Computer Science

CSC 665-1: Advanced Topics in Probabilistic Graphical Models

Graphical Models

Instructor: Prof. Jason Pacheco

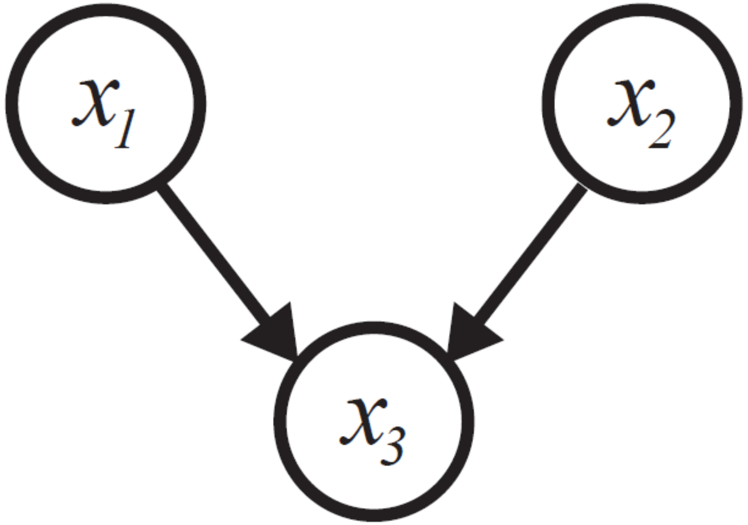
From Probabilities to Pictures

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

Graphical Model:

Probability Model:

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

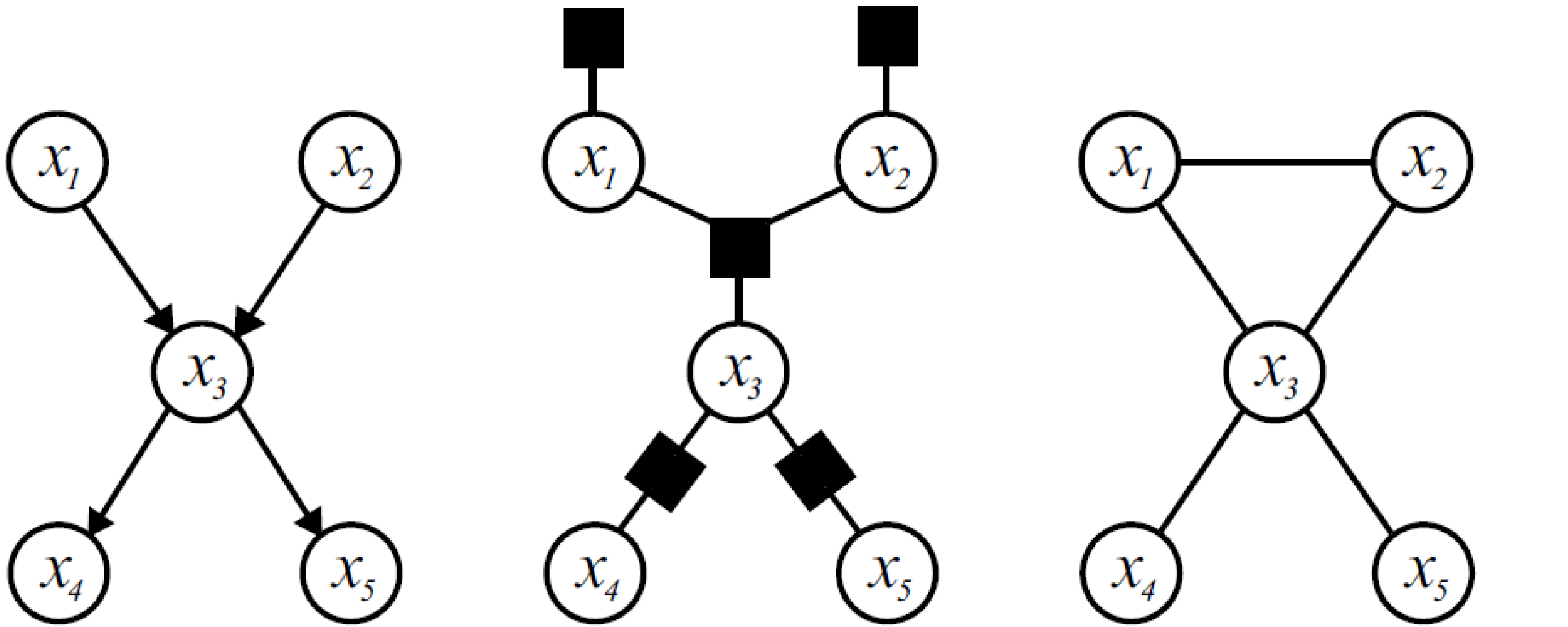


The graphical model structure *obeys* the factorization of the probability function in a sense we will formalize later

*** We will use the term "distribution" loosely to refer to a CDF / PDF / PMF**

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]

