EECS 545: Machine Learning

Lecture 14. Markov Networks

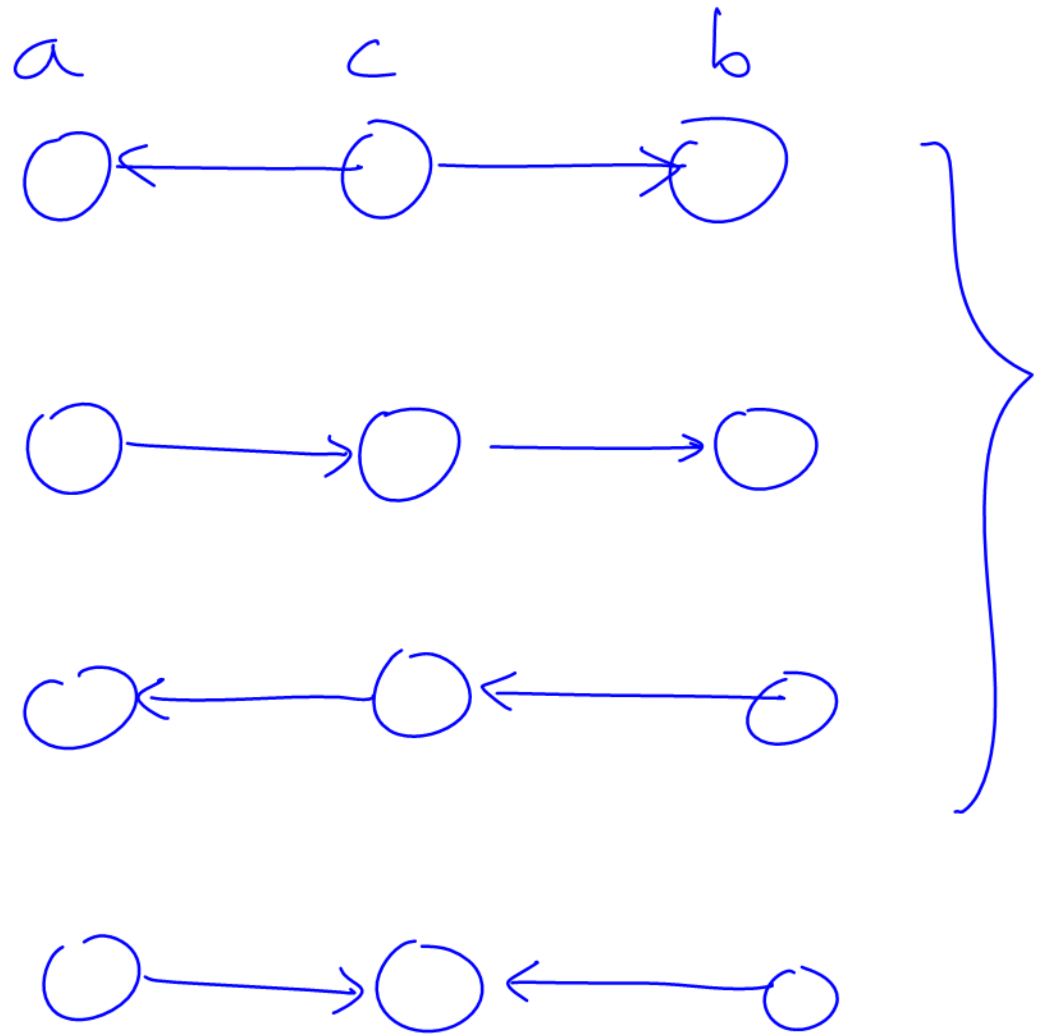
Honglak Lee

2/23/2011

Outline

- Bayesian Networks (cont'd)
  - D separation
  - Markov Blanket
- Markov Networks
  - (aka Markov Random Fields, Undirected graphical models)
  - Representation
  - Conditional Independence
  - Examples
- Directed vs Undirected graphical models

D-Separation & Markov Blankets



$a \perp\!\!\!\perp b \mid c$

$a \not\perp\!\!\!\perp b \mid \emptyset$

$a \perp\!\!\!\perp b \mid \emptyset$

$a \not\perp\!\!\!\perp b \mid c$

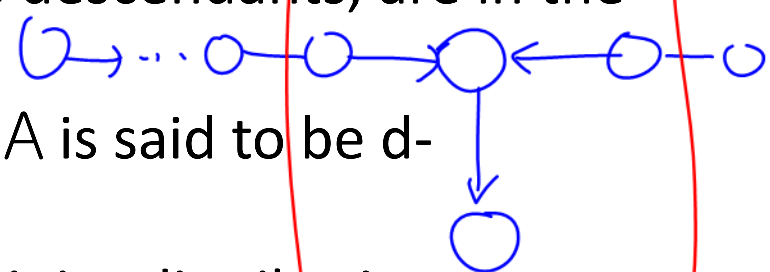
"head-to-head" or "v-structure"

$A \perp\!\!\!\perp B \mid C$?

