



Computer
Science

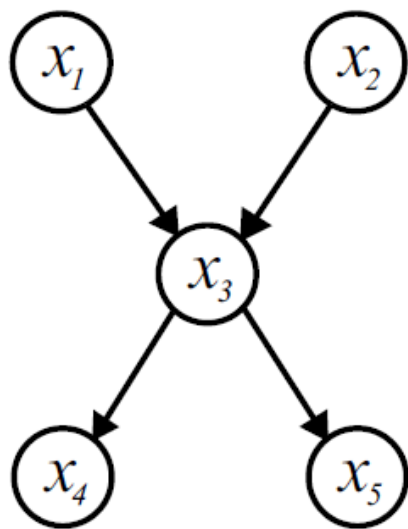
CSC535: Probabilistic Graphical Models

Probabilistic Graphical Models

Prof. Jason Pacheco

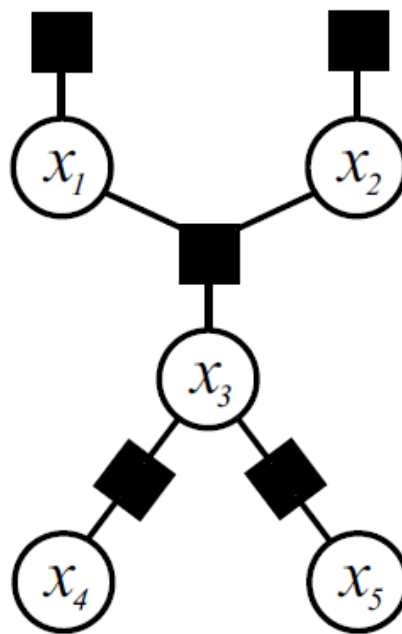
Graphical Models

A variety of graphical models can represent the same probability distribution

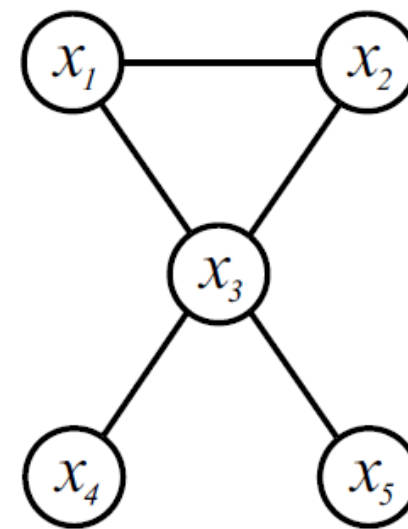


Bayes Network

Directed Models



Factor Graph

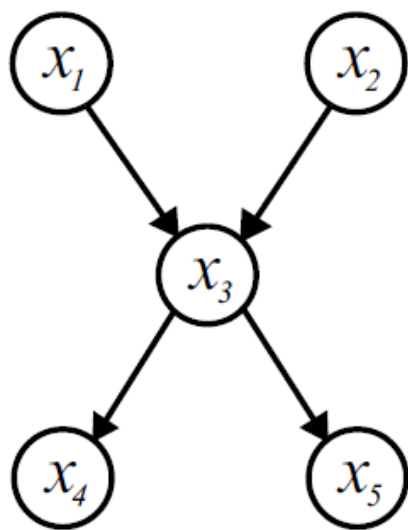


Markov Random Field

Undirected Models

Graphical Models

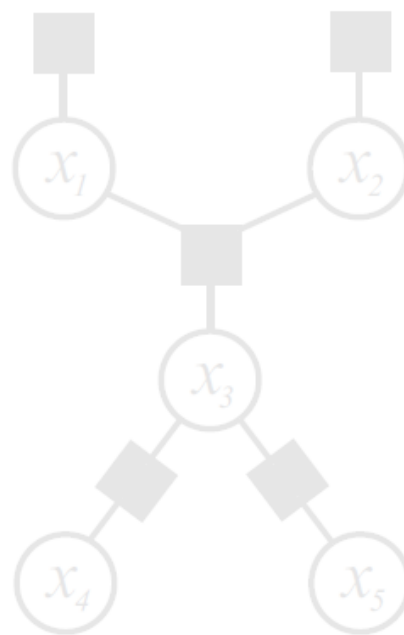
A variety of graphical models can represent the same probability distribution



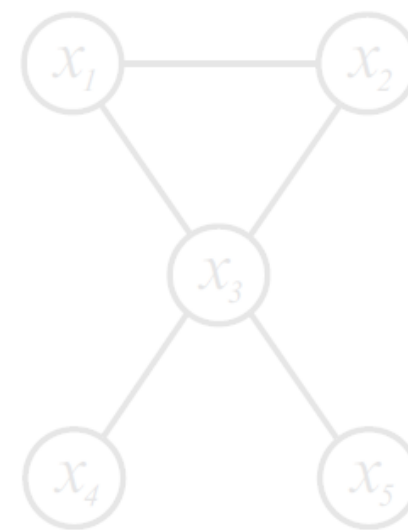
Bayes Network



Directed Models



Factor Graph



Markov Random Field

Undirected Models

Outline

Directed graphical models

- Bayes Nets
- Conditional dependence

Undirected graphical models

- Markov random fields (MRFs)
- Factor graphs

Outline

Directed graphical models

- Bayes Nets
- Conditional dependence

Undirected graphical models

- Markov random fields (MRFs)
- Factor graphs

From Probabilities to Pictures

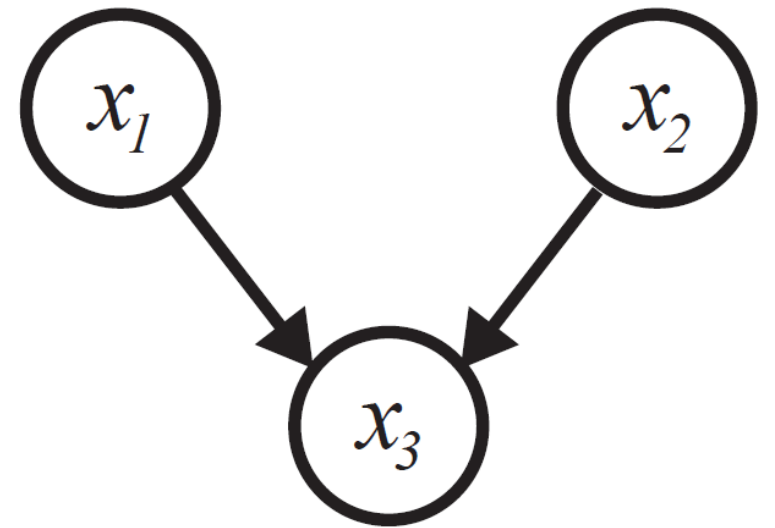
A probabilistic graphical model allows us to pictorially represent a probability distribution

Probability Model:

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



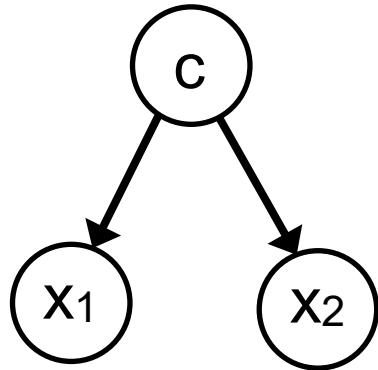
Graphical Model:



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

Directed Graphical Models

Directed models are generative models...



$$p(C, X_1, X_2) = p(C)p(X_1 | C)p(X_2 | C)$$

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

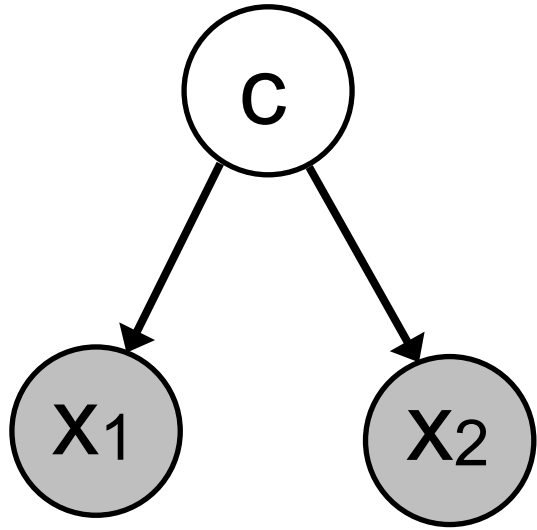
...tells how data are generated (called ***ancestral sampling***)

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

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

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

Inference



Denote observed data with shaded nodes,

$$X_1 = x_1 \quad X_2 = x_2$$

Infer *latent* variable C via Bayes' rule:

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

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

Probability Chain Rule

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

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 | x_1)p(x_3 | x_1, x_2)p(x_4 | x_1, x_2, x_3)$$

Valid for *any ordering* of the random variables...

$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 | x_3)p(x_4 | x_1, x_3)p(x_2 | x_1, x_3, x_4)$$

For a collection of N RVs and any permutation ρ :

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

Conditional Independence

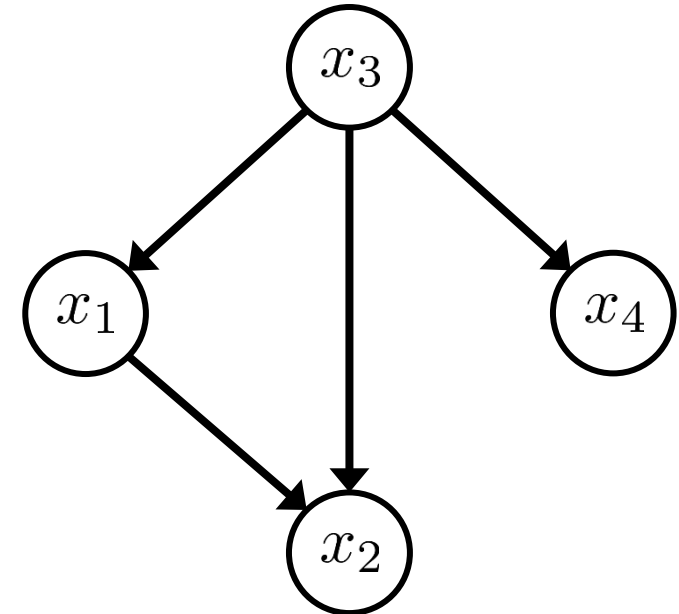
Recall two RVs X and Y are **conditionally independent** given Z (or $X \perp Y \mid Z$) iff:

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

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:

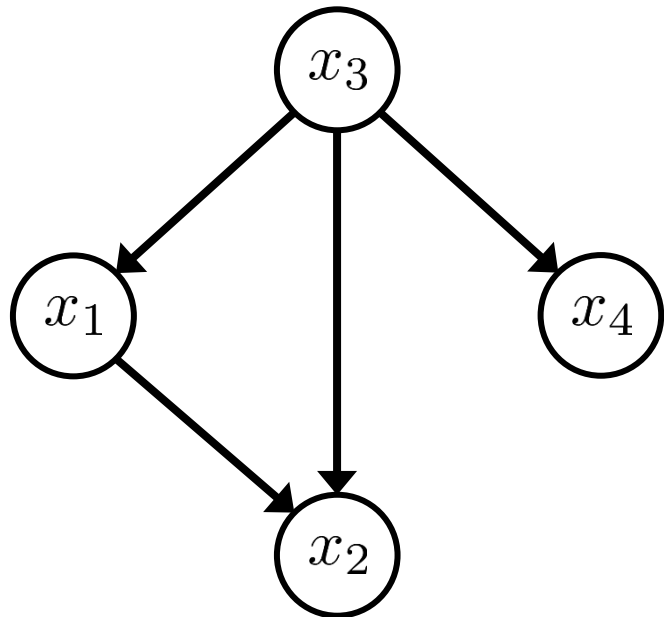
$$\begin{aligned} p(x) &= p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_1, x_3)p(x_2 \mid x_1, x_3, x_4) \\ &= p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3) \end{aligned}$$



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 ,

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

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

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

Ancestors are parents-of-parents.

Descendants are children-of-children.

Directed PGM = Bayes Network

Model factors are normalized conditional distributions:

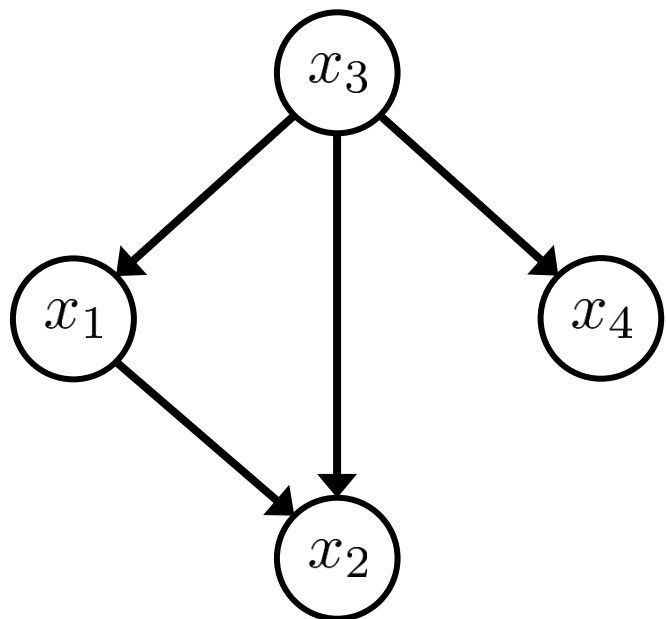
$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\text{Pa}(s)})$$

 Parents of node s

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

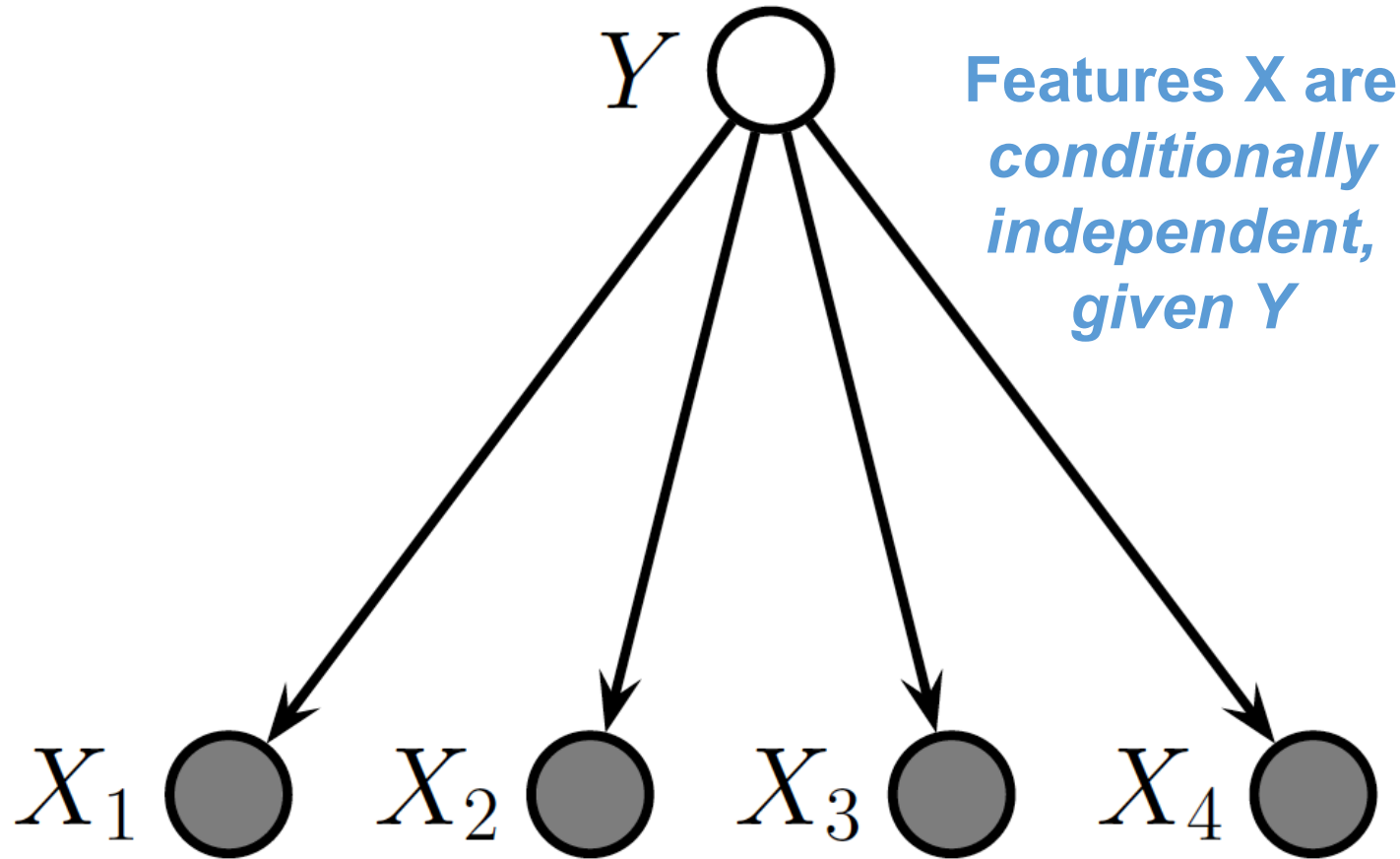
$$p(x) = p(x_3)p(x_1 \mid x_3)p(x_4 \mid x_3)p(x_2 \mid x_1, x_3)$$