Factorized Probability Distributions

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

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

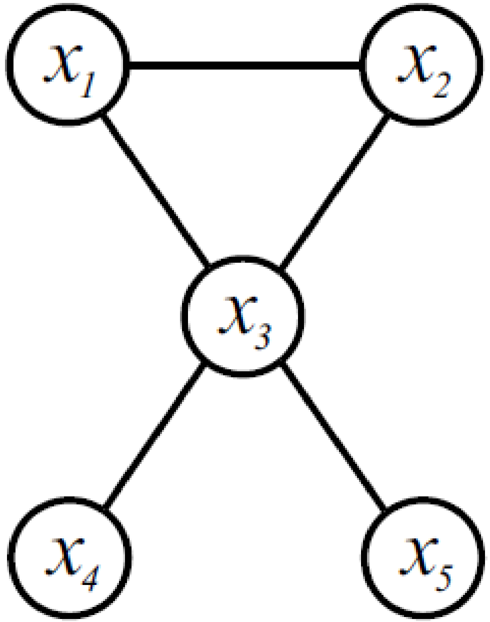
<u>Where:</u>

- $\mathcal{C}$ 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,

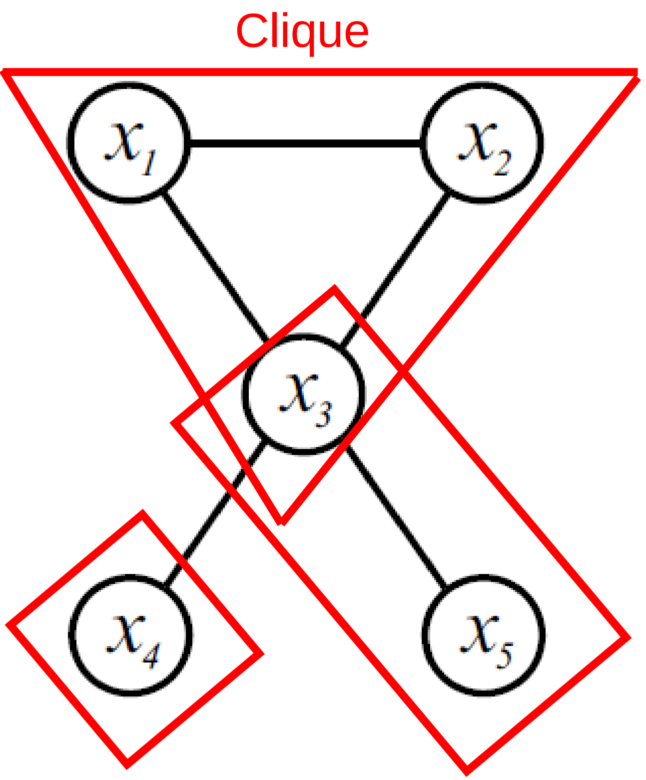
$$(s,t) \in \mathcal{E} \Leftrightarrow (t,s) \in \mathcal{E}$$

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*":

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

A factorization is *valid* if it satisfies the *Global Markov property*, defined by conditional independencies

Conditional Independence (Undirected)

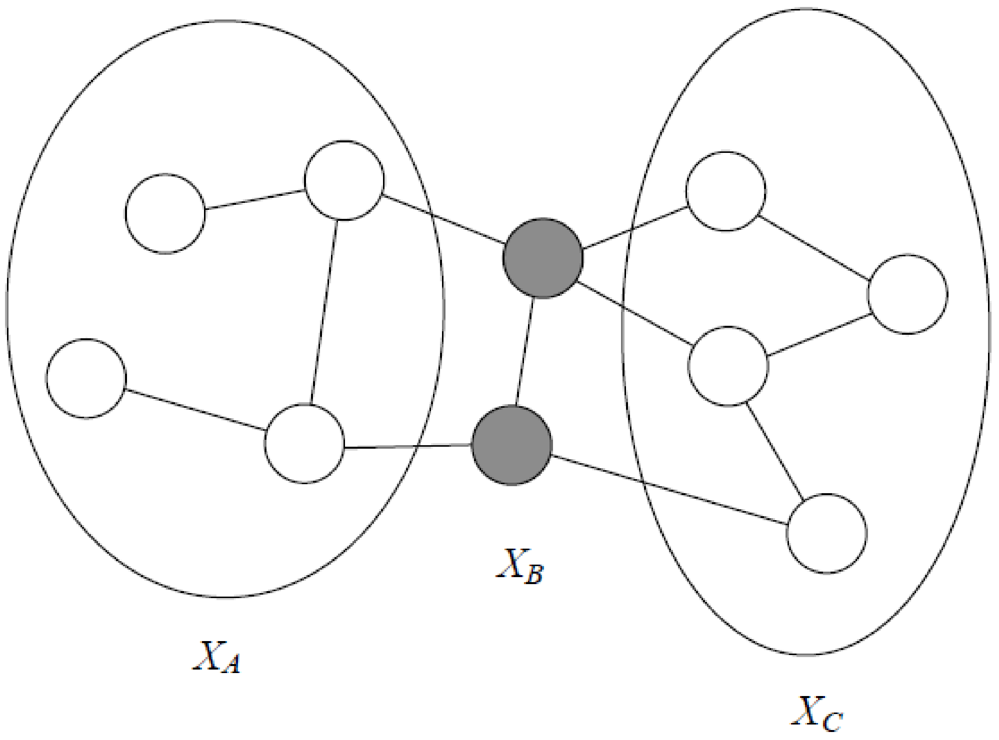
We say x_A and x_C are *independent* or $x_A \perp\!\!\!\perp x_C$ if:

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

We say they are *conditionally independent* or $x_A \perp\!\!\!\perp x_C \mid x_B$ 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 $x_A \perp\!\!\!\perp x_C \mid x_B$ in any separating set of $\mathcal{G}$.



Conditional independence in undirected graphical models is defined by separating sets

[Source: Michael I. Jordan]

Hammersley-Clifford Theorem

Thorem (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 $\mathcal{G}$:*

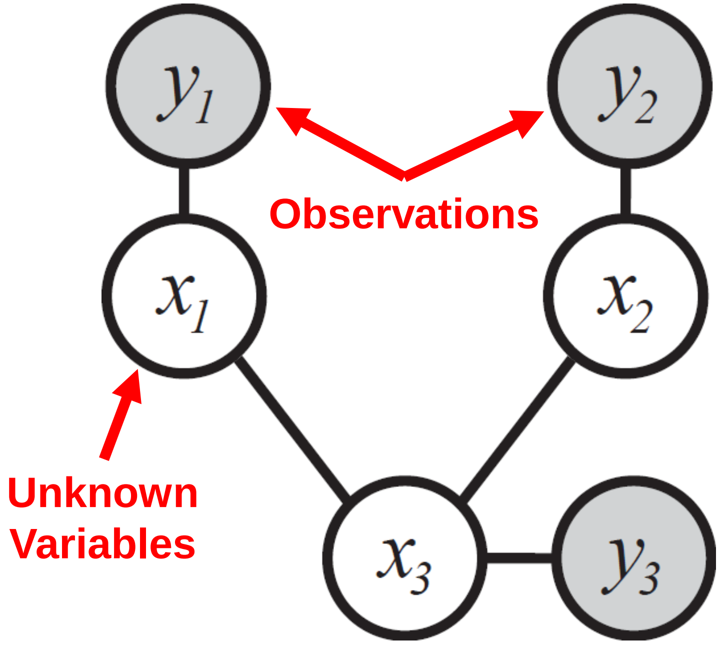
$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