D-separation

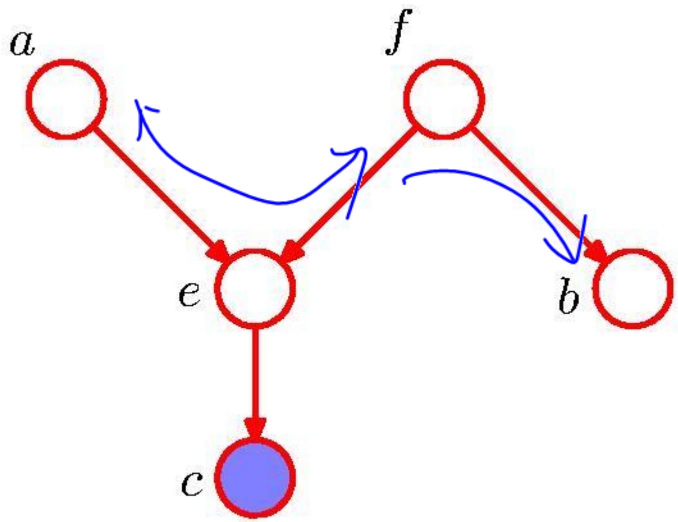
- A, B, and C are non-intersecting subsets of nodes in a directed graph.
- A path from A to B is blocked if it contains a node such that either

  a) the arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set C, or
  b) the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set C.

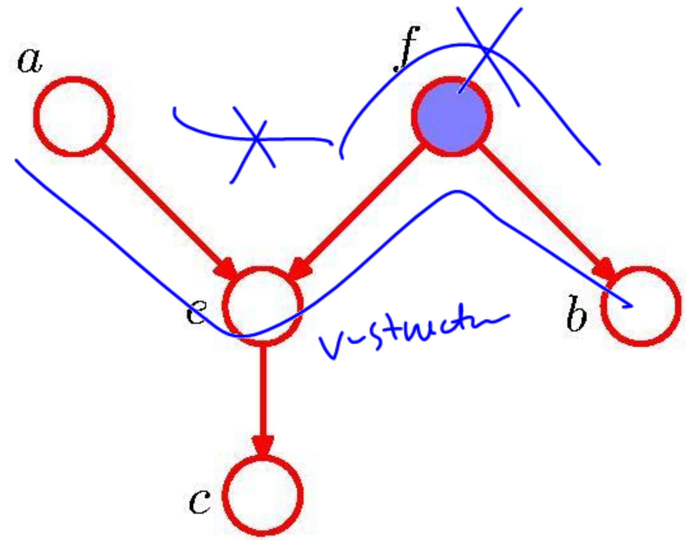


- If all paths from A to B are blocked, A is said to be d-separated from B by C.
- If A is d-separated from B by C, the joint distribution over all variables in the graph satisfies $A \perp\!\!\!\perp B \mid C$.

D-separation: Example



$$\boxed{a \not\perp\!\!\!\perp b \mid c}$$

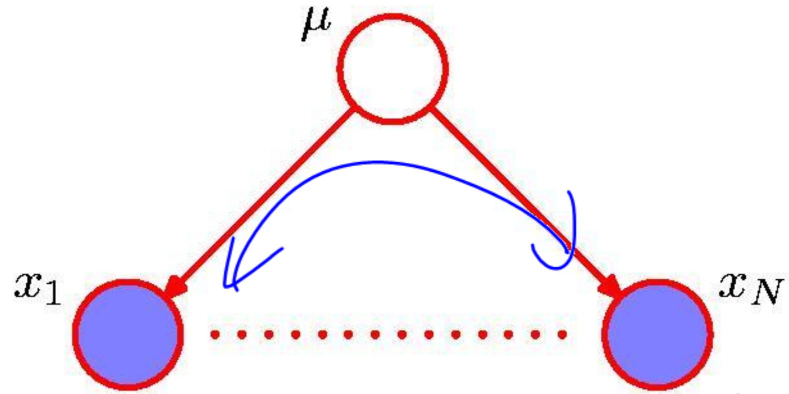


$$a \perp\!\!\!\perp b \mid f$$

D-separation: I.I.D. Data

Naive Bayes

$X_i \perp\!\!\!\perp X_j \mid \mu$



class label

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu)$$

NB assumption

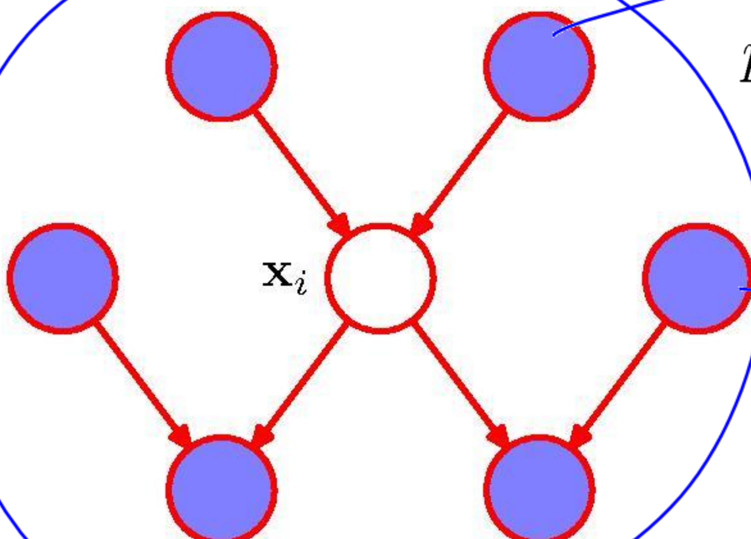
Bayesian NB

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D}|\mu)p(\mu)\,\mathrm{d}\mu \neq \prod_{n=1}^{N} p(x_n)$$

The Markov Blanket

$\mathcal{X}$: all variables

A set S is called a Markov Blanket of X_i iff
X_i is conditionally independent of all other variables given S.