Locally normalized factors yield globally normalized joint probability

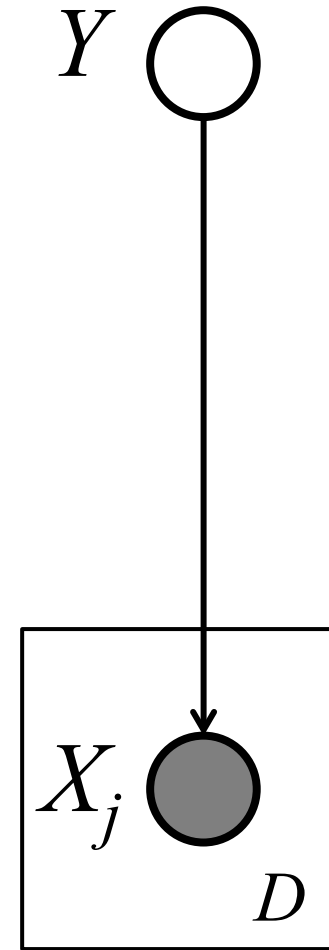


Shading & Plate Notation

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**

Example: Gaussian Mixture Model

Bayes nets are easily simulated via ancestral sampling...

Probability Model

$$\pi \sim \text{Dirichlet}(\cdot)$$

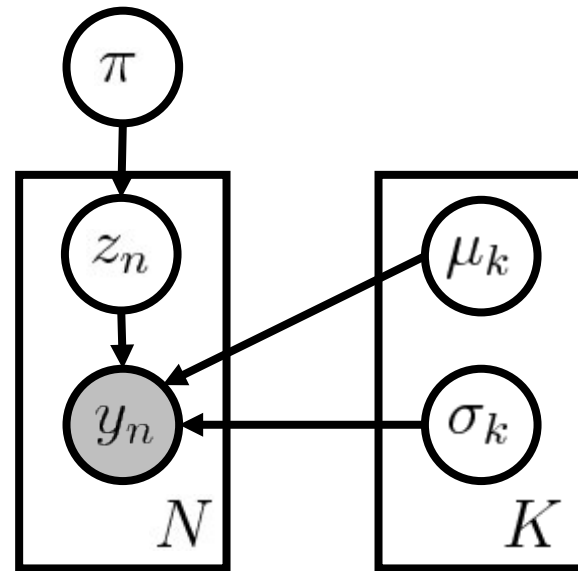
$$\mu_k \sim \mathcal{N}(\cdot)$$

$$\sigma_k \sim \text{Inv-Gamma}(\cdot)$$

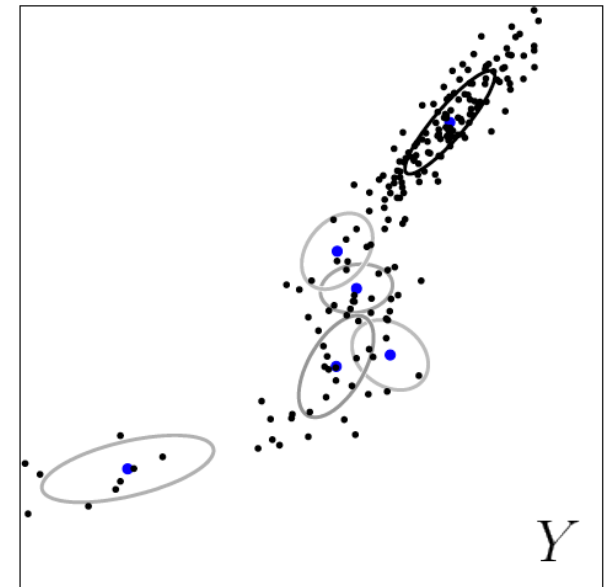
$$z_n \mid \pi \sim \text{Cat}(\pi)$$

$$y_n \mid z_n, \mu_{z_n}, \sigma_{z_n} \sim \mathcal{N}(\mu_{z_n}, \sigma_{z_n})$$

Bayes Net

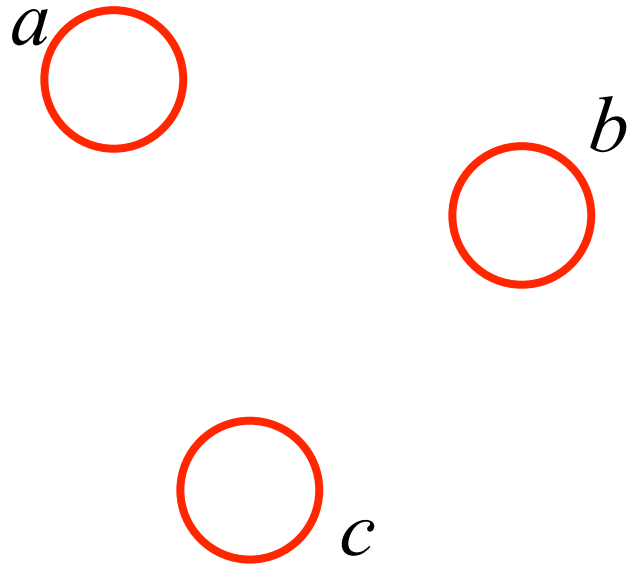


Joint Sample

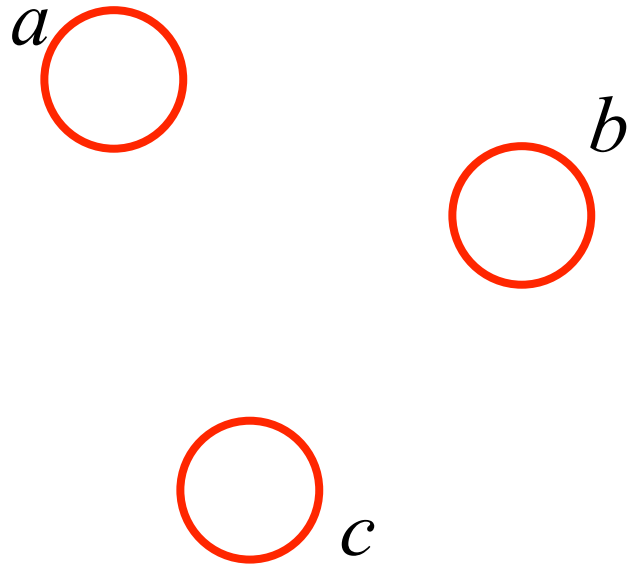


Sample all nodes with no parents, then children, etc., to terminals. Can sample nodes at same level in parallel.

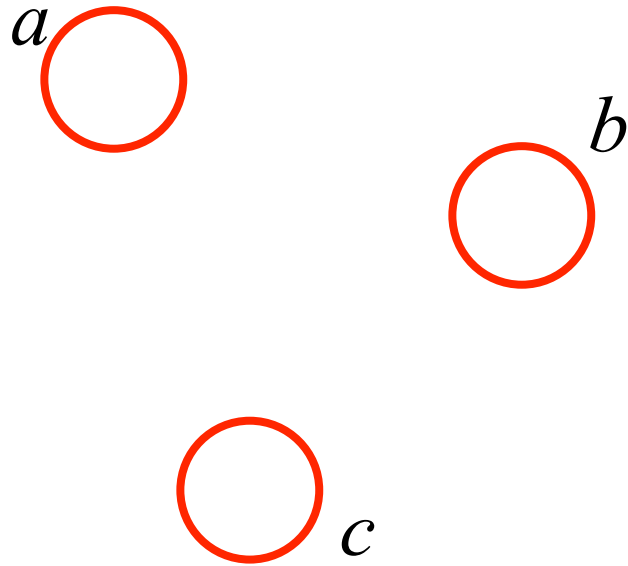
What is the joint factorization?



$$p(\mathbf{a}, \mathbf{b}, \mathbf{c}) = p(\mathbf{a})p(\mathbf{b})p(\mathbf{c})$$

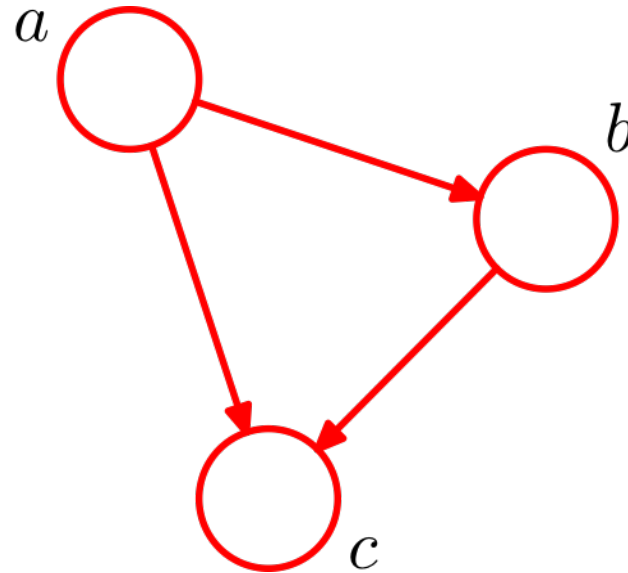


Are a and b independent ($a \perp b$)?



$$\mathbf{p(a,b,c) = p(a)p(b)p(c)}$$

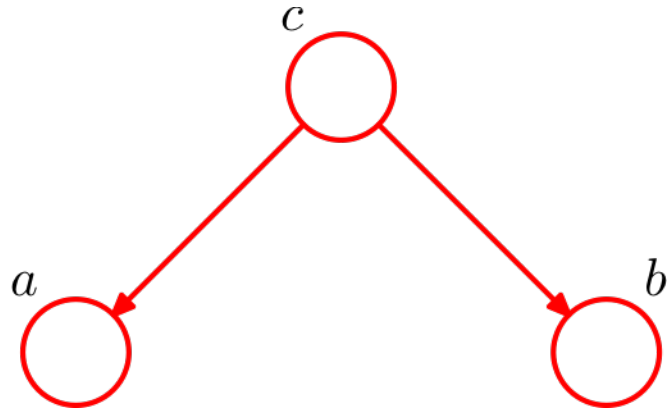
$$p(a,b,c) = p(a)p(b|a)p(c|a,b)$$



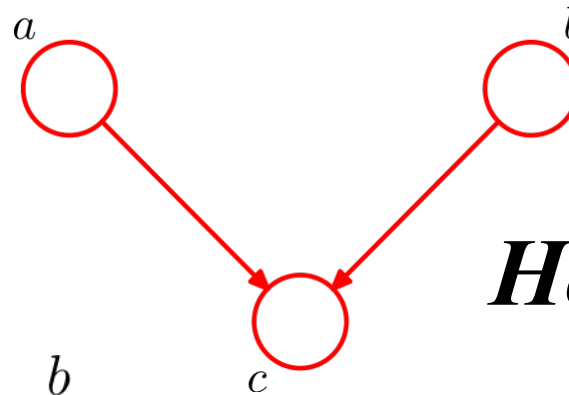
Note there are **no conditional independencies**

Three interesting cases

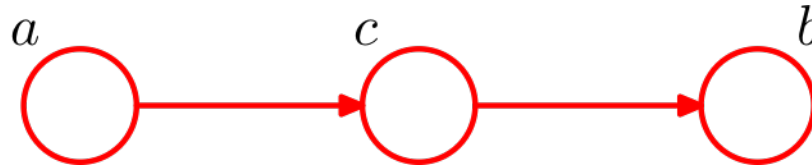
Tail-to-tail



Head-to-head

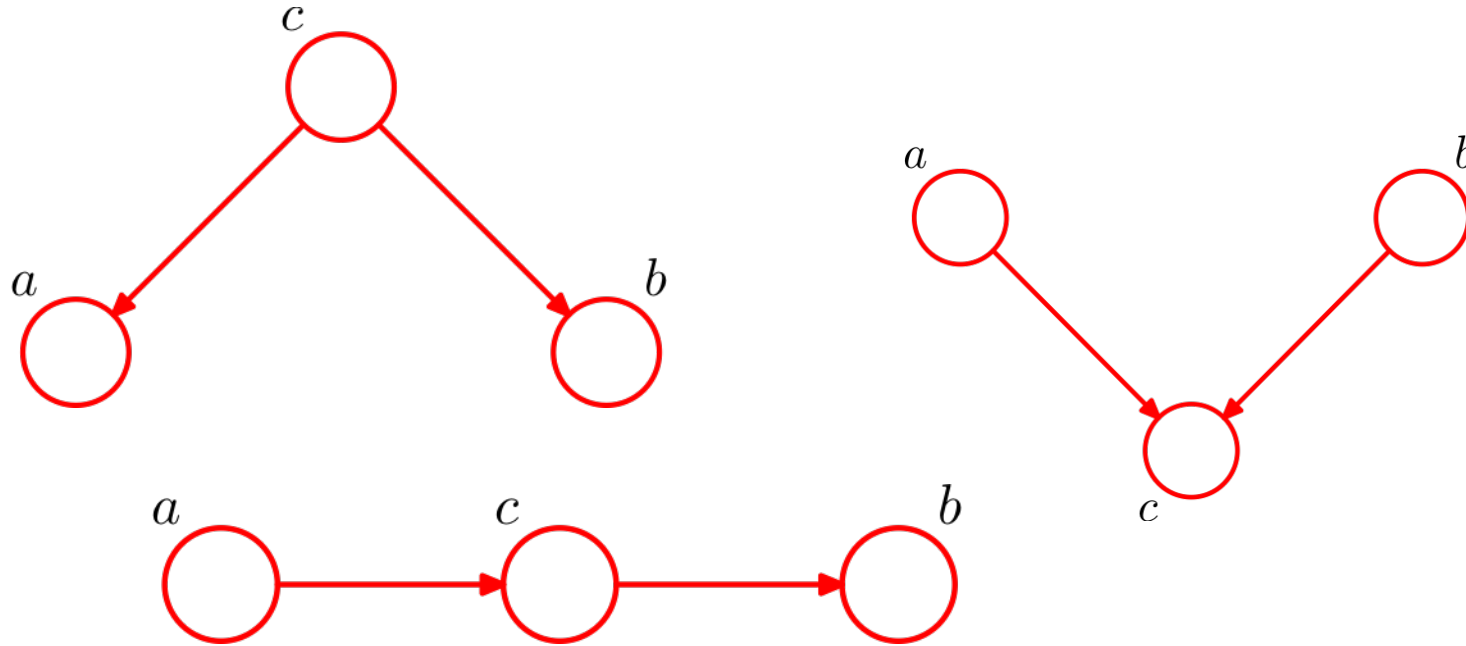


Head-to-tail



D-Separation: Use these cases to determine dependence / independence properties

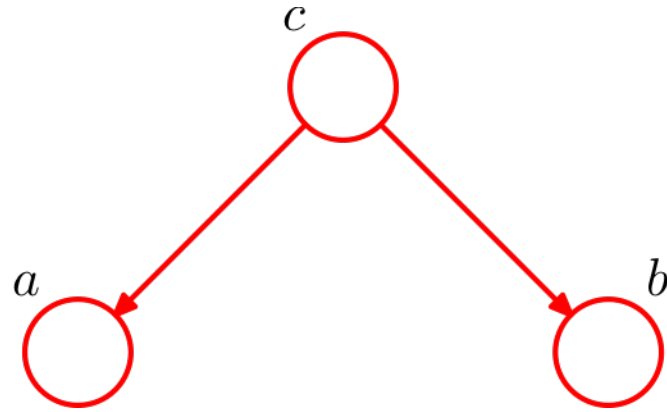
Three interesting cases



For each case, consider two questions:

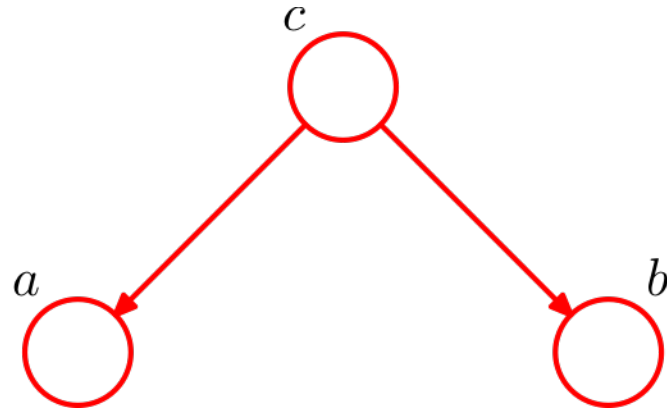
- 1) Is $a \perp b$?
- 2) Is $a \perp b \mid c$? (i.e. *c* is observed)

Case one (tail-to-tail)



Is $a \perp b$?

Case one (tail-to-tail)

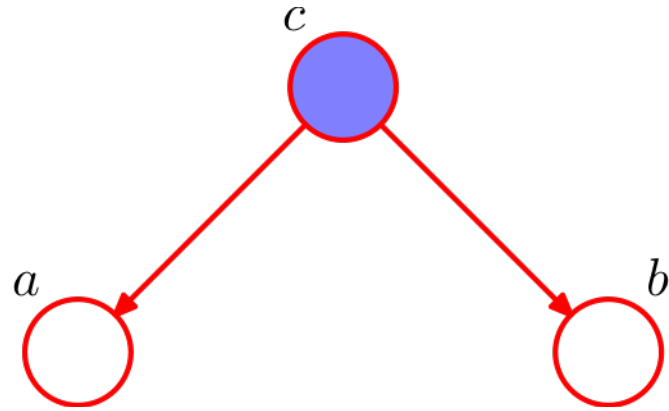


$$a \not\perp b$$

Intuition c generates, both, a and b . Knowing a tells you something about c (via Bayes rule $p(c|a)$) which in turn generates b ...information is exchanged

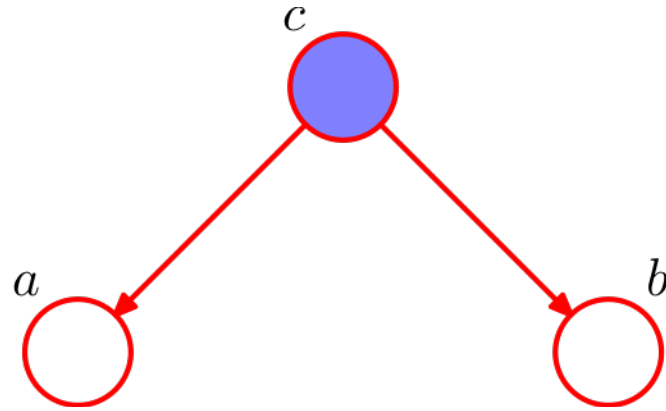
Can prove by counterexample (HW problem)

Case one where c is observed



Is $a \perp b \mid c$?

Case one where c is observed



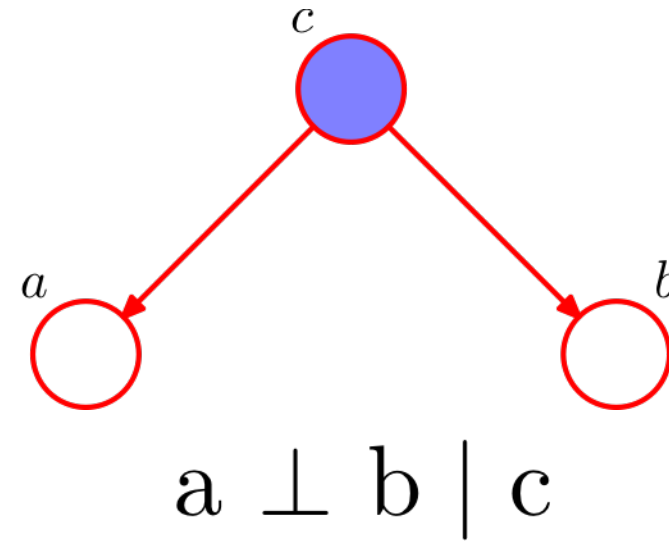
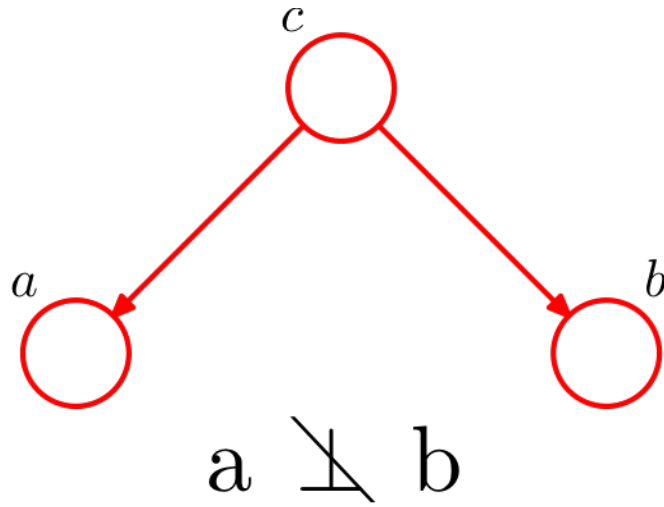
$$a \perp b \mid c$$

$$p(a, b, c) = p(c)p(a|c)p(b|c) \quad (\text{what the graph represents in general})$$

$$p(a, b|c) = p(a|c)p(b|c) \quad (\text{with } c \text{ observed})$$

This is the definition of $a \perp b \mid c$

Case one (tail-to-tail) summary

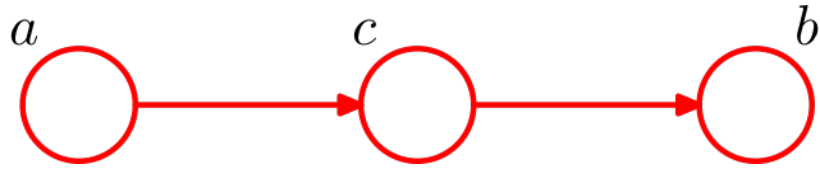


Tail-to-tail case

With no conditioning = no independence

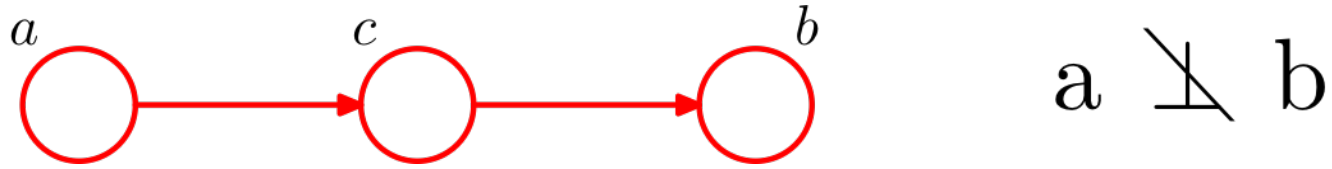
With conditioning = independence

Case two (head-to-tail)



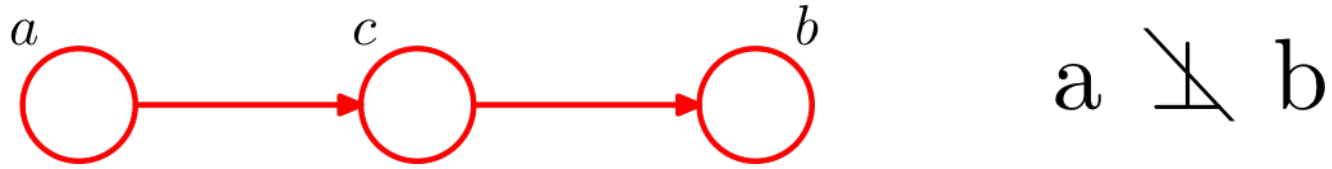
Is $a \perp b$?

Case two (head-to-tail)



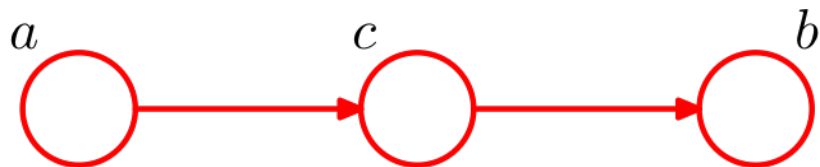
If you know a , that informs you about c , which informs you about b .

Case two (head-to-tail)



The graph represents $p(a, b, c) = p(a)p(c|a)p(b|c)$

Case two (head-to-tail)



$a \not\perp b$

The graph represents $p(a, b, c) = p(a)p(c|a)p(b|c)$

Algebraically,

$$p(a, b) = \sum_c p(a, b, c) = p(a) \sum_c p(c|a) p(b|c)$$

If $a \perp b$ then the above would also have to be equal to $p(a)p(b)$

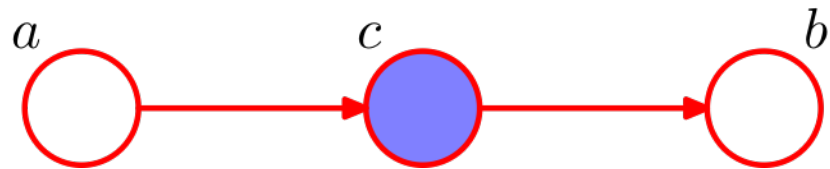
$$p(a, b) = \sum_c p(a, b, c) = p(a) \sum_c p(c|a) p(b|c)$$

If $a \perp b$ then the above **also** equals $p(a)p(b)$

To prove the claim that $a \not\perp b$ we can construct a counter example where the above is false.

Homework Question

Case two where c is observed

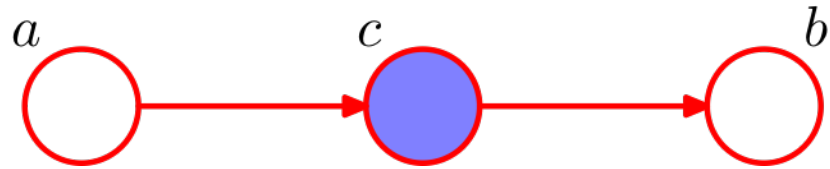


$$a \perp b \mid c$$

$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)}$$

(why?)

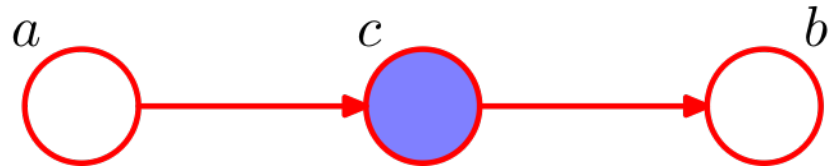
Case two where c is observed



$a \perp b \mid c$

$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} \quad \text{(definition)}$$
$$= \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad \text{(why?)}$$

Case two where c is observed



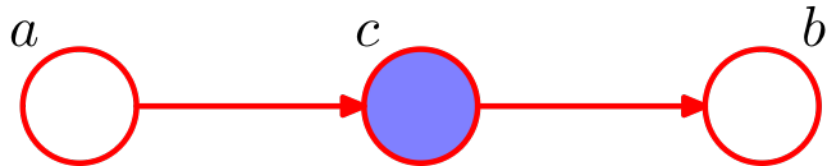
$$a \perp b \mid c$$

$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} \quad \text{(definition)}$$

$$= \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad \text{(from graph)}$$

$$= \frac{p(a)p(a|c)p(c)p(b|c)}{p(a)p(c)} \quad \text{(why?)}$$

Case two where c is observed



$$a \perp b \mid c$$

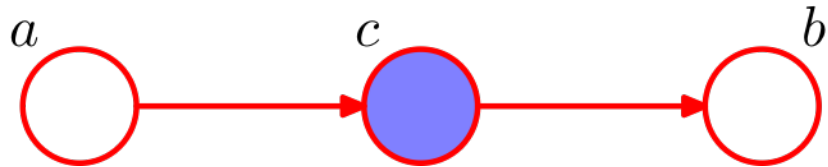
$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} \quad \text{(definition)}$$

$$= \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad \text{(from graph)}$$

$$= \frac{p(a)p(a|c)p(c)p(b|c)}{p(a)p(c)} \quad \text{(Bayes on } p(c|a))$$

$$= p(a|c)p(b|c) \quad \text{(why?)}$$

Case two where c is observed



$$a \perp b \mid c$$



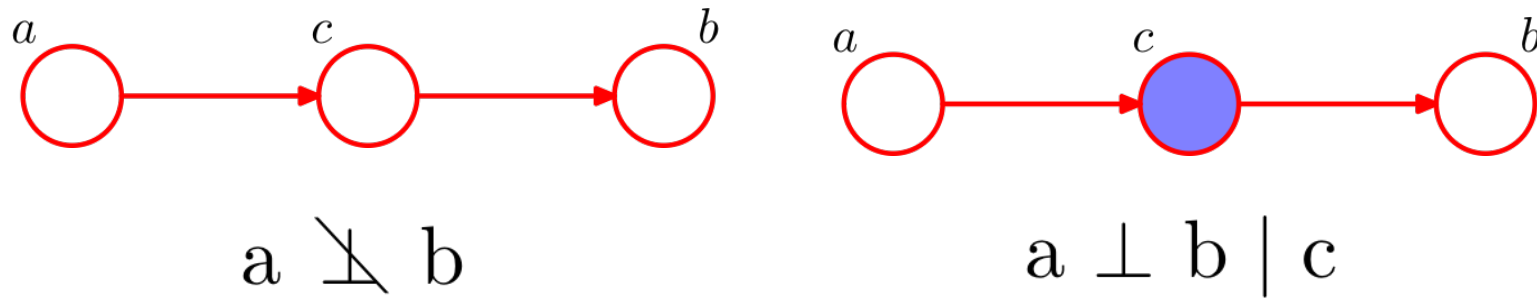
$$p(a, b \mid c) = \frac{p(a, b, c)}{p(c)} \quad \text{(definition)}$$

$$= \frac{p(a)p(c|a)p(b|c)}{p(c)} \quad \text{(from graph)}$$

$$= \frac{p(a)p(a|c)p(c)p(b|c)}{p(a)p(c)} \quad \text{(Bayes on } p(c|a))$$

$$= p(a|c)p(b|c) \quad \text{(canceling factors)}$$

Case two (head-to-tail) summary



Head-to-tail case

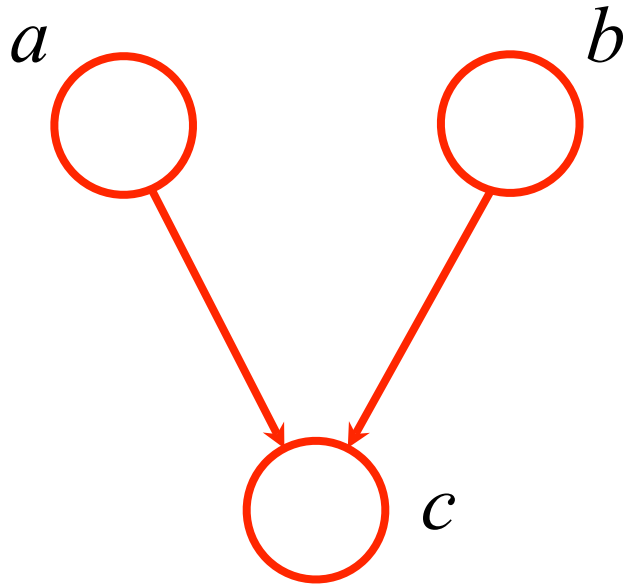
With no conditioning = no independence

With conditioning = independence

(Same as tail-to-tail case!)