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

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

Pairwise Markov Random Field

Often easier to specify and do inference on pairwise model

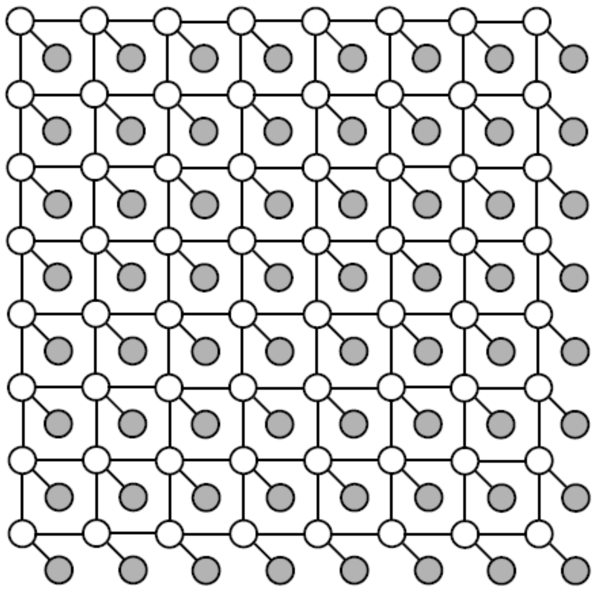


$$p(x, y) \propto \prod_{s \in \mathcal{V}} \underbrace{\psi_s(x_s, y)}_{\text{Likelihood}} \prod_{(s,t) \in \mathcal{E}} \underbrace{\psi_{st}(x_s, x_t)}_{\text{Prior}}$$

Restricted class of MRFs

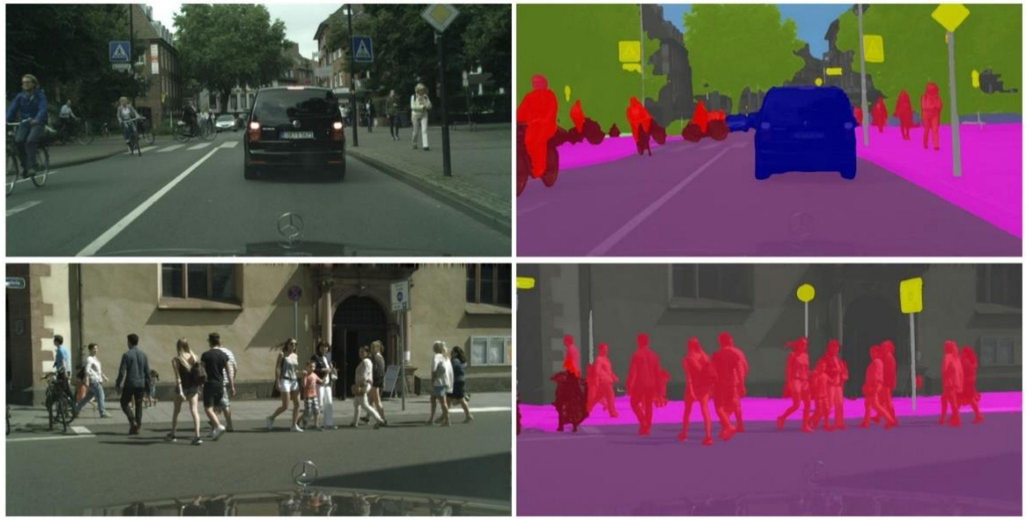
- 2-node factor exists for every edge
- Explicit factorization of joint distribution
- High-order factors not always easily decomposed into pairwise terms

Example: Image Segmentation



Notional figure only!

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



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

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

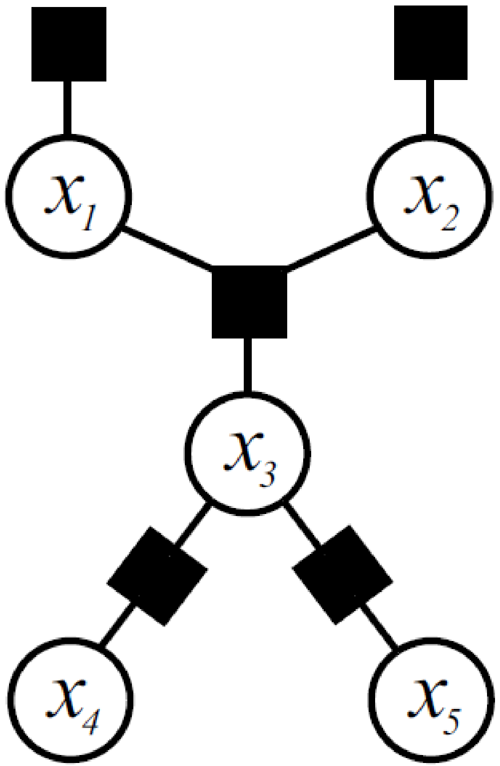
Low energy configurations = High probability

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

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 graphs explicitly encode factorization of distribution:

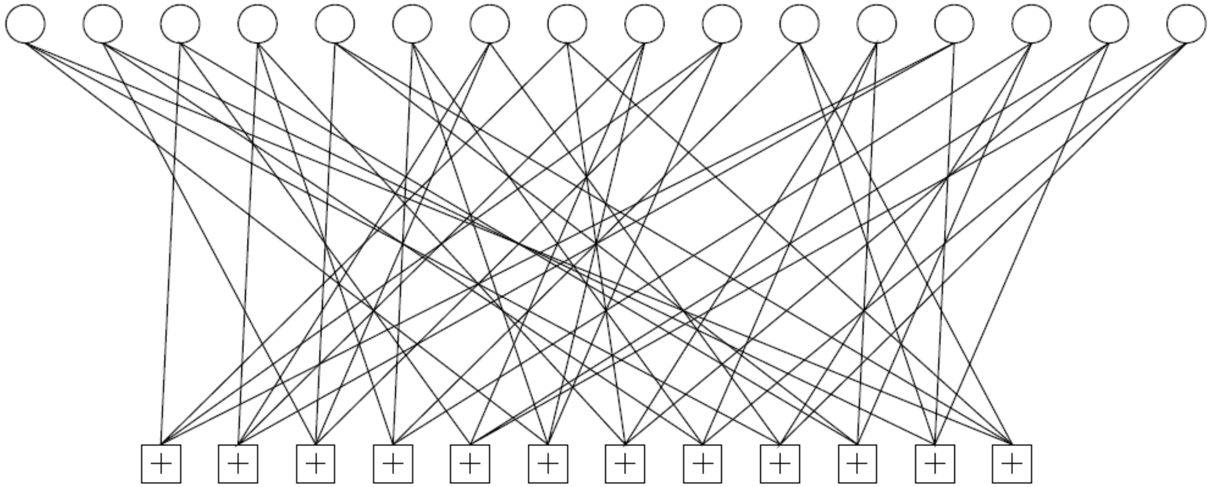
$$p(x) \propto \prod_{f \in \mathcal{F}} \psi_f(x_f)$$