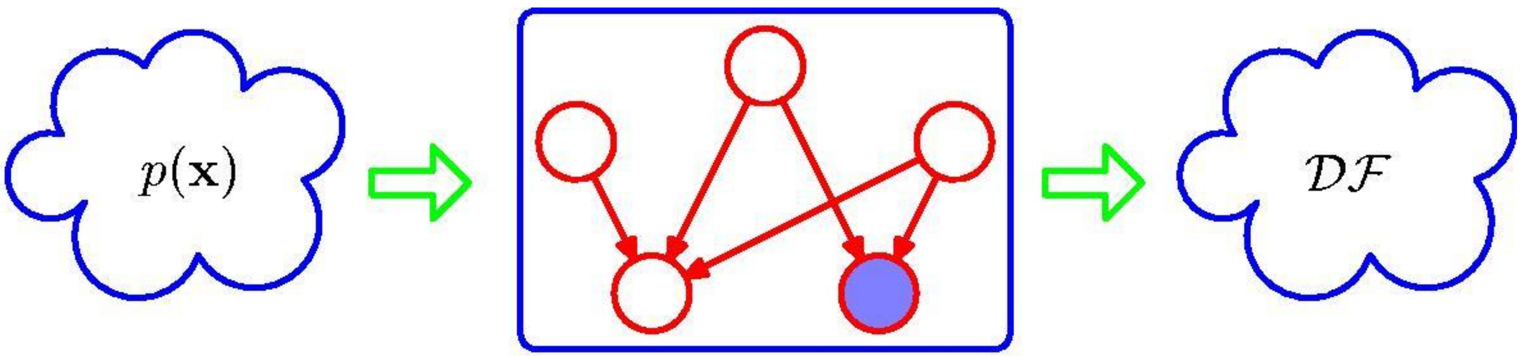


$$p(\mathbf{x}_i|\mathbf{x}_{\{j\neq i\}}) = \frac{p(\mathbf{x}_1,\ldots,\mathbf{x}_M)}{\int p(\mathbf{x}_1,\ldots,\mathbf{x}_M)\,\mathrm{d}\mathbf{x}_i} = \frac{\prod_k p(\mathbf{x}_k|\mathrm{pa}_k)}{\int \prod_k p(\mathbf{x}_k|\mathrm{pa}_k)\,\mathrm{d}\mathbf{x}_i}$$

$X_i \perp\!\!\!\perp \mathcal{X} \setminus \{X_i\} \cup MB \mid MB$

In <u>Bayesian Networks</u>, a Markov Blanket of Xi
= $\mathrm{Pa}_{X_i} \cup \mathit{Children}_{X_i} \cup \mathit{Coparents}$

Directed Graphs as Distribution Filters

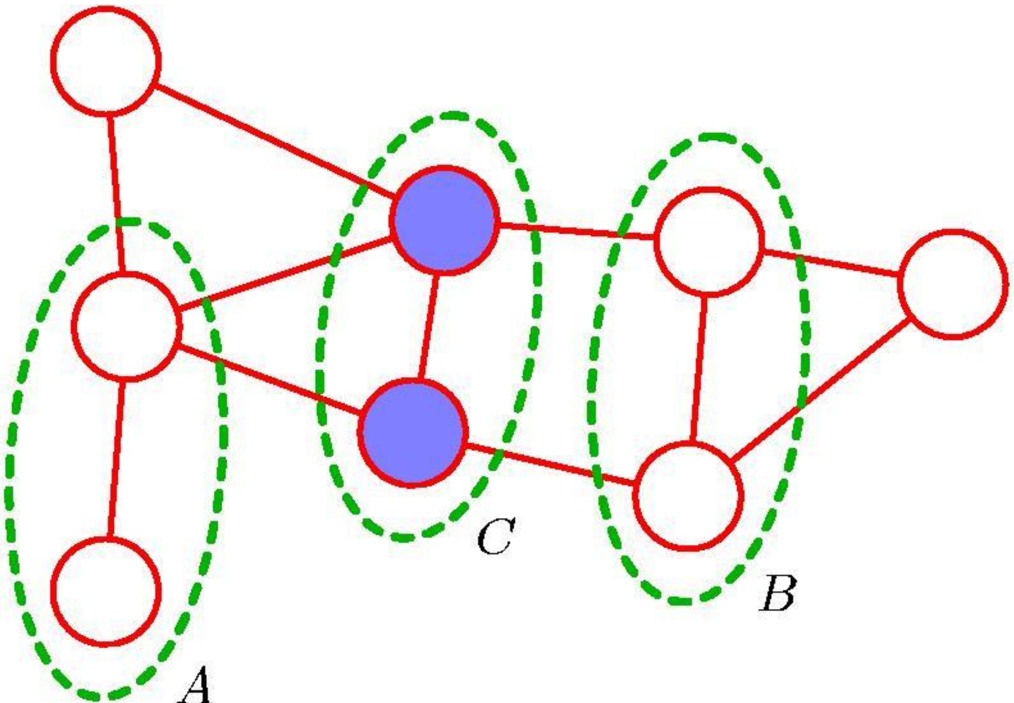


A set of distribution that satisfies a set of conditional independence represented by directed graphs (via d-separation) <u>is equivalent to</u>
a set of distribution that is represented by Bayesian Network joint-probability factorization.