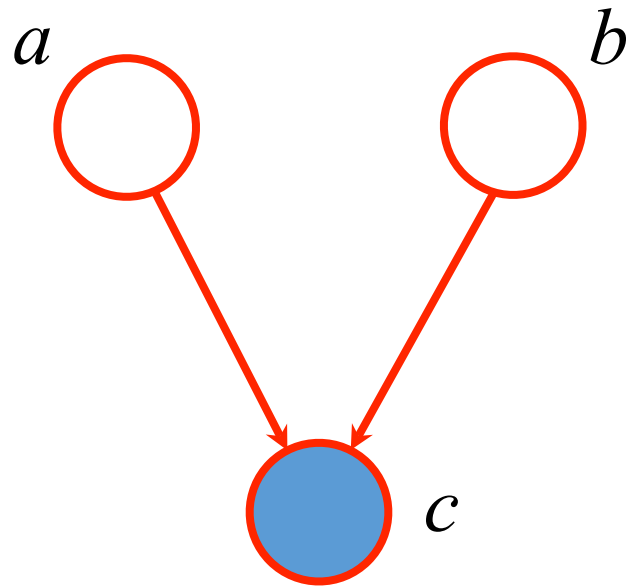
Case three (head-to-head)

Are a and b independent ($a \perp b$)?



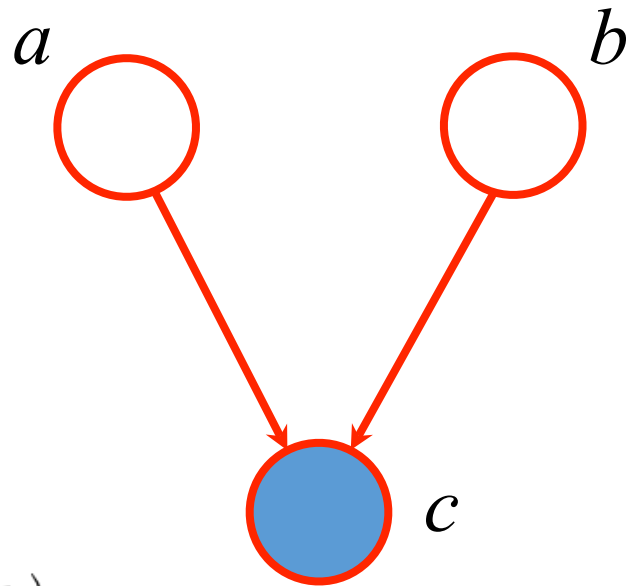
$$p(a, b) = \sum_c p(a)p(b)p(c | a, b) = p(a)p(b)$$

Are a and b conditionally independent ($a \perp b \mid c$)?



$$\mathbf{p(a,b,c) = p(a)p(b)p(c|a,b)}$$

Are a and b conditionally independent ($a \perp b \mid c$)?

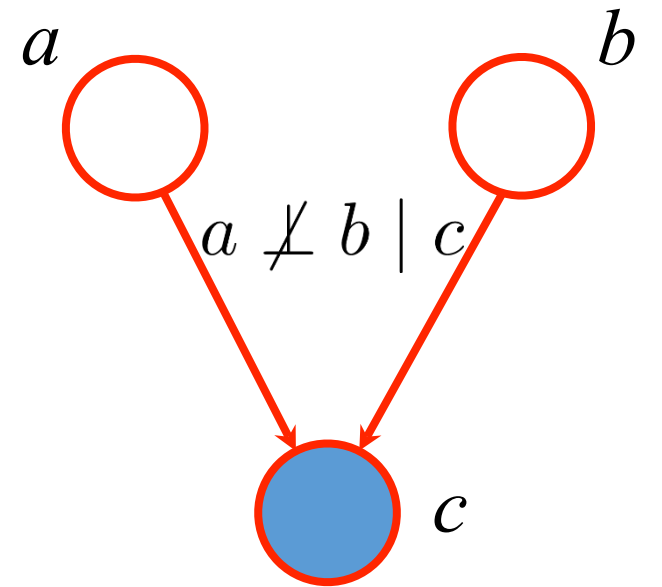
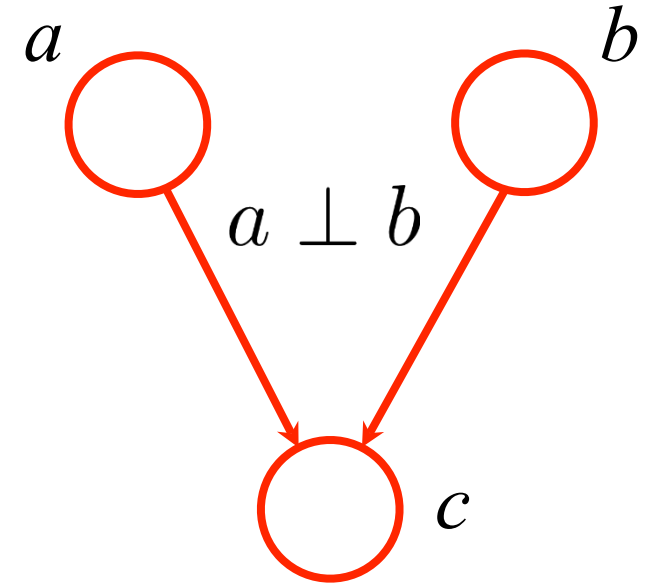
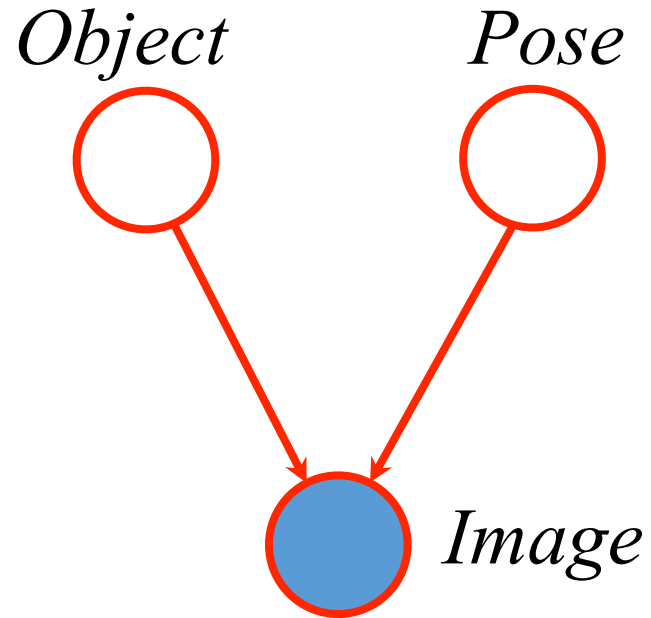


$$\begin{aligned} p(a, b \mid c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c \mid a, b)}{p(c)} \\ &\neq p(a \mid c)p(b \mid c) \quad (\text{in general}) \end{aligned}$$

Attempt at algebraic proof.

Unless the algebra reduces to something obviously false, we typically look for a counter example

Both latent variables must explain same observed data so become dependent



Phenomenon in Bayes networks known as **explaining away**

Markov Properties

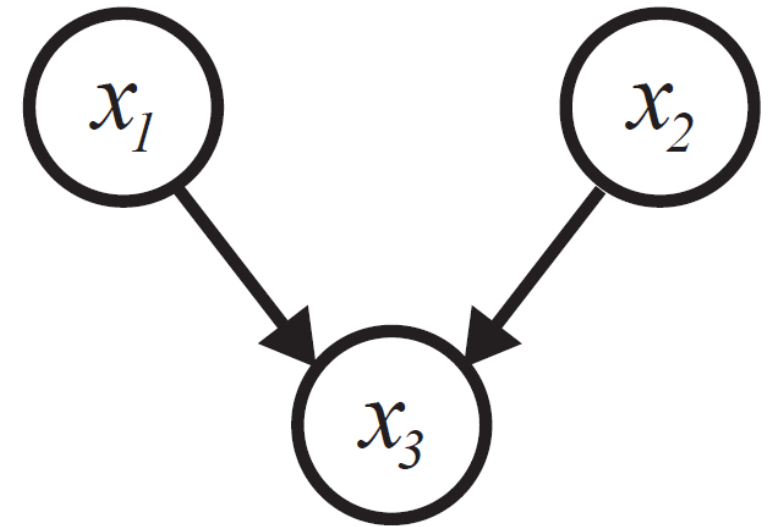
*How can we be sure a PGM is **correct** for a distribution $p(x)$?*

Probability Model:

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



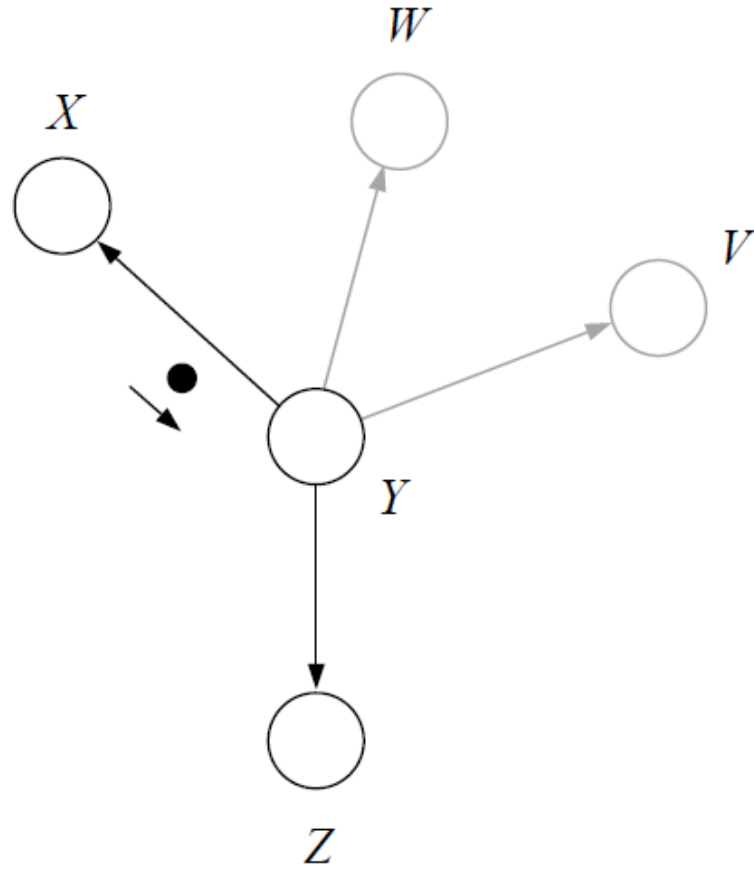
Graphical Model:



It must satisfy **all** of the conditional independencies of $p(x)$, then we say $p(x)$ is **Markov with respect to** the graph.

Bayes Ball Algorithm

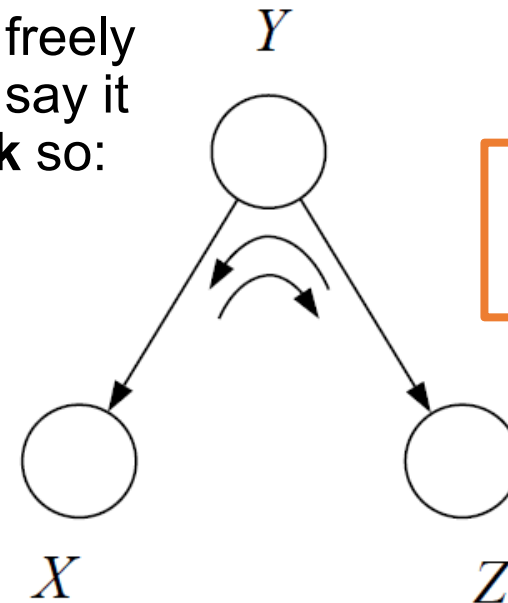
To test if $X \perp Z \mid Y$ imagine rolling a “ball” from X towards Z



The ball follows rules defined by the canonical 3-node subgraphs we’ve discussed

The ball passes freely from X-to-Z, we say it **does not block** so:

