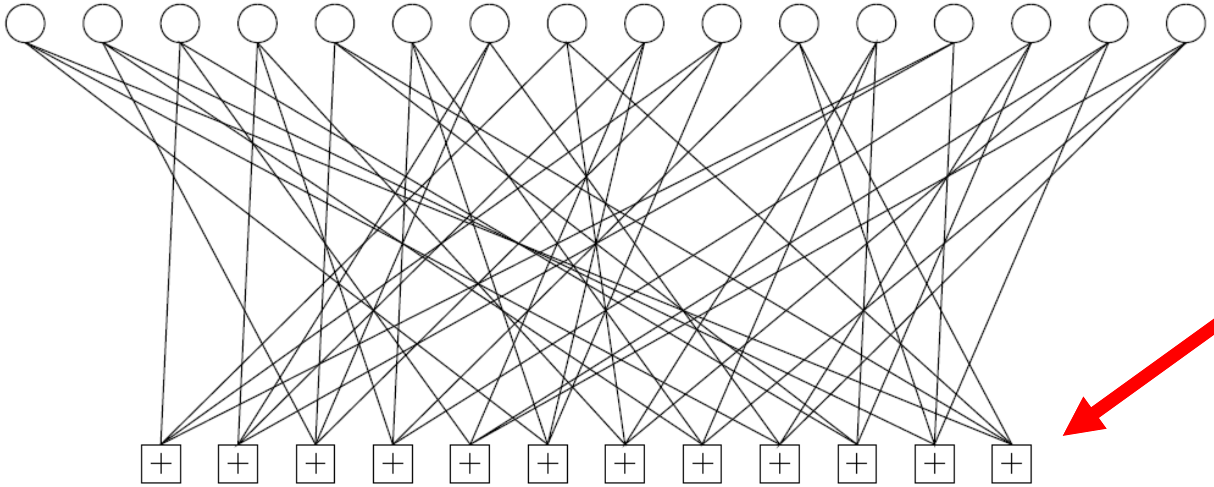
Decoder

$t^* = \arg\max_t p(t \mid r)$

[Source: David MacKay]

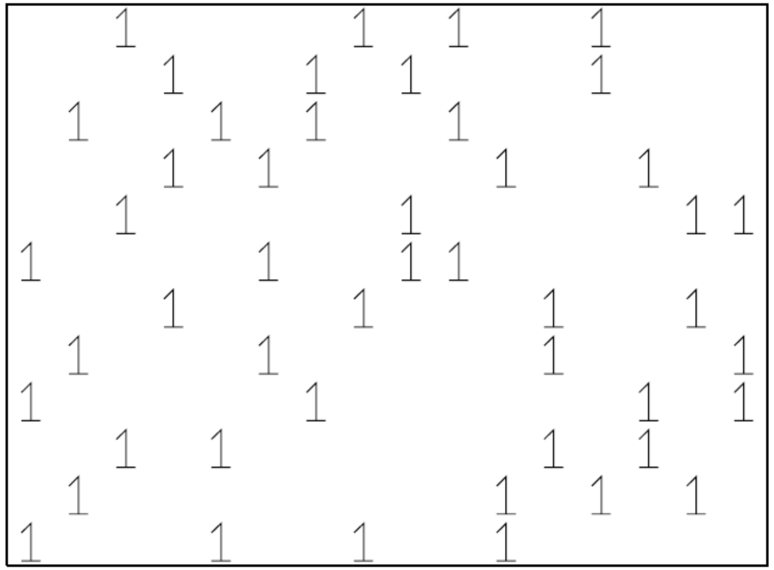
Example: Low Density Parity Check Codes

Factor Graph Representation



Sparse Parity Check Matrix

$\mathbf{H} =$



- Valid codes have zero parity: $p(t) \propto \mathbb{I}(Ht = 0 \bmod 2)$
- Chanel noise model arbitrary, e.g. flip bits w/ ϵ probability:

$$p(r \mid t) = \prod_n p(r_n \mid t_n) = \prod_n (1-\epsilon)^{\mathbb{I}(r_n = t_n)} \epsilon^{\mathbb{I}(r_n \neq t_n)}$$

n-th bit

[Source: David MacKay]

Recap: Directed Models

- Distribution factorized as product of conditionals via chain rule

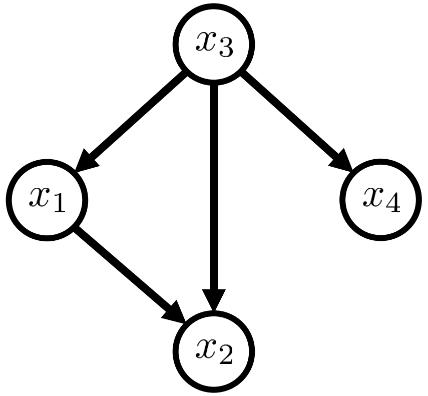
$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4)$$

- Choose ordering where terms simplify due to conditional independence

Eg. Suppose $x_4 \perp x_1 \mid x_3$ and $x_2 \perp x_4 \mid x_1$ then:

$$p(x) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3)$$

- Directed graph encodes factorized distribution via conditional independence properties



- Straightforward simulation via **ancestral sampling**
- **Factorization is unique** for a Bayes net

Recap: Undirected Model

- Joint factorization as nonnegative factors (potentials) over subsets:

$$p(x) \propto \prod_{f \in \mathcal{F}} \psi_f(x_f)$$

- Easier to specify models compared to Bayes nets since:
  - Factors do not need to be normalized conditional probabilities
  - May specify up to unknown normalization constant

- **Factorization ambiguous** in MRFs, but **explicit in factor graphs** (factor graphs generally preferred)

- Simulation is not easy in general. Can't do ancestral sampling because no partial ordering.