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

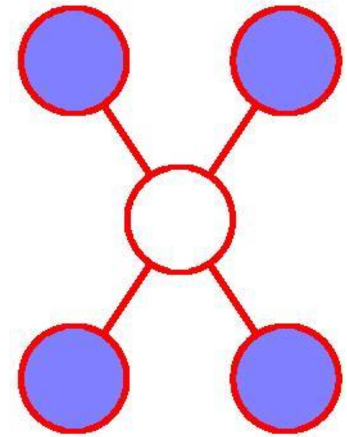
Markov Networks

Markov Networks (Markov Random Fields)



$$A \perp\!\!\!\perp B \mid C$$

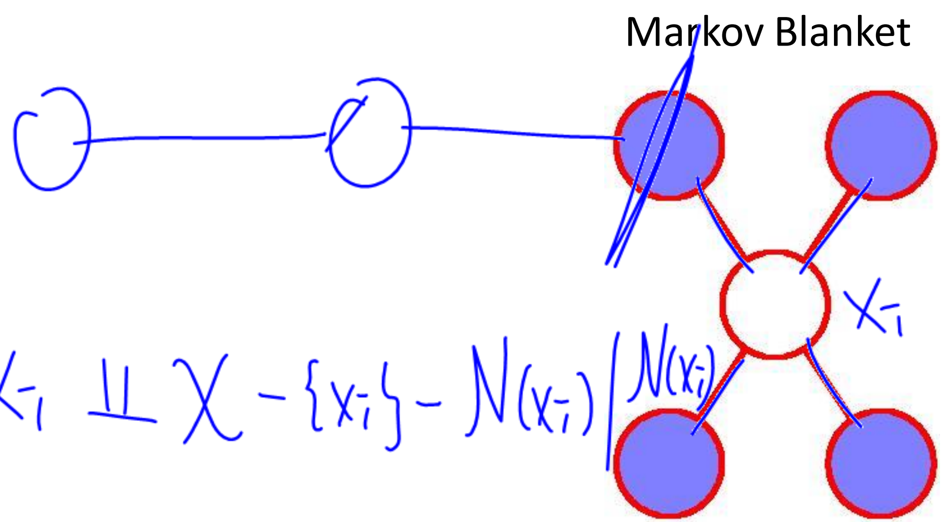
Markov Blanket



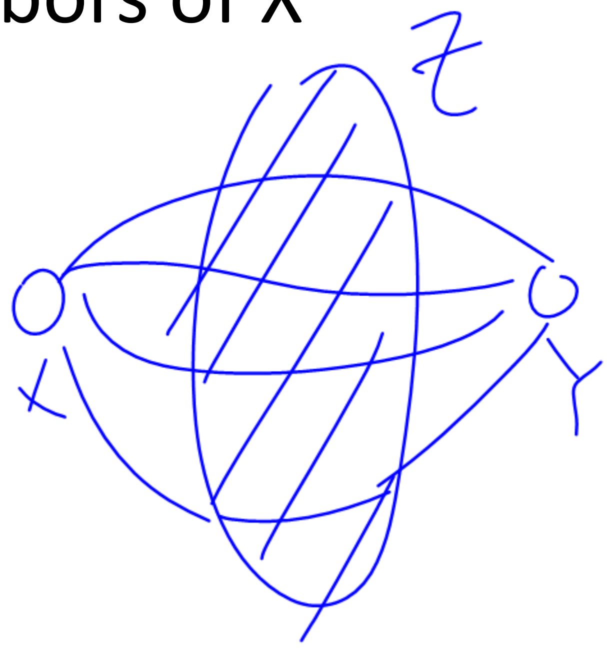
Markov networks are represented using undirected edges.
<u>Note</u>: there is no explicit definition of conditional probability in the network. (cf. CPD is the basis of Bayesian Network)

Conditional Independence in MN

- A path from X to Y is active if it is not blocked by observed variables.
- X and Y are conditionally independent given Z iff all path from X to Y are blocked by Z.
- Markov Blanket of X = neighbors of X