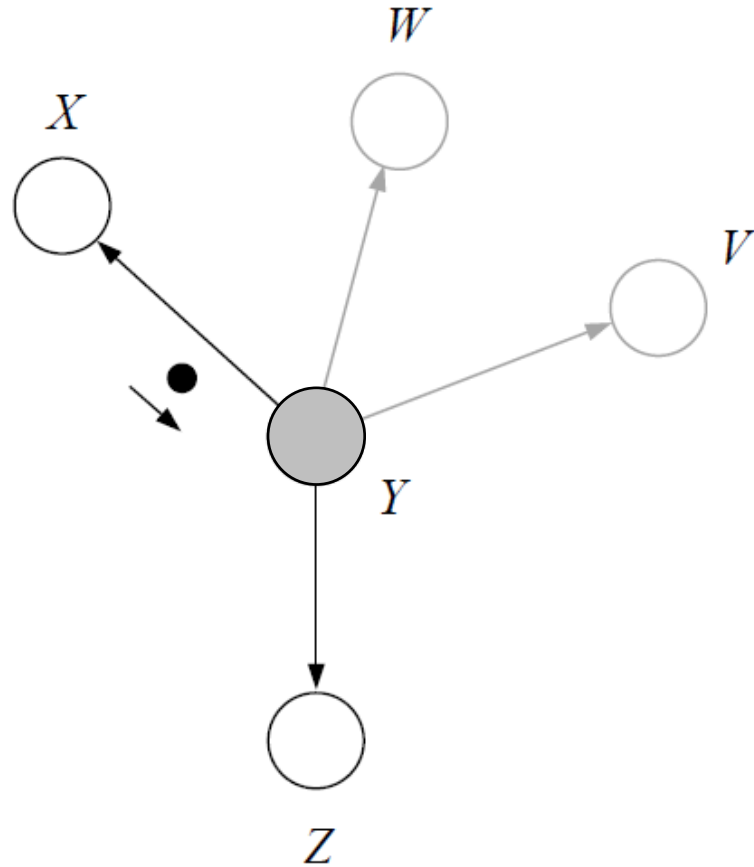
$$X \not\perp Z$$



**Reading:
Murphy Sec. 10.5**

Bayes Ball Algorithm

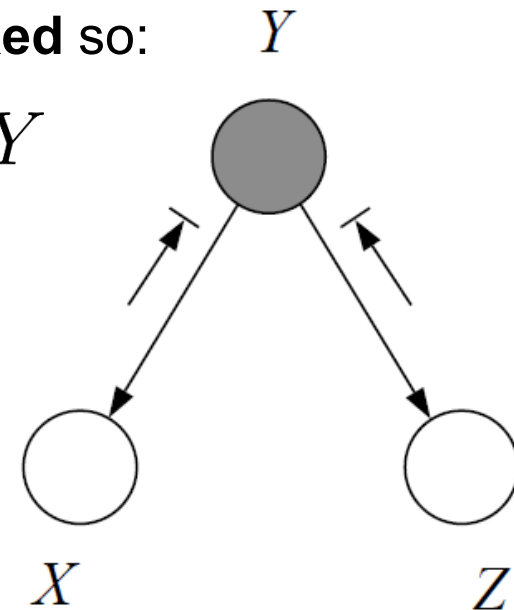
To test if $X \perp Z \mid Y$ imagine rolling a “ball” from X towards Z



The ball follows rules defined by the canonical 3-node subgraphs we’ve discussed

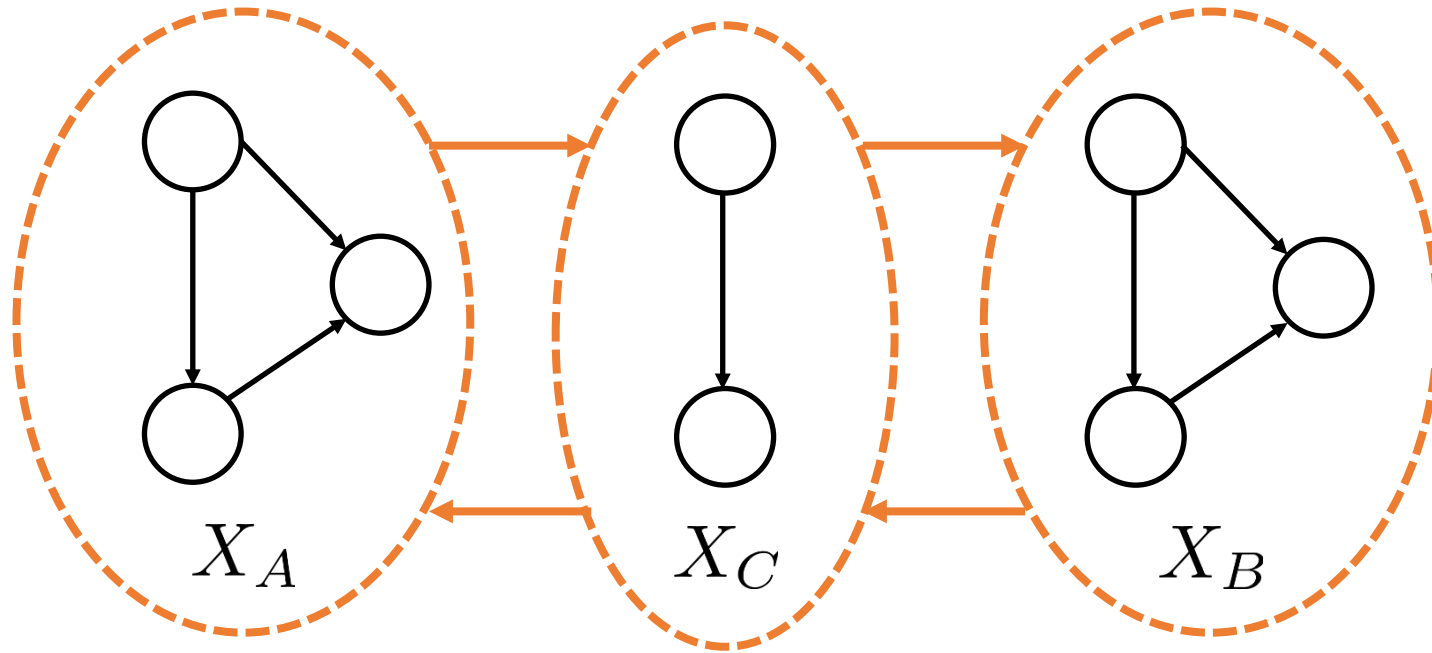
The ball is blocked so:

$$X \perp Z \mid Y$$



Directed Separation (d-Separation)

To test if $X_A \perp X_B \mid X_C$ roll ball from *every node in* $X_A \dots$



If *any* ball reaches *any* node in X_B then

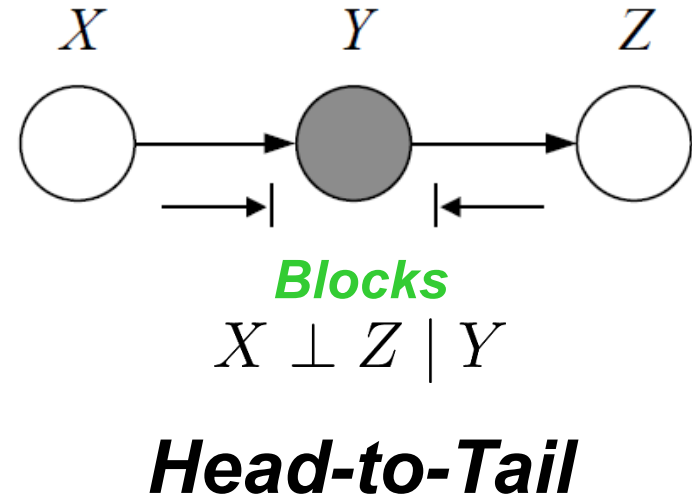
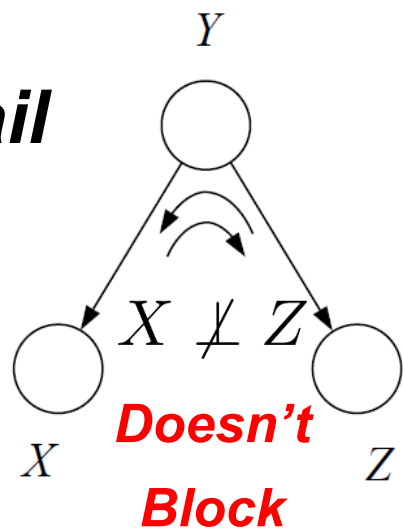
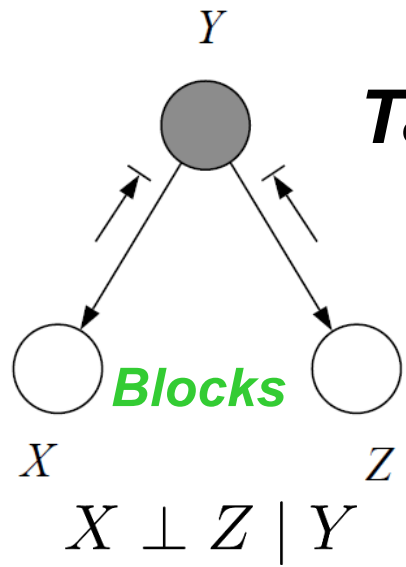
$$X_A \not\perp X_B \mid X_C$$

Otherwise:

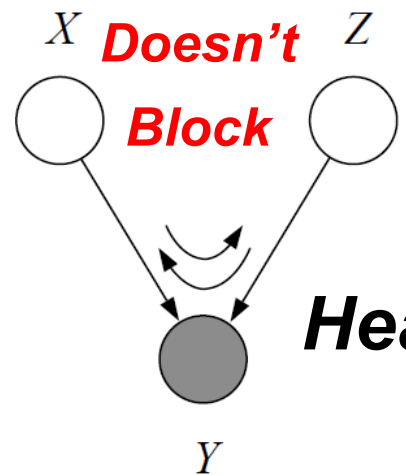
$$X_A \perp X_B \mid X_C$$

Tests for property of **directed separation (d-separation)**: if X_C separates / blocks X_A from X_B then $X_A \perp X_B \mid X_C$

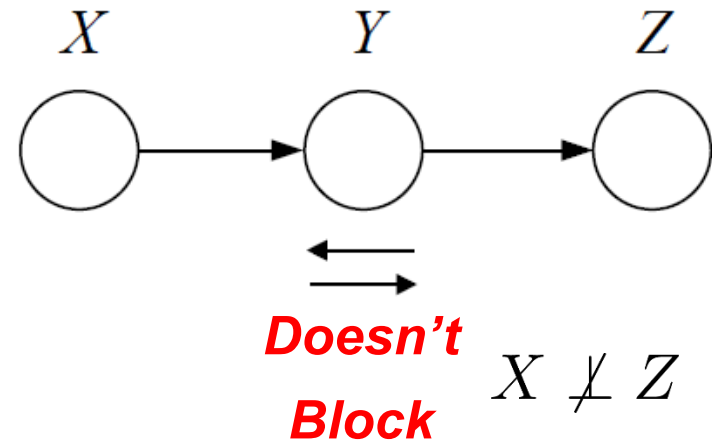
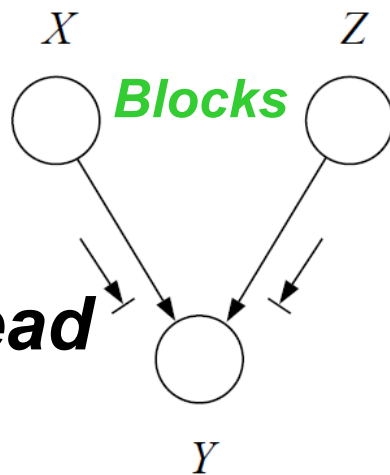
Bayes Ball Algorithm



$X \not\perp Z | Y$

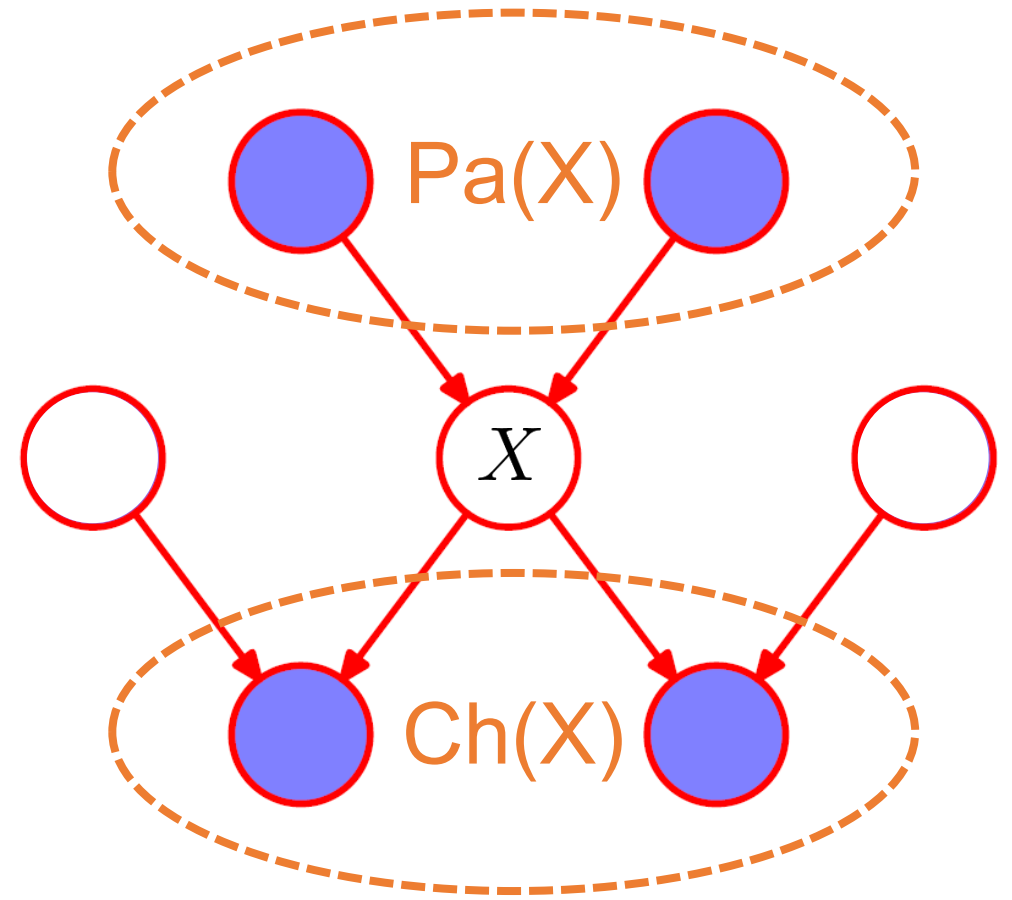


$X \perp Z$



Markov Blanket

Question Is X conditionally independent of all other nodes in graph given its **parents** and **children**?

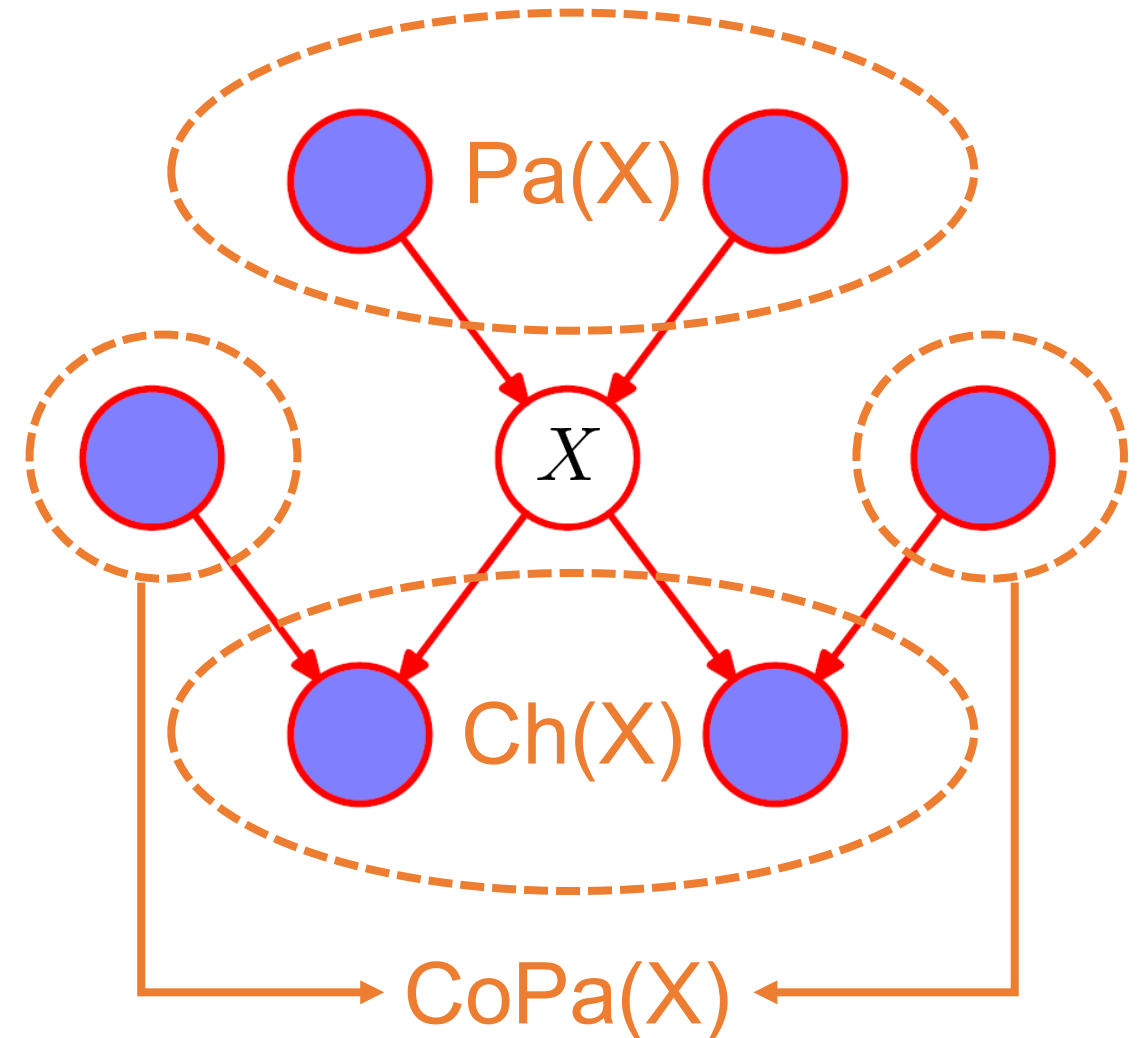


Markov Blanket

Question Is X conditionally independent of all other nodes in graph given its **parents** and **children**?