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

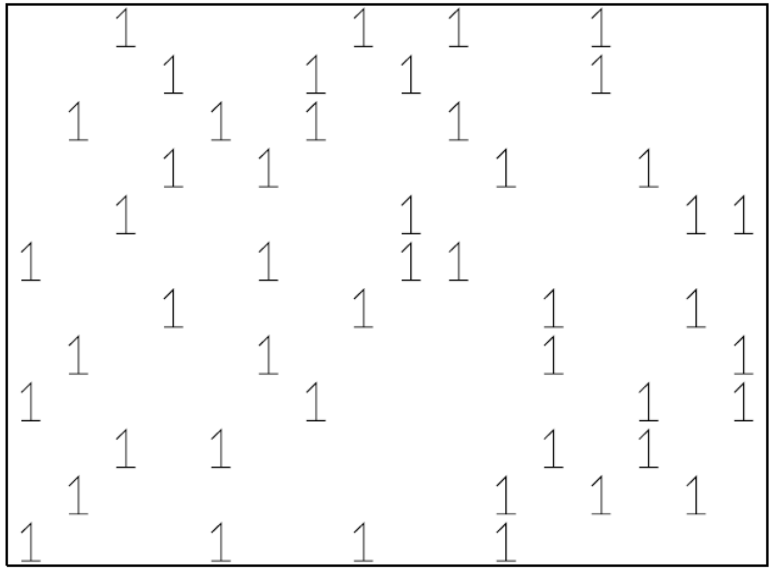
$$\psi(x_1)\psi(x_2)\psi(x_1, x_2, x_3)\psi(x_3, x_4)\psi(x_3, x_5)$$

Example: Low Density Parity Check Codes

Factor Graph Representation



Sparse Parity Check Matrix

$\mathbf{H} =$ 

Transmitted Code

$t \sim p(t)$

Noisy Channel

Received Code

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

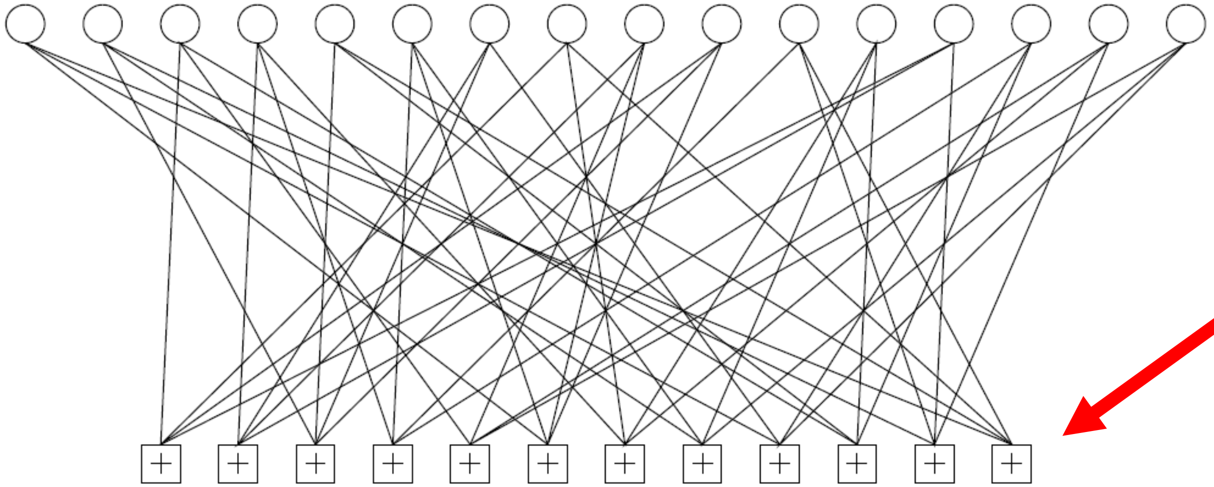
Decoder

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

[Source: David MacKay]

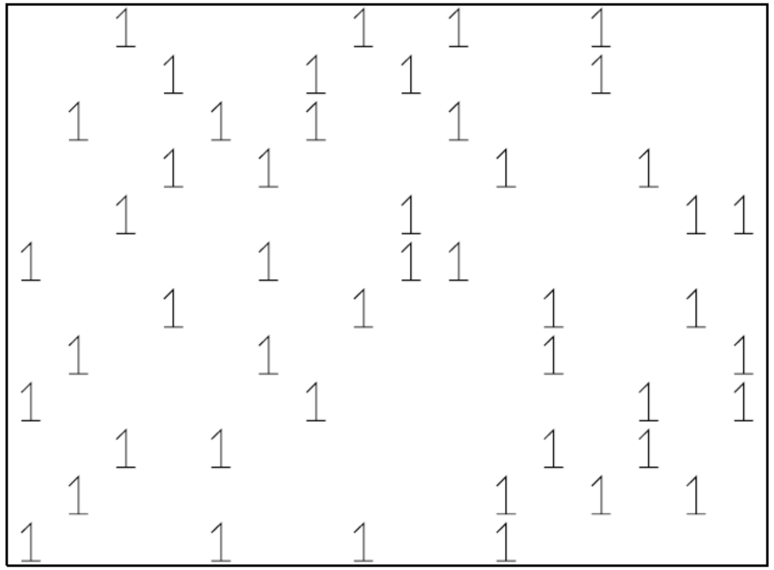
Example: Low Density Parity Check Codes

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:

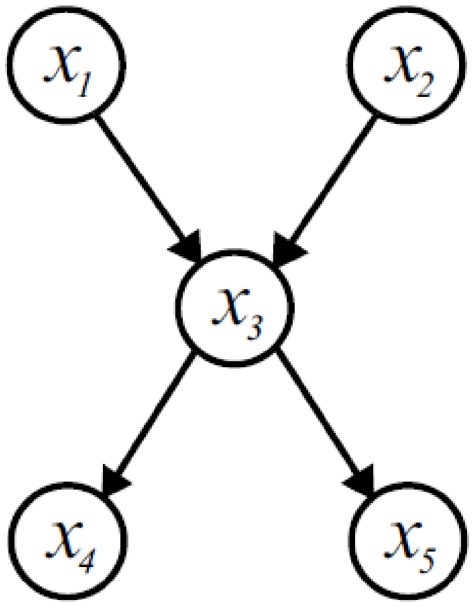
$$p(r \mid t) = \prod_n p(r_n \mid t_n) = \prod_n (1-\epsilon)^{\mathbb{I}(r_n = t_n)} \epsilon^{\mathbb{I}(r_n \neq t_n)}$$

n-th bit

[Source: David MacKay]

Directed Graphs

Def. A <u>directed graph</u> 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. <u>Parents</u> of vertex $t \in \mathcal{V}$ are given by the set of nodes with arcs pointing to t ,

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

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

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

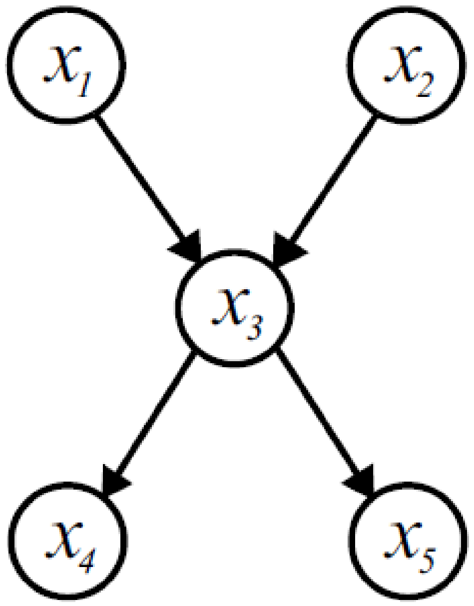
<u>Ancestors</u> are parents-of-parents.
<u>Descendants</u> are children-of-children.

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:

$$p(x_1)p(x_2)p(x_3 \mid x_1, x_2)p(x_4 \mid x_3)p(x_5 \mid x_3)$$

Locally normalized factors yield *globally normalized* joint probability

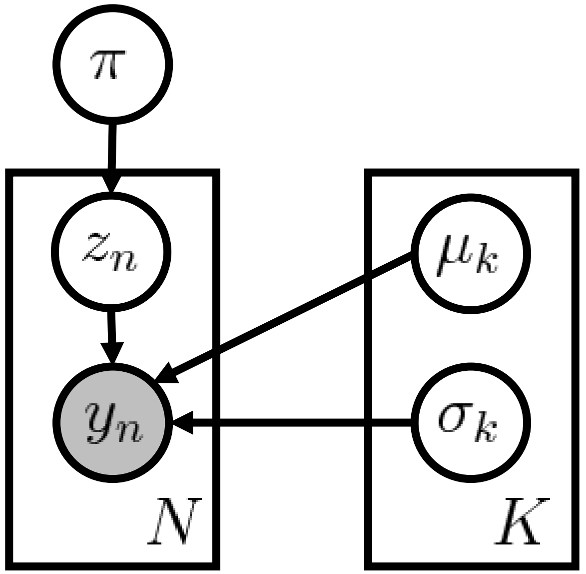
Example: Gaussian Mixture Model

Bayes nets are easily simulated via ancestral sampling

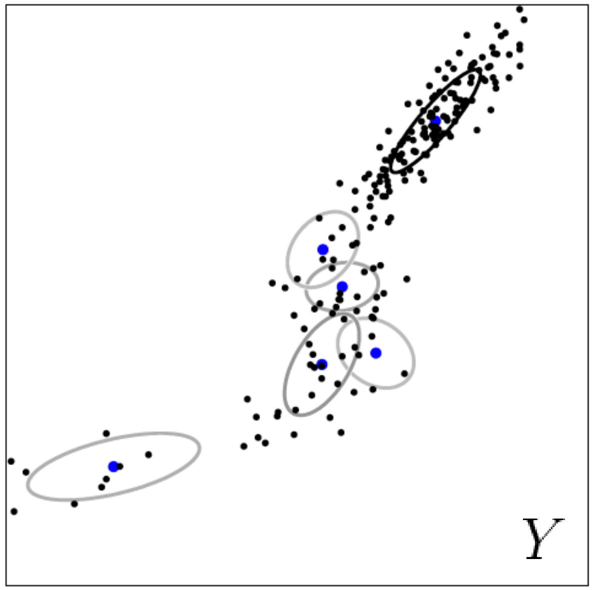
<u>Probability Model</u>

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

<u>Bayes Net</u>



<u>Joint Sample</u>

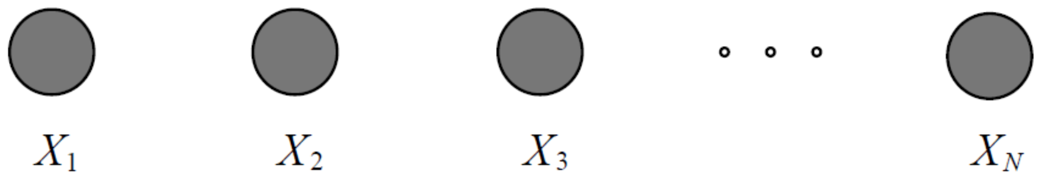


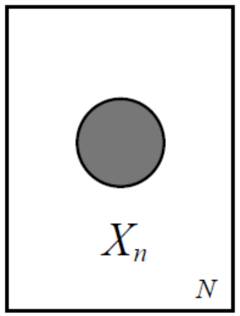
Specification is more difficult than undirected models since each factor must be a normalized probability measure