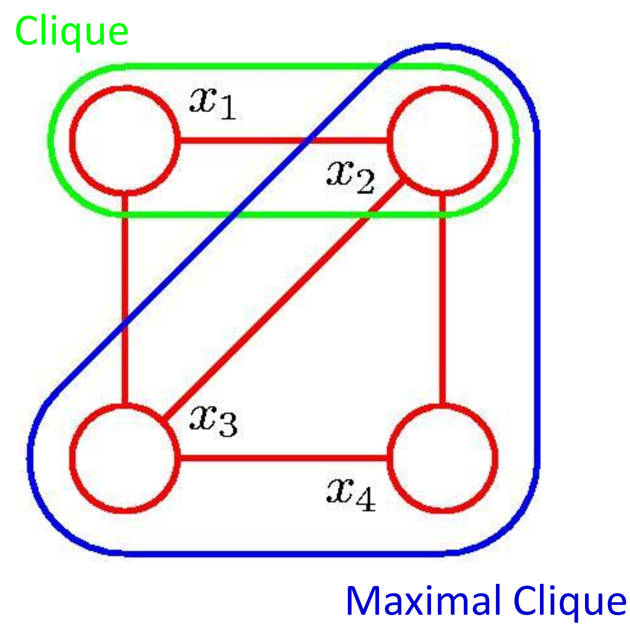


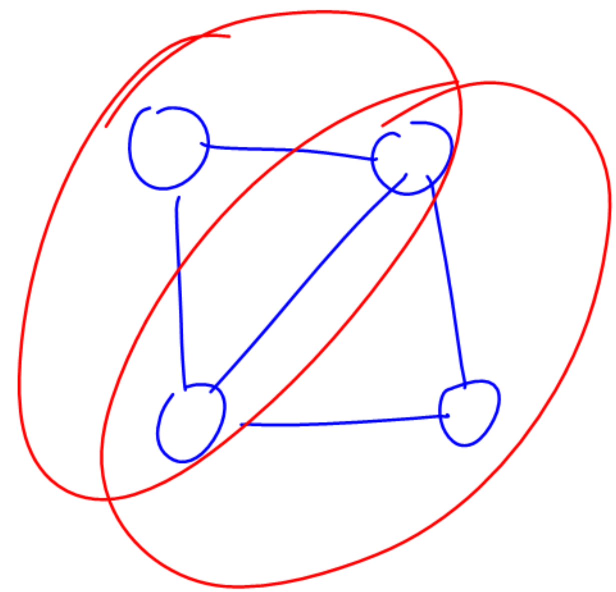
$X_i \perp X - \{X_i\} - N(X_i) \mid N(X_i)$



Cliques and Maximal Cliques

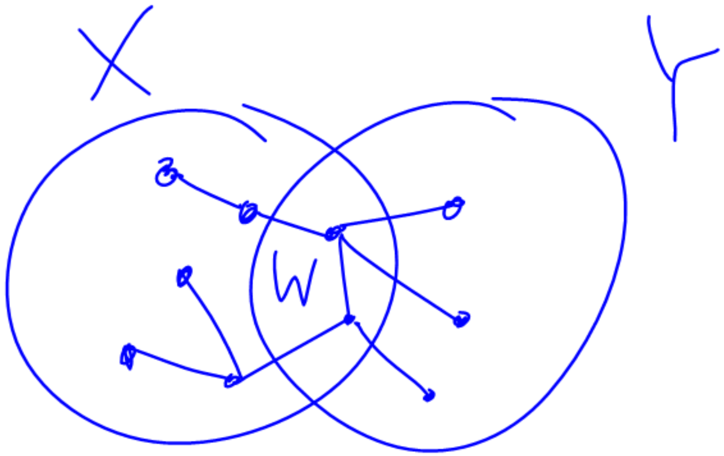
A *clique* is a subset C of S where every pair of elements of C are neighbors.





Any clique is subset of a maximal clique. Joint probability is defined as a product of nonnegative potential functions (for maximal cliques).

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C) \propto \psi_1(x_1, x_2, x_3)\, \psi_2(x_2, x_3, x_4)$$



X and Y are separated by W

⇒ any node in X and Y cannot belong to the same maximal clique

⇒ MN factorization $P(X,Y,W) = \frac{1}{Z}\prod_{c:X\cup W}\psi(x_c)\prod_{c:Y\cup W}\psi_c(x_c)$

$$P(X,Y,W) = \frac{1}{Z}\psi_1(X,W)\,\psi_2(Y,W)$$

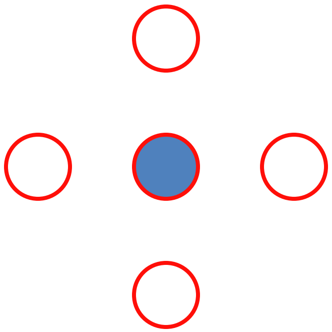
$$\Rightarrow\quad P(X,Y\mid W) = \frac{P(X,Y,W)}{P(W)} = \frac{\frac{1}{Z}\psi_1(X,W)\,\psi_2(Y,W)}{\frac{1}{Z}\sum_x\psi_1(x,W)\sum_y\psi_2(y,W)}$$

$$= P(X\mid W)\,P(Y\mid W)$$

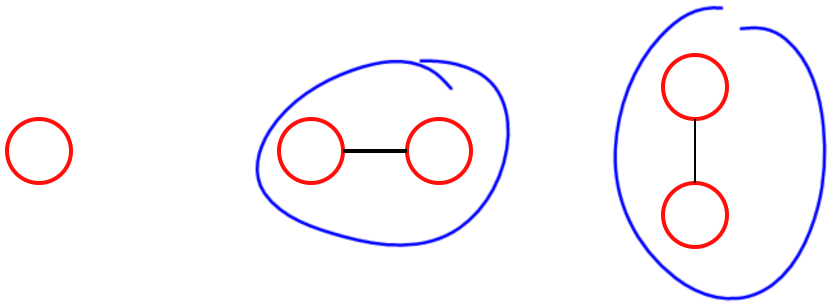
$$\Rightarrow\quad X \perp\!\!\!\perp Y \mid W \quad \square$$

Cliques (for c=1 neighborhoods)

- Sets where every pair are neighbors.