Answer No. It still depends on **co-parents**.

WHY?



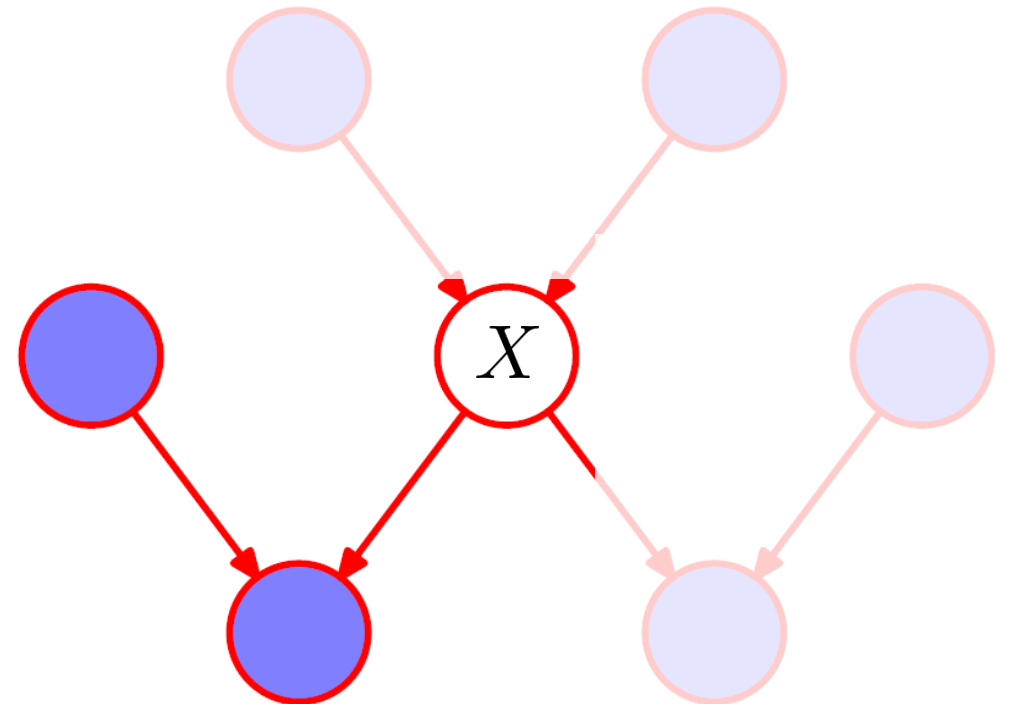
Markov Blanket

Question Is X conditionally independent of all other nodes in graph given its **parents** and **children**?

Answer No. It still depends on **co-parents**.

WHY?

Head-to-head conditional dependence from *explaining away* property



We refer to this conditioning set as the *Markov Blanket* of X ...

Markov Blanket

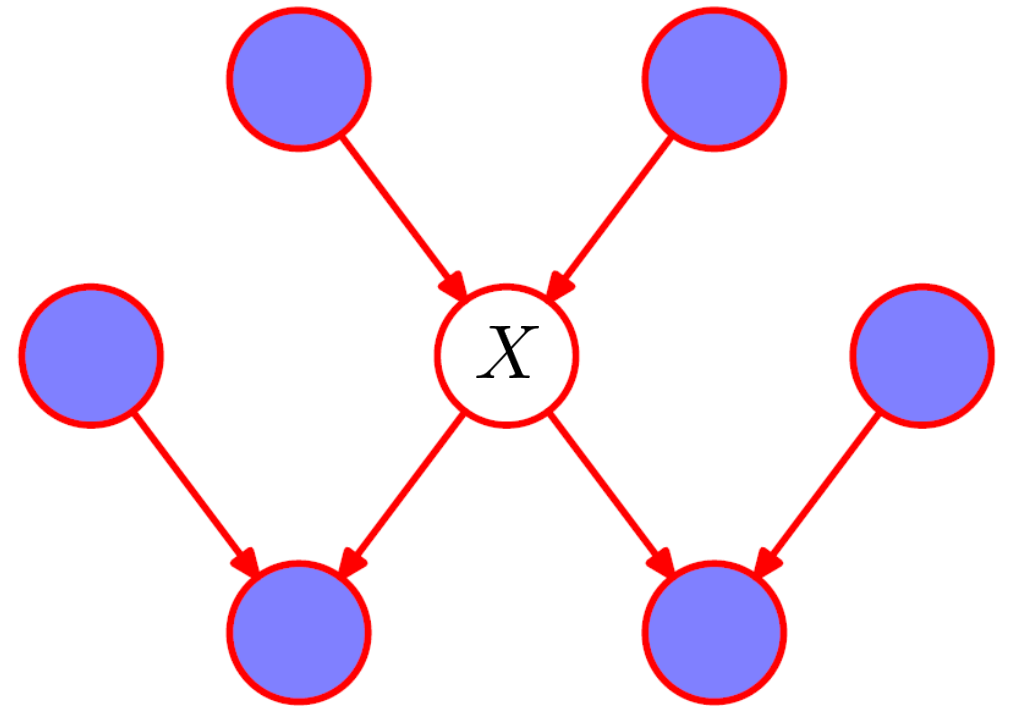
X conditionally independent of all other nodes, given its Markov blanket

Definition A RV X with distribution $p(x)$ that is Markov w.r.t. Bayes Net with graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has a **Markov blanket** given by:

$$\text{Mb}(X) = \text{Pa}(X) \cup \text{Ch}(X) \cup \text{CoPa}(X)$$

For any $Y \notin \text{Mb}(X)$:

$$X \perp Y \mid \text{Mb}(X)$$



Markov blanket used to simplify inference and distribute computation (e.g. Gibbs sampler, variational inference, etc.)

Outline

Directed graphical models

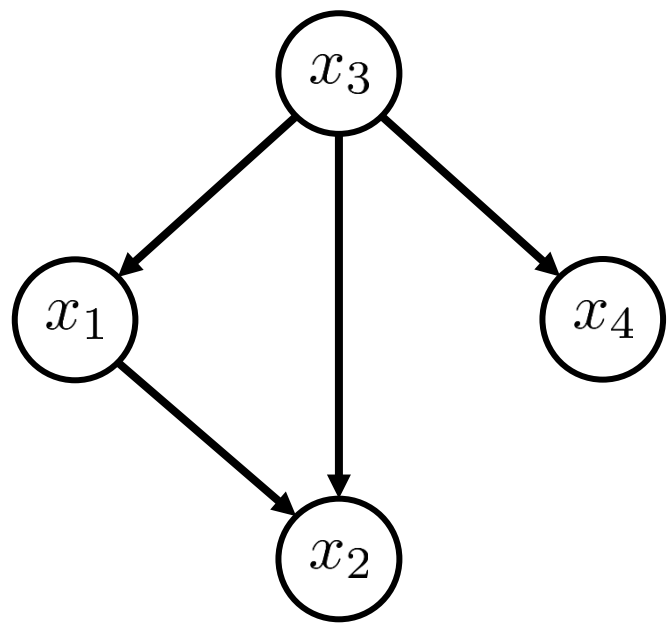
- Bayes Nets
- Conditional dependence

Undirected graphical models


- Markov random fields (MRFs)
- Factor graphs

Directed PGM = Bayes Network

Model factors are normalized conditional distributions:



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

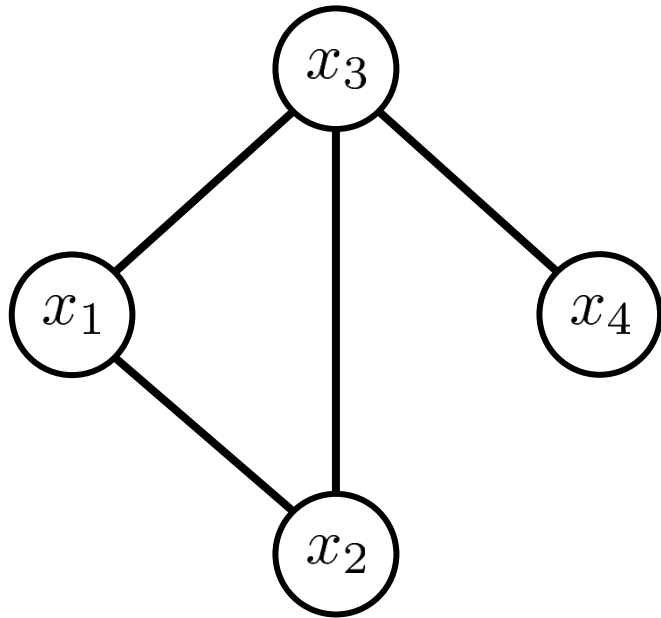
 Parents of node s

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...



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

Functions model how variables interact

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 convertes to MRF by dropping local normalizers
- MRF \rightarrow Bayes Net not straightforward

Factorized Probability Distributions

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

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

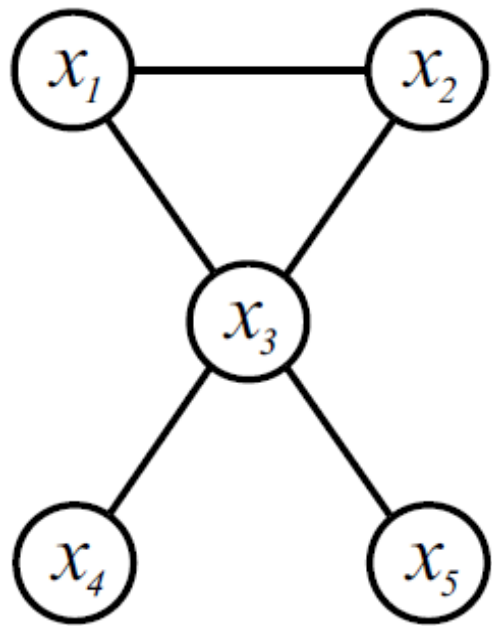
Where:

- \mathcal{C} a collection of subsets of indices $\{1, \dots, d\}$
- $\psi(\cdot)$ 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),

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

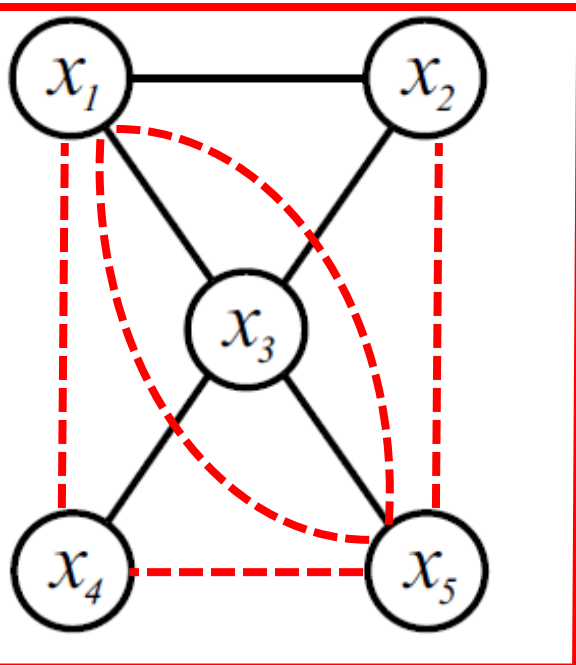
Markov Random Fields (MRFs)

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*”:

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

Complete Graph



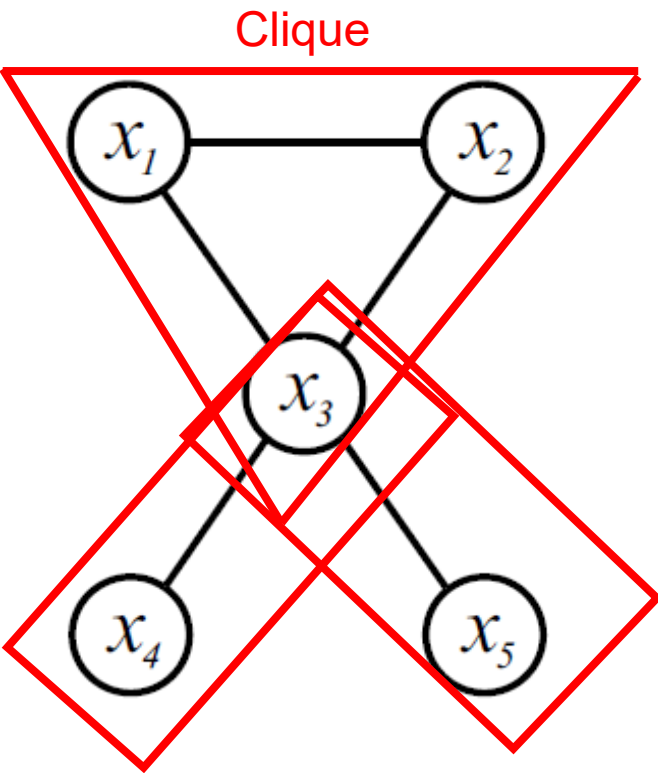
Markov Random Fields (MRFs)

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*”:

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

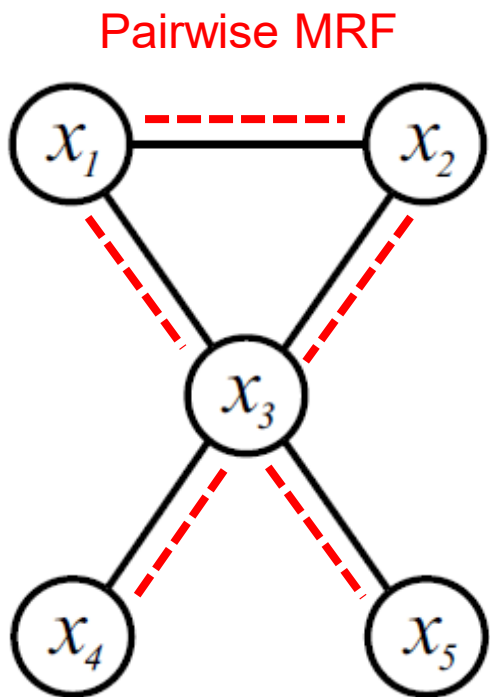
$$\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$



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*”:



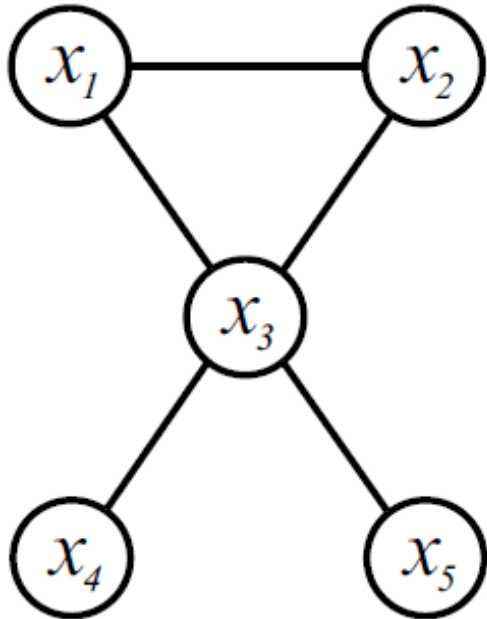
$$\psi(x_1, x_2, x_3, x_4, x_5)$$

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

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

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

Example



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,

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

Question What named distribution is $p(x)$?