Plate Notation

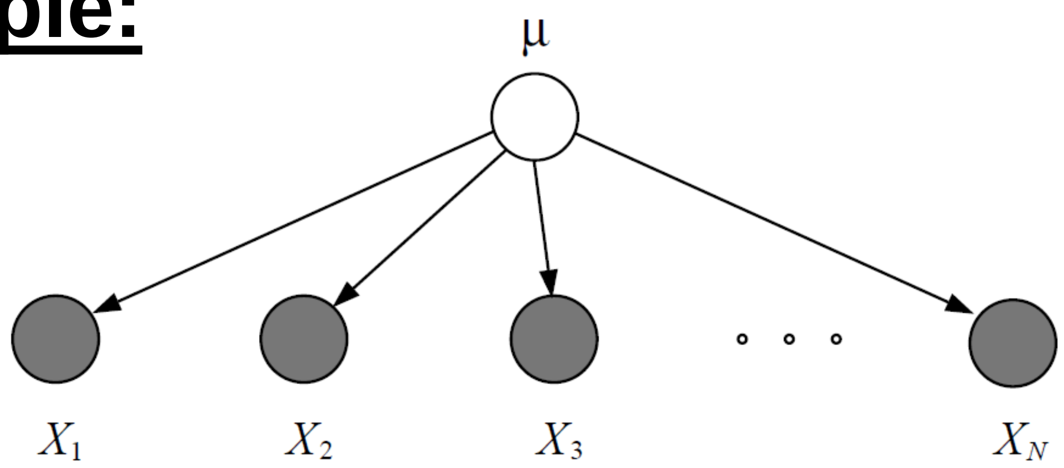
Plates denote replication of elements

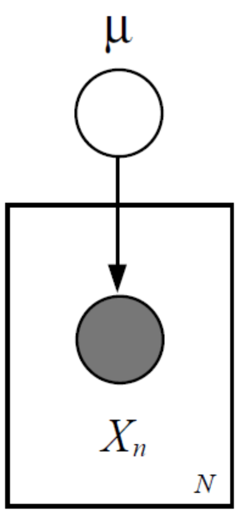
Example:



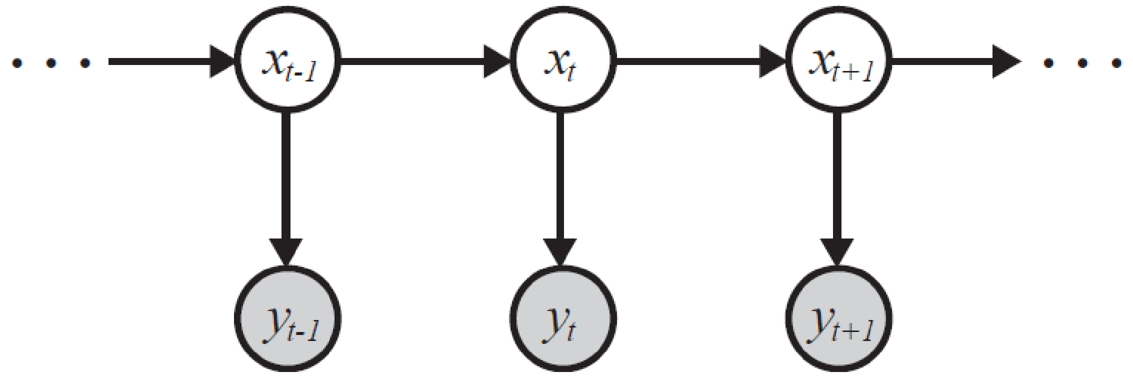


Example:





Example: Linear Gaussian Dynamics System



Latent state $x \in \mathbb{R}^D$ evolves according to linear dynamics.

Observations $y \in \mathbb{R}^M$ are linear functions of the state.

Conditional Probability Model:

$$x_t \mid x_{t-1} \sim \mathcal{N}(Ax_{t-1}, Q)$$

State Dynamics

Process Noise

$$y_t \mid x_t \sim \mathcal{N}(Cx_t, R)$$

Measurement Model

Observation Noise

State-Space Model (equivalent):

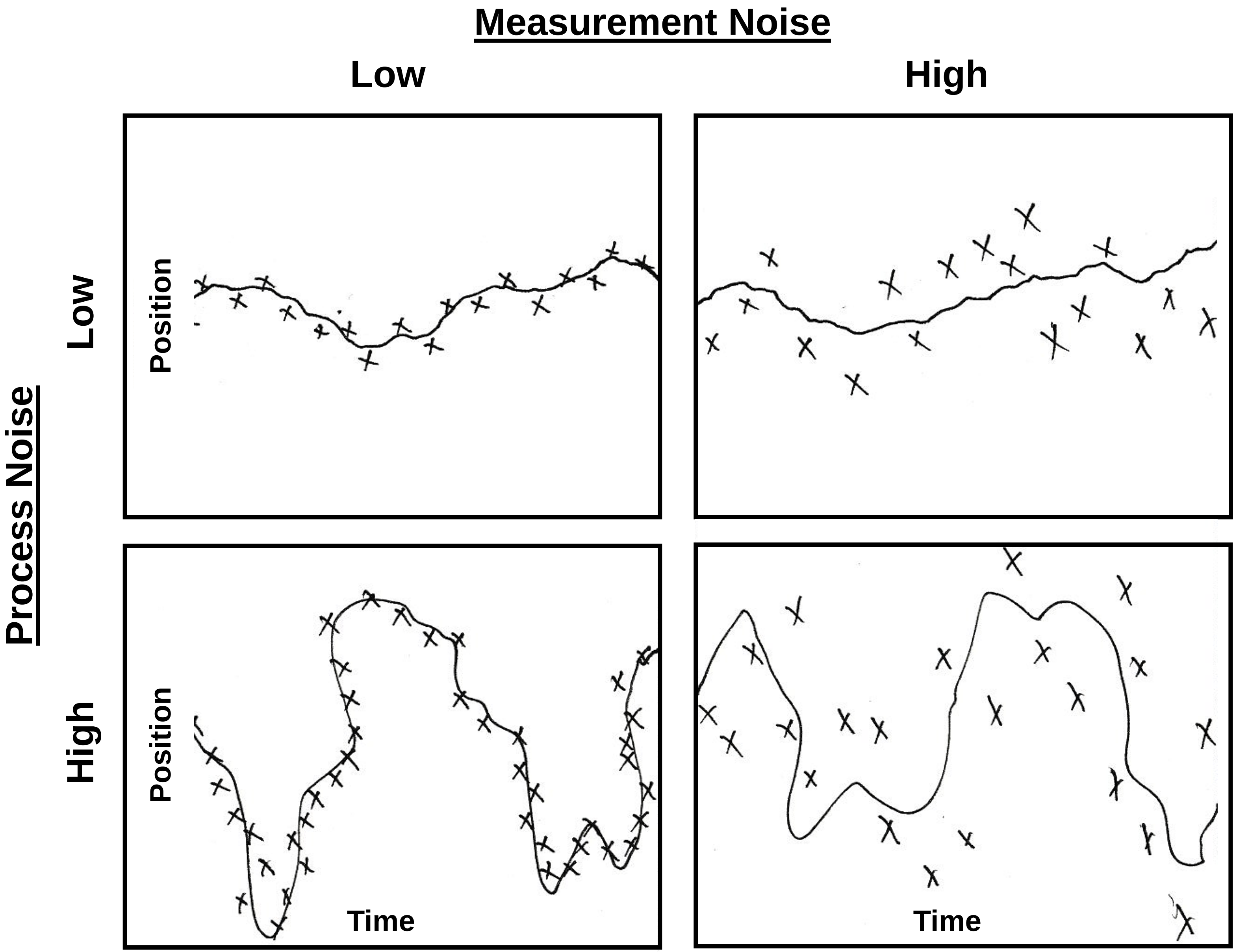
$$x_t = Ax_{t-1} + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, Q)$$