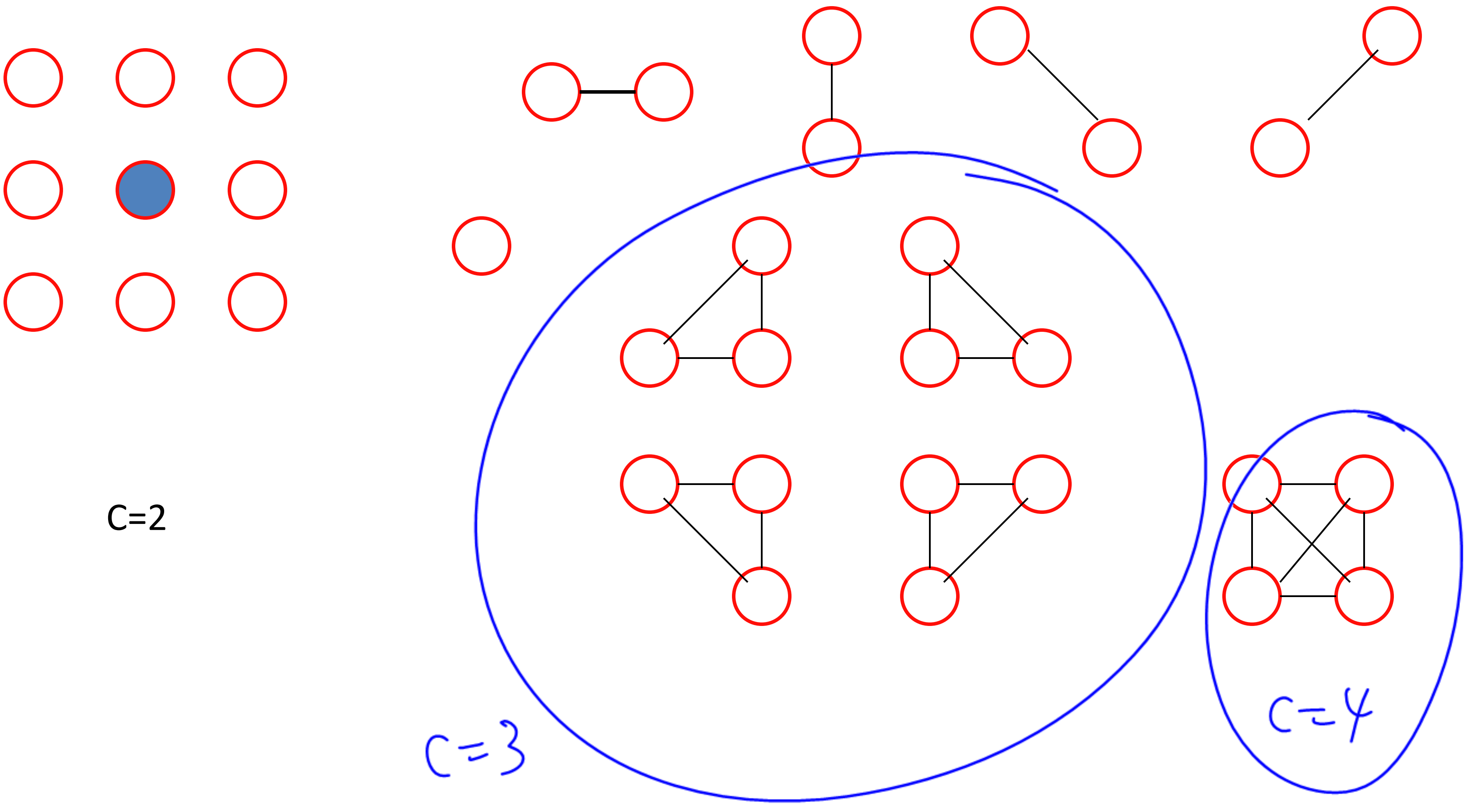
C=1



c=2

Cliques (for c=2 neighborhoods)

- Maximal clique: if enlarged, it's not a clique.



Joint Distribution

$$p(\mathbf{x}) = \frac{1}{Z}\prod_C \psi_C(\mathbf{x}_C)$$

where $\psi_C(\mathbf{x}_C)$ is the potential over clique C and

$$Z = \sum_{\mathbf{x}}\prod_C \psi_C(\mathbf{x}_C)$$

$\sum_x P(x) = 1$

is the normalization coefficient; note: M K-state variables → K^M terms in Z.

Energies and the Boltzmann distribution

$$\psi_C(\mathbf{x}_C) = \exp\left\{-E(\mathbf{x}_C)\right\}$$

Examples

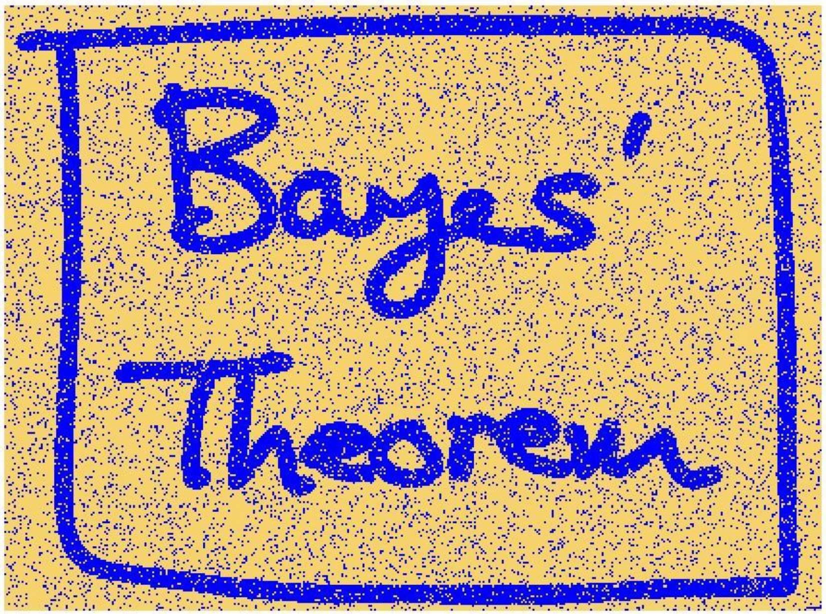
Fields of Random Variables

- In a Bayesian network, nodes represent the probabilities of propositions.
  - Layout of the nodes is arbitrary.