Answer Multivariate Gaussian

$$p(x) = \mathcal{N}(x \mid \mu, \Sigma)$$

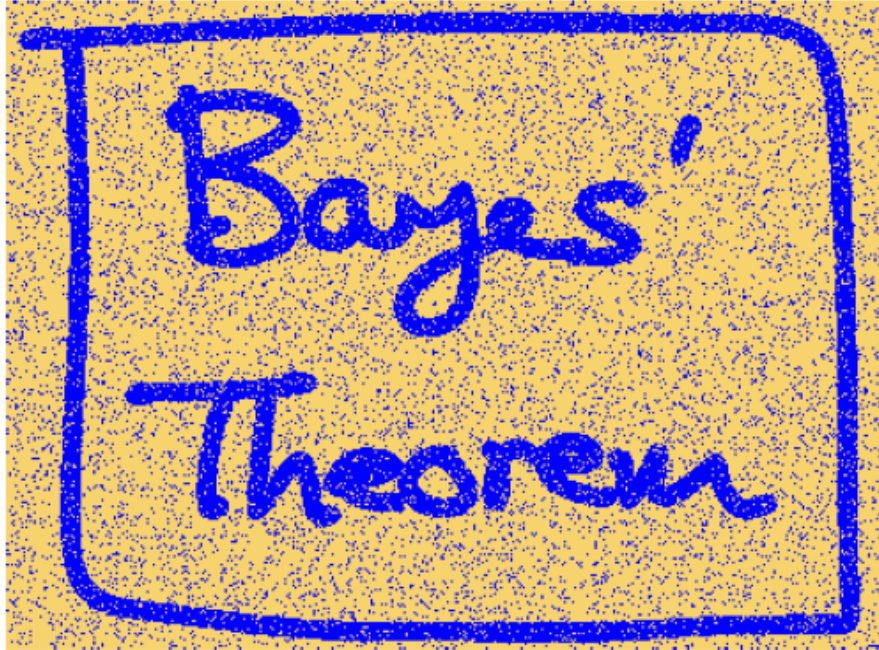
$$Z = (2\pi)^{5/2} |\Sigma|^{1/2}$$

Can easily read off
inverse covariance...

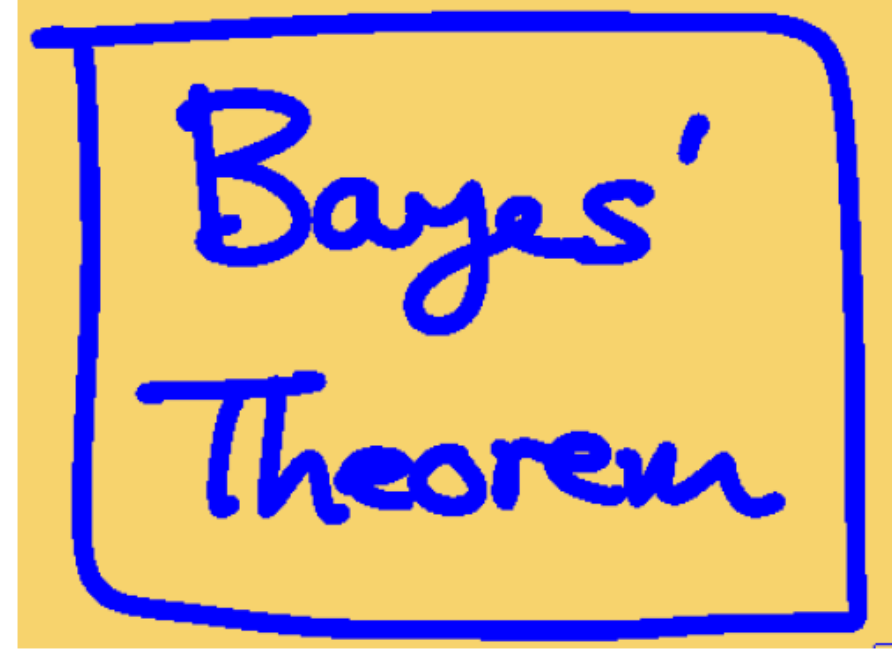
$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 4 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Example: Image Denoising

Noisy Image



Latent Image



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

Example: Image Denoising

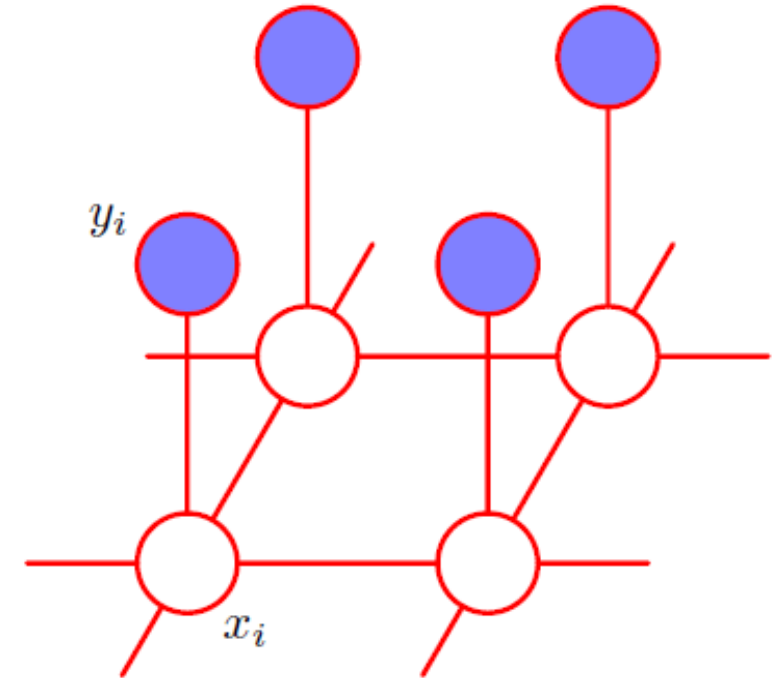
Model Assume binary image with latent pixels $x_i \in \{-1, +1\}$ and observed $y_i \in \{-1, +1\}$

Observed pixels randomly flipped i.i.d.,

$$\log \phi_i(x_i) = \eta x_i y_i \quad \text{Eta parameter controls noise}$$

Neighboring pixels should appear similar,

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



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 \rightarrow maximize probability)

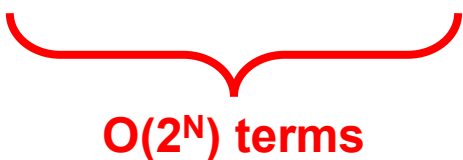
Normalizing MRFs

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:

$$Z = \sum_{x_1} \sum_{x_2} \dots \sum_{x_N} \exp \{-E(x, y)\}$$

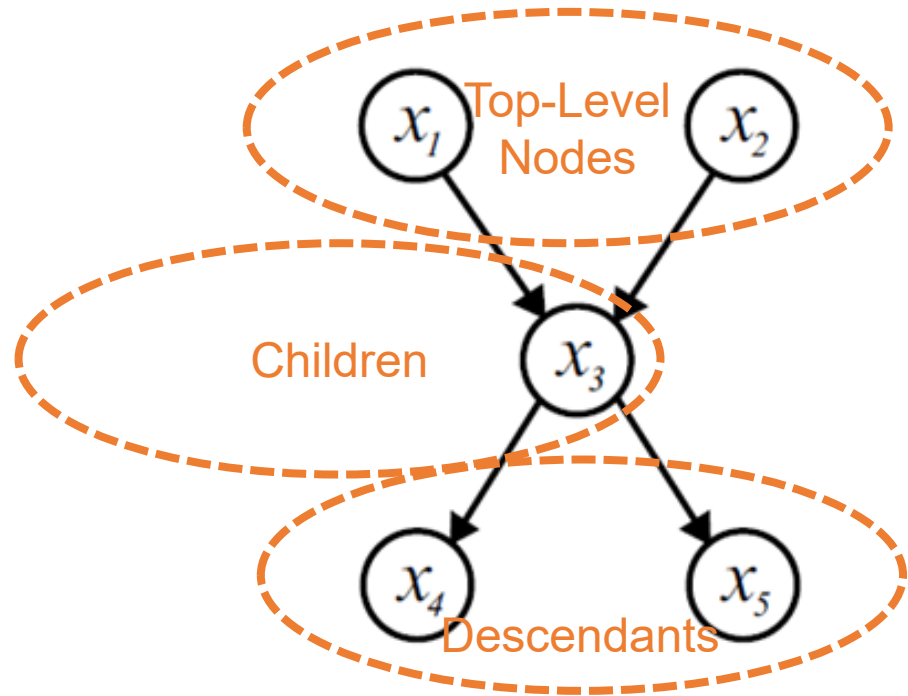


O(2^N) terms

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:

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

so

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

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

Conditional Independence (Undirected)

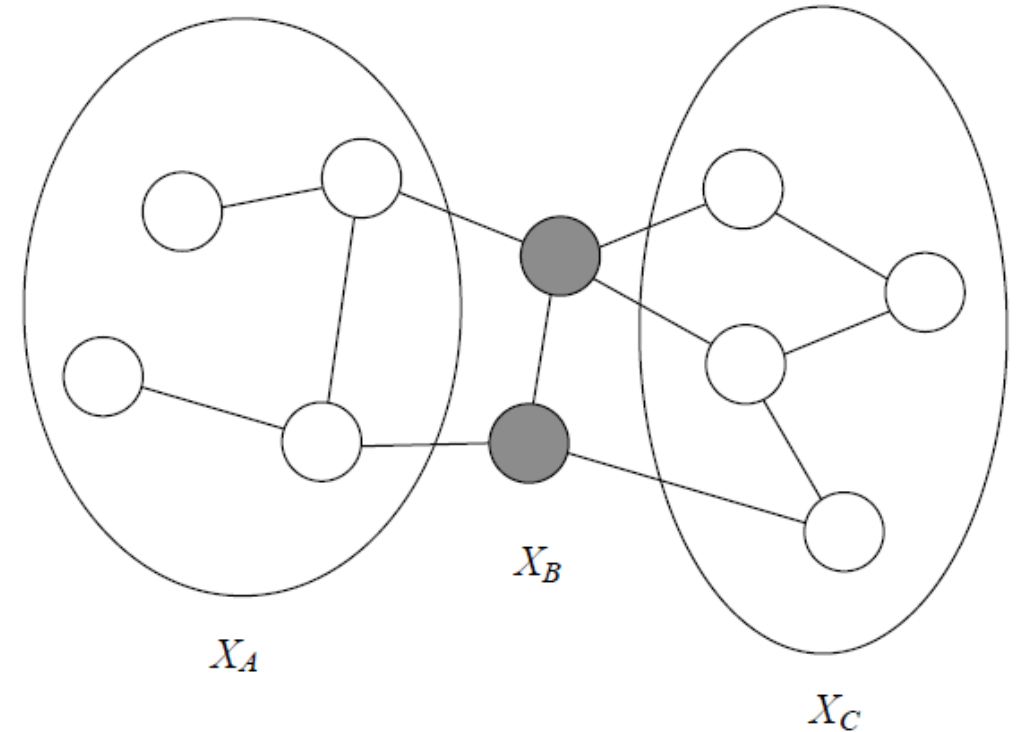
We say x_A and x_C are *independent* or $x_A \perp x_C$ if:

$$p(x_A, x_C) = p(x_A)p(x_C)$$

We say they are *conditionally independent* or $x_A \perp x_C \mid x_B$ if:

$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B)$$

Def. We say $p(x)$ is *globally Markov* w.r.t. \mathcal{G} if $x_A \perp x_C \mid x_B$ in any separating set of \mathcal{G} .



Conditional independence in undirected graphical models is defined by separating sets

Global & Local Markov Properties

Global Markov Property

- Set B **separates** A from C if all paths from A to C pass through B
- By definition, **distribution is Markov** if and only if for any B separating A and C:

$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B)$$

$$p(x_A \mid x_B, x_C) = p(x_A \mid x_B) \quad p(x_C \mid x_B, x_A) = p(x_C \mid x_B)$$

Local Markov Property

- Given its **neighbors**, each node is independent of all other variables

$$p(x_s \mid x_{\mathcal{V} \setminus s}) = p(x_s \mid x_{\Gamma(s)})$$

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

Markov blanket only includes immediate neighbors (we needed co-parents in Bayes nets)

- This local Markov property is a special case of the global Markov property

Hammersley-Clifford Theorem

Theorem (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} :*

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

Conversely, any strictly positive density which is Markov with respect to \mathcal{G} can be represented in this factored form.