Plant Equations

"White" Noise

$$y_t = Cx_t + \omega \quad \text{where} \quad \omega \sim \mathcal{N}(0, R)$$

Example: Linear Gaussian Dynamical System



Conditional Independence (Directed)

Not as simple as graph separation in directed graphs…

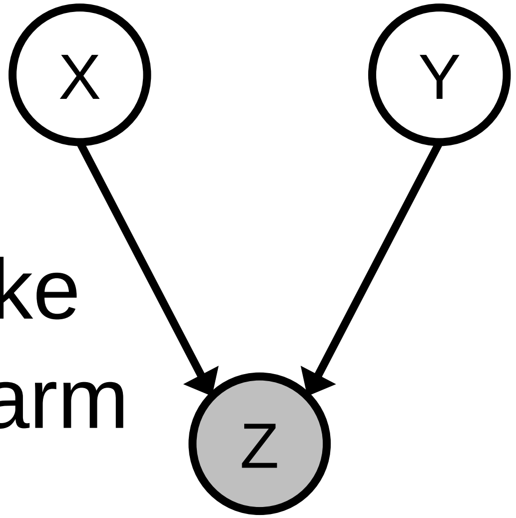
"Explaining Away Evidence"

<u>Binary RVs:</u>

X : Robber

Y : Earthquake

Z : House Alarm



$$X \not\perp\!\!\!\perp Y \mid Z$$

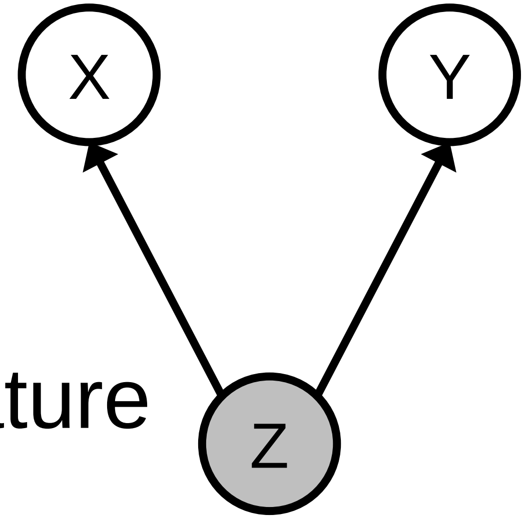
$$p(z)\boxed{p(x \mid z)p(y \mid z)}$$

<u>Binary RVs:</u>

X : Heat On

Y : A/C On

Z : Temperature

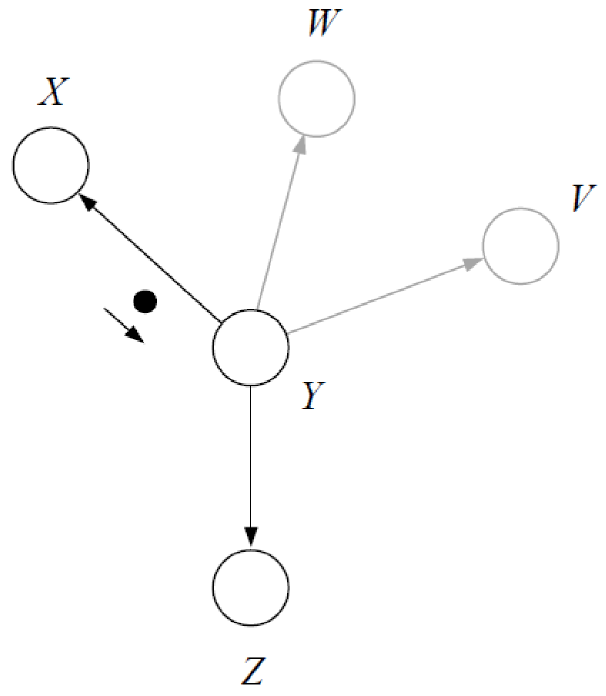


$$X \perp\!\!\!\perp Y \mid Z$$

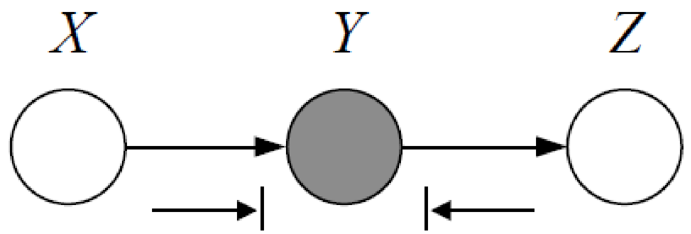
Directed separation (d-separation) property indicates conditional independence in directed models.