- For problems in image analysis, physics, etc, spatial structure is important.
  - Many interactions are *local*, so we need to define the *neighborhood* structure on the set (field) of random variables.

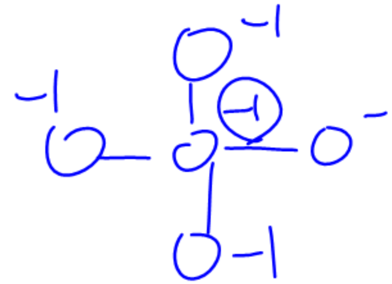
Illustration: Image De-Noising (1)

Original Image

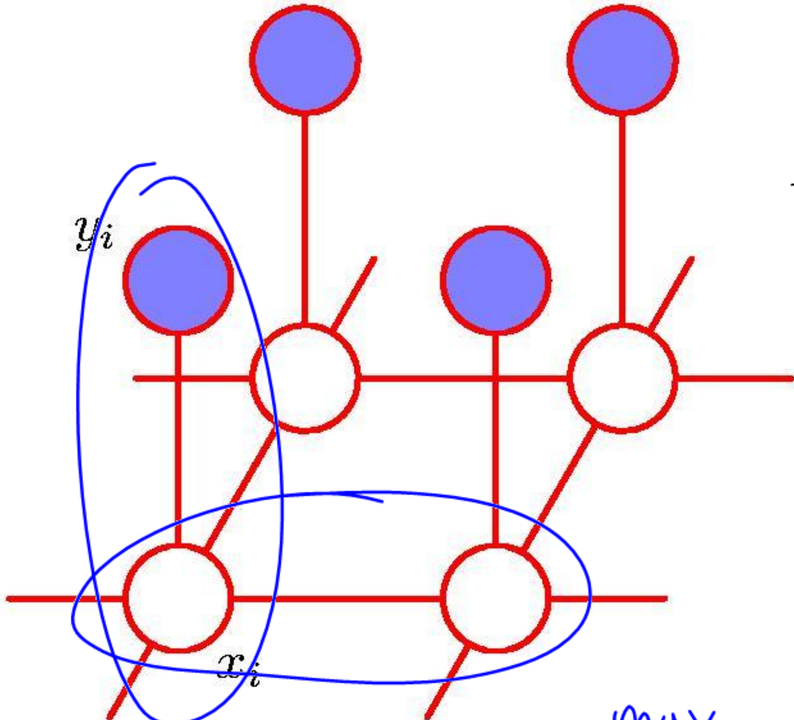


Noisy Image

Illustration: Image De-Noising (2)



$X_i = X_j \Leftrightarrow X_i X_j = 1$



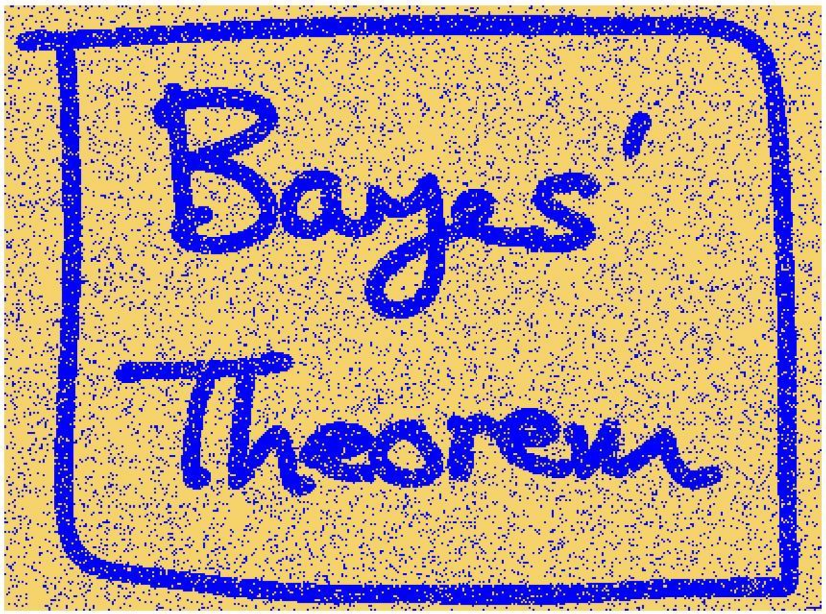
$$E(\mathbf{x},\mathbf{y}) = h\sum_i x_i - \beta \underline{\sum_{\{i,j\}} x_i x_j}$$
$$\underline{-\eta \sum_i x_i y_i}$$