Global Markov Property

(Graph Separation Implies Independence)



Joint Factorization

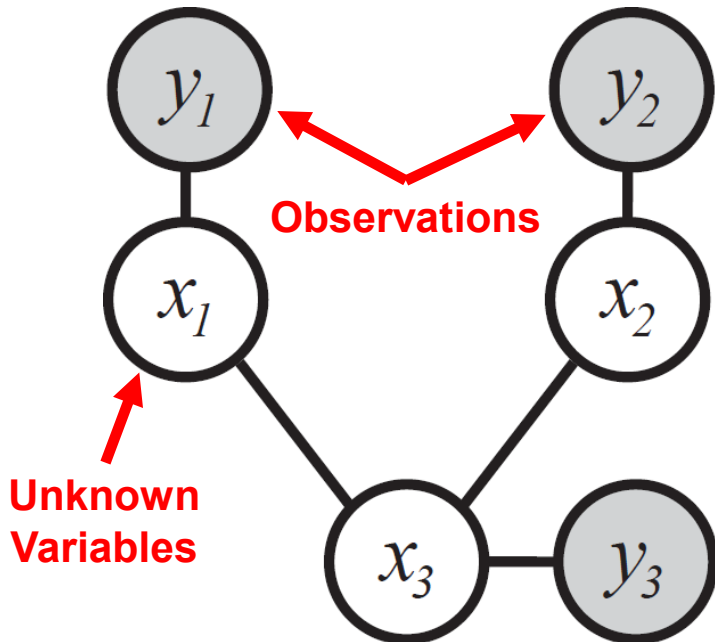
(Potential Function for Each Clique)

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**

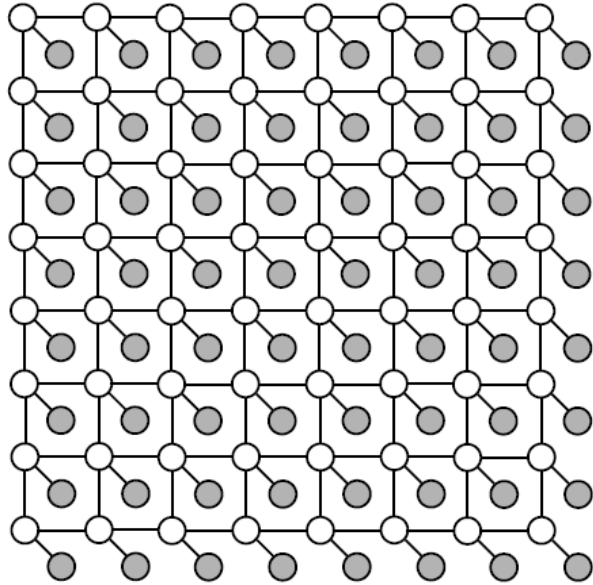


Restricted class of MRFs

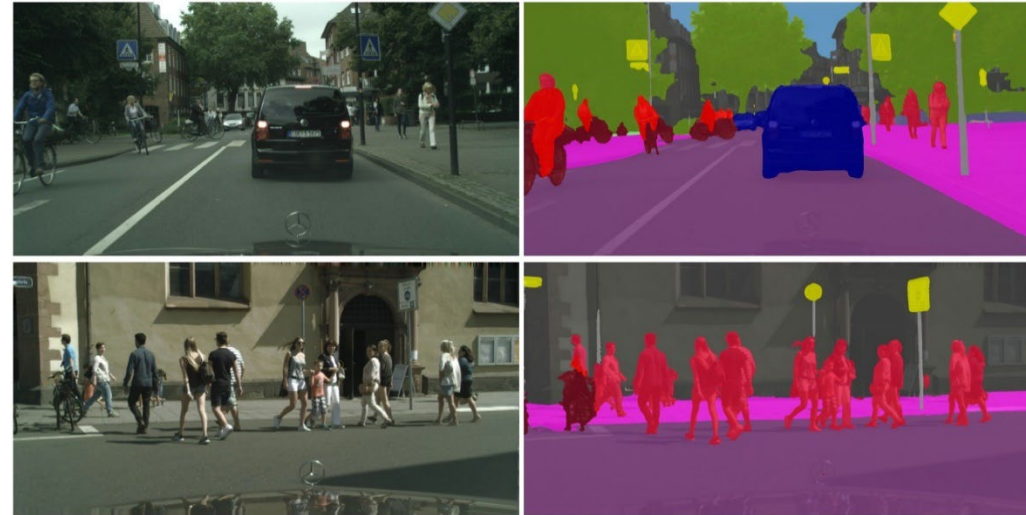
- 2-node factor exists for every edge
- Explicit factorization of joint distribution
- High-order factors not always easily decomposed into pairwise terms

Example: Image Segmentation

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



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



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

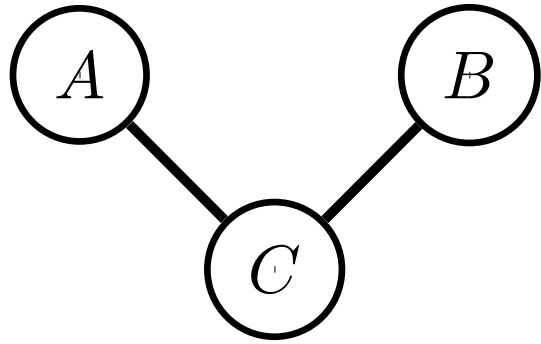
Don't need to specify normalized conditionals as in Bayes Nets

Low energy configurations = High probability

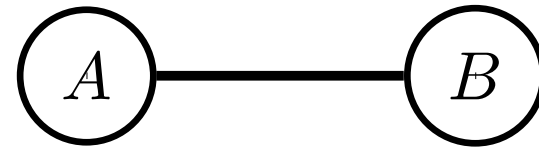
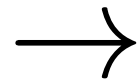
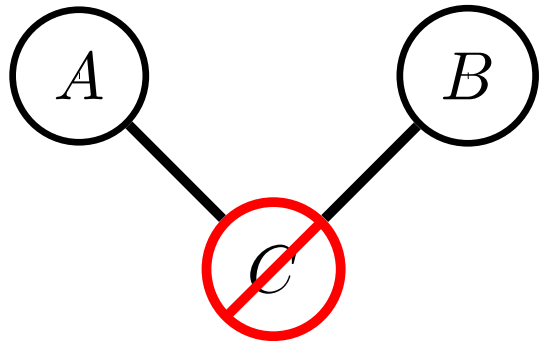
L2 Likelihood: $\psi_i(x_i, y_i) = \|x_i - y_i\|^2$ **Potts model:** $\psi_{i,j}(x_i, x_j) = \mathbb{I}(x_i \neq x_j)$

MAP (minimum energy) configuration = Piecewise constant regions

Transformations of Undirected Models



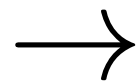
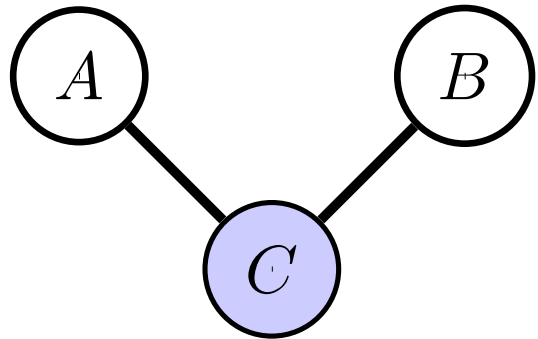
$$p(A, B, C) = \psi_{AC}(A, C)\psi_{BC}(B, C)/Z$$



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

Marginalization: Join all nodes that have path through C

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



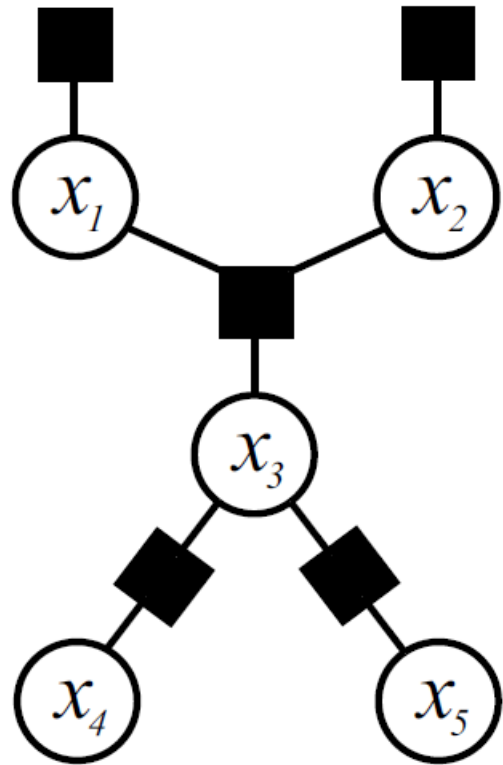
$$p(A, B|C) = p(A|C)p(B|C)$$

Conditioning: Drop all edges on path through C

Conditioning on C makes A and B independent:

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

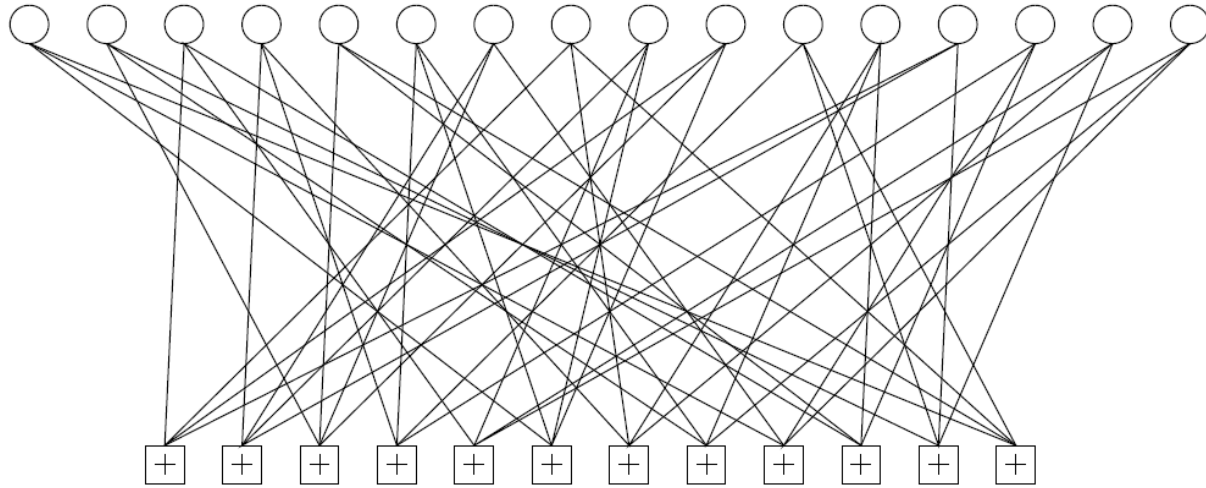
$$p(x) \propto \prod_{f \in \mathcal{F}} \psi_f(x_f)$$

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*)

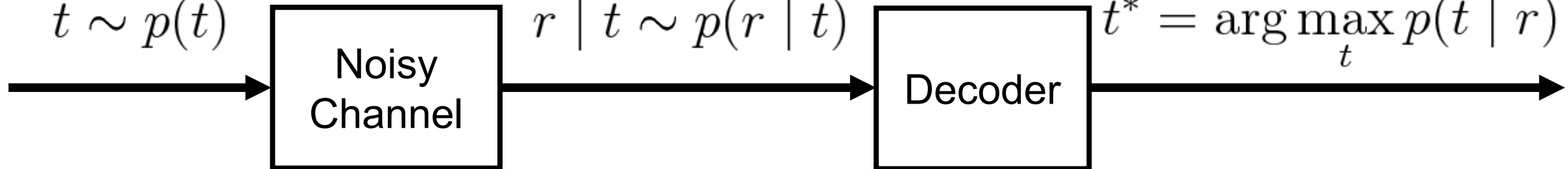
Transmitted Code

$$t \sim p(t)$$

Received Code

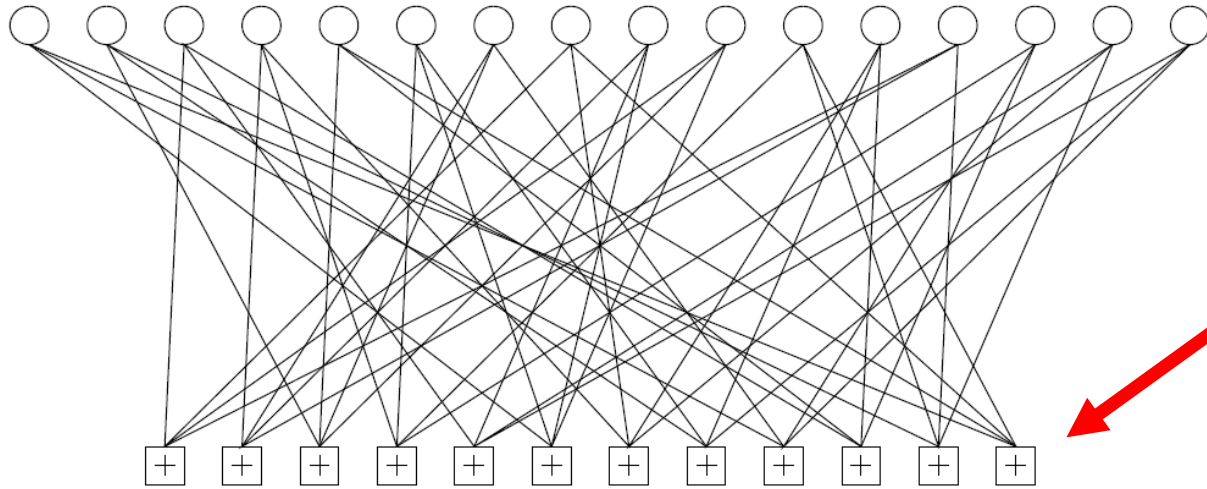
$$r \mid t \sim p(r \mid t)$$

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



Example: Low Density Parity Check Codes

Factor Graph Representation



Sparse Parity Check Matrix

$$\mathbf{H} = \begin{bmatrix}
 1 & & & & & & & & & & & & & & \\
 & 1 & & & & & & & & & & & & & \\
 & & 1 & & & & & & & & & & & & \\
 & & & 1 & & & & & & & & & & & \\
 & & & & 1 & & & & & & & & & & \\
 & & & & & 1 & & & & & & & & & \\
 & & & & & & 1 & & & & & & & & \\
 & & & & & & & 1 & & & & & & & \\
 & & & & & & & & 1 & & & & & & \\
 & & & & & & & & & 1 & & & & & \\
 & & & & & & & & & & 1 & & & & \\
 & & & & & & & & & & & 1 & & & \\
 & & & & & & & & & & & & 1 & & \\
 & & & & & & & & & & & & & 1 & \\
 & & & & & & & & & & & & & & 1
 \end{bmatrix}$$

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

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

n-th bit \longrightarrow n

[Source: David MacKay]

Recap: Directed Models

- Distribution factorized as product of conditionals via chain rule

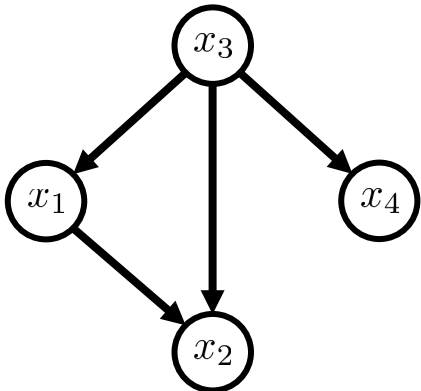
$$p(x_1, x_2, x_3, x_4) = p(x_3)p(x_1 | x_3)p(x_4 | x_1, x_3)p(x_2 | x_1, x_3, x_4)$$

- Choose ordering where terms simplify due to conditional independence

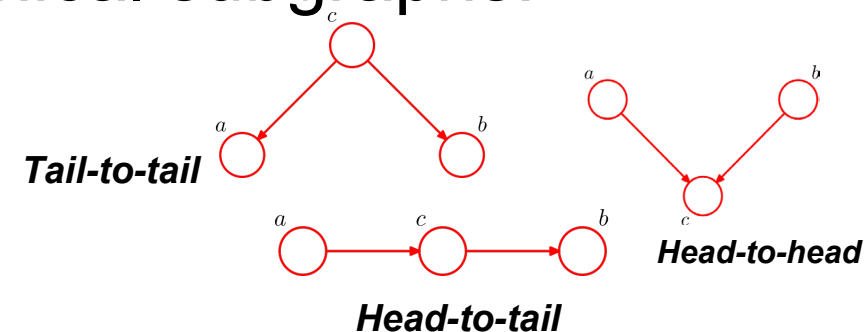
Eg. Suppose $x_4 \perp x_1 | x_3$ and $x_2 \perp x_4 | x_1$ then:

$$p(x) = p(x_3)p(x_1 | x_3)p(x_4 | x_3)p(x_2 | x_1, x_3)$$

- Directed graph encodes factorized distribution via conditional independence properties



- Test independence using canonical subgraphs:
- Straightforward simulation via **ancestral sampling**



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
- Easier to verify Markov independence via *separating sets*
- Factorization ambiguous in MRFs, but explicit in factor graphs (factor graphs generally preferred)
- Simulation is not easy in general. Can't do ancestral sampling.