Bayes Ball Algorithm

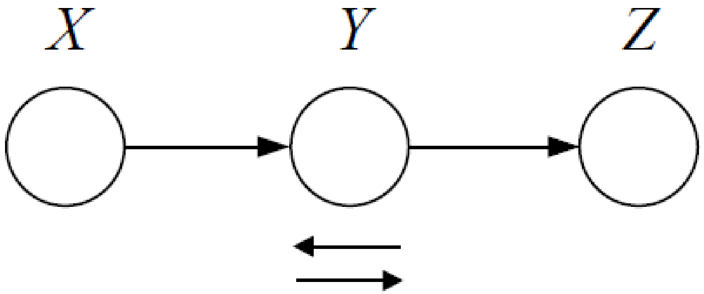
To test if $x_A \perp\!\!\!\perp x_C \mid x_B$ imagine rolling a "ball" from each node in x_A. The "ball" follows certain rules defined by canonical 3-node subgraphs:



Incoming & outgoing edges



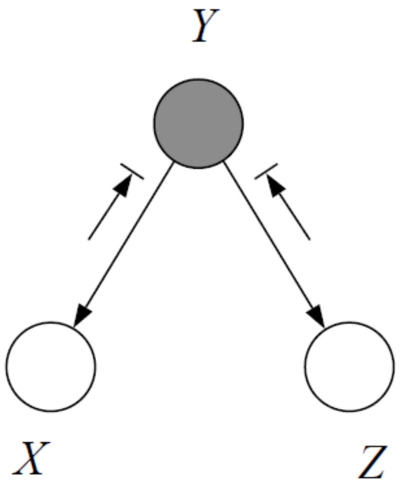
Y blocks the Bayes ball, acting as a d-separator.



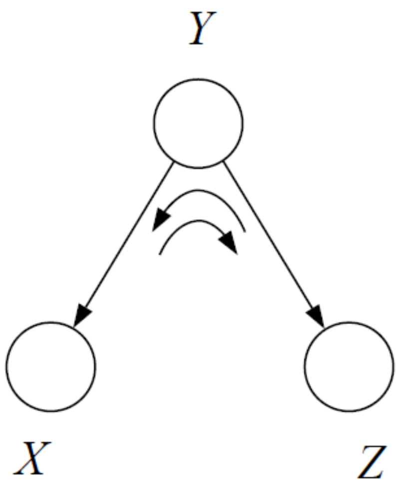
Y does not block. It is not a d-separator.

[Source: Michael I Jordan]

Two Outgoing Arrows

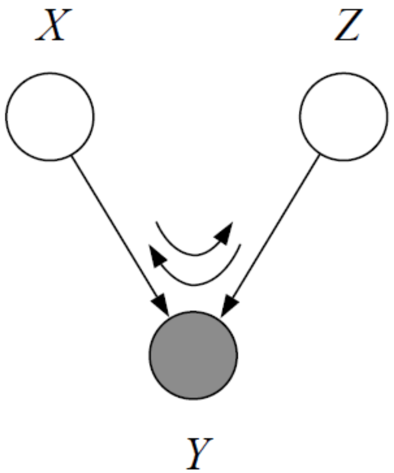


Y blocks

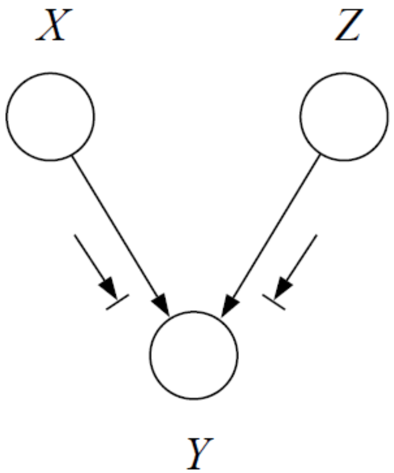


Y does not block

Two Incoming Arrows



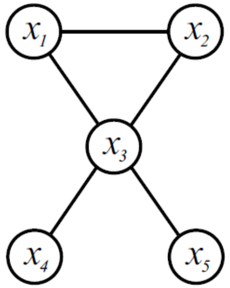
Y blocks

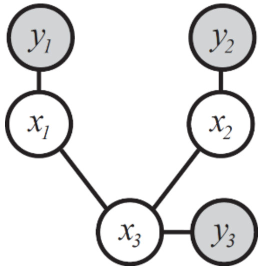


Y does not block

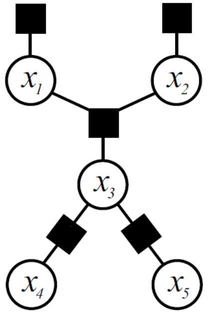
If a set x_B blocks for every node in x_C then $x_A \perp\!\!\!\perp x_C \mid x_B$. Conversely, if a ball reaches *any* node in x_C then they are **not** conditionally independent.

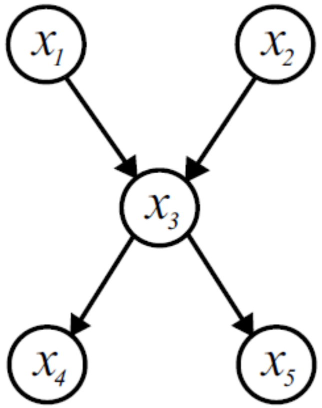
Summary



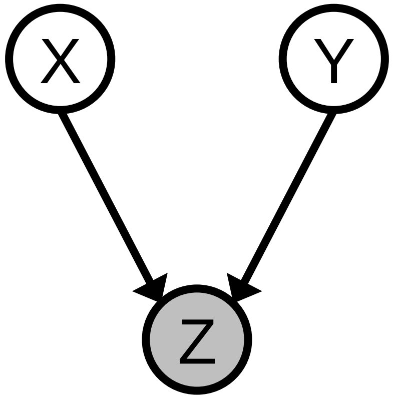


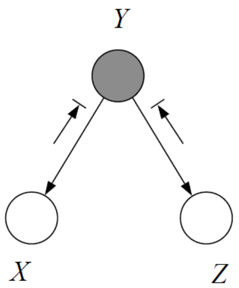
Undirected models may be specified up to normalization. Factorization may not be unique for MRFs.

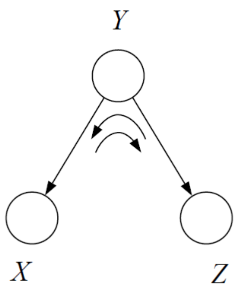




Directed models useful for product of locally-normalized conditional probabilities. Simplifies simulation via ancestral sampling. Conditional independence more difficult.







Conditional independence given by graph separation and d-separation for undirected / directed models.