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

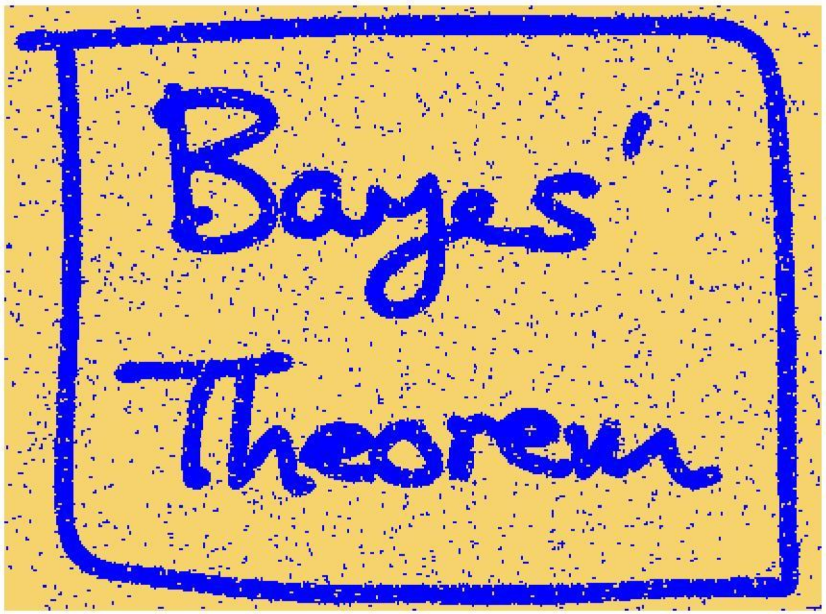
$\max_{X} P(X \mid Y)$

Y: noisy binary image (observed), $Y_i \in \{1, -1\}$
X: underlying binary image (to infer), $X_i \in \{1, -1\}$

Illustration: Image De-Noising (3)

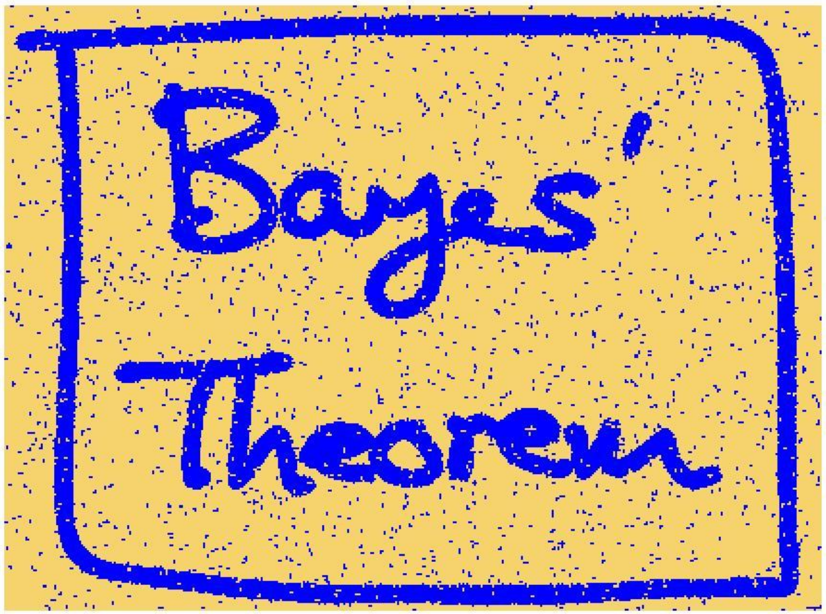


Noisy Image

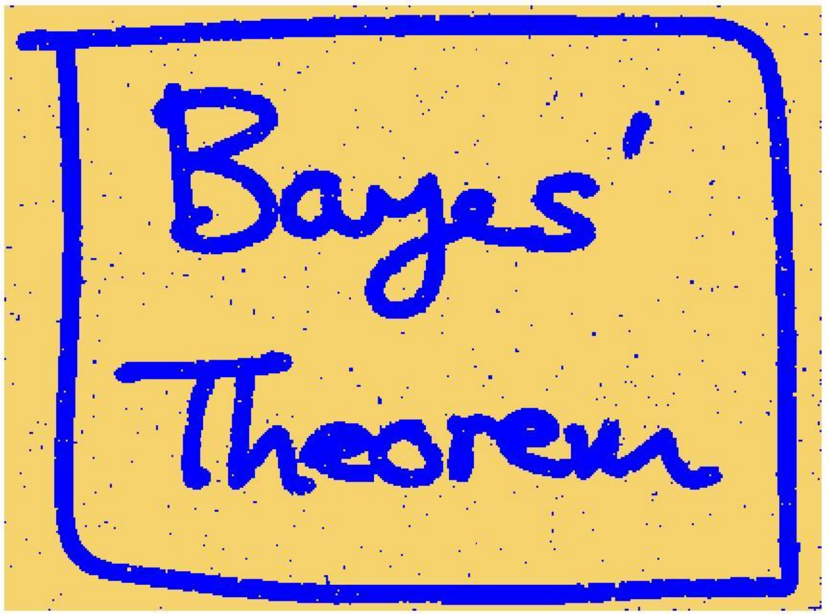


Restored Image (ICM)

Illustration: Image De-Noising (4)



Restored Image (ICM)

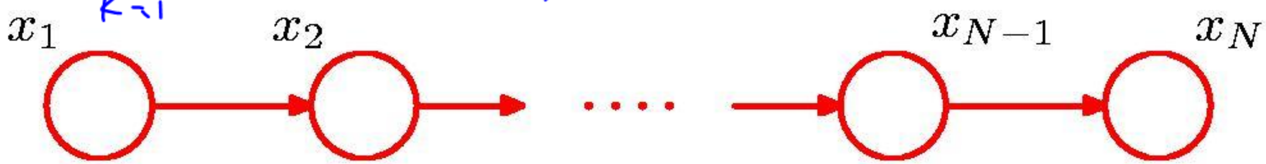


Restored Image (Graph cuts)

Directed vs. Undirected Graphs

Converting Directed to Undirected Graphs (1)

$P(x) = \prod_{k=1}^{N} P(x_k \mid Pa\, x_k)$



$$p(\mathbf{x}) = \underbrace{p(x_1)p(x_2|x_1)}\, p(x_3|x_2) \cdots p(x_N|x_{N-1})$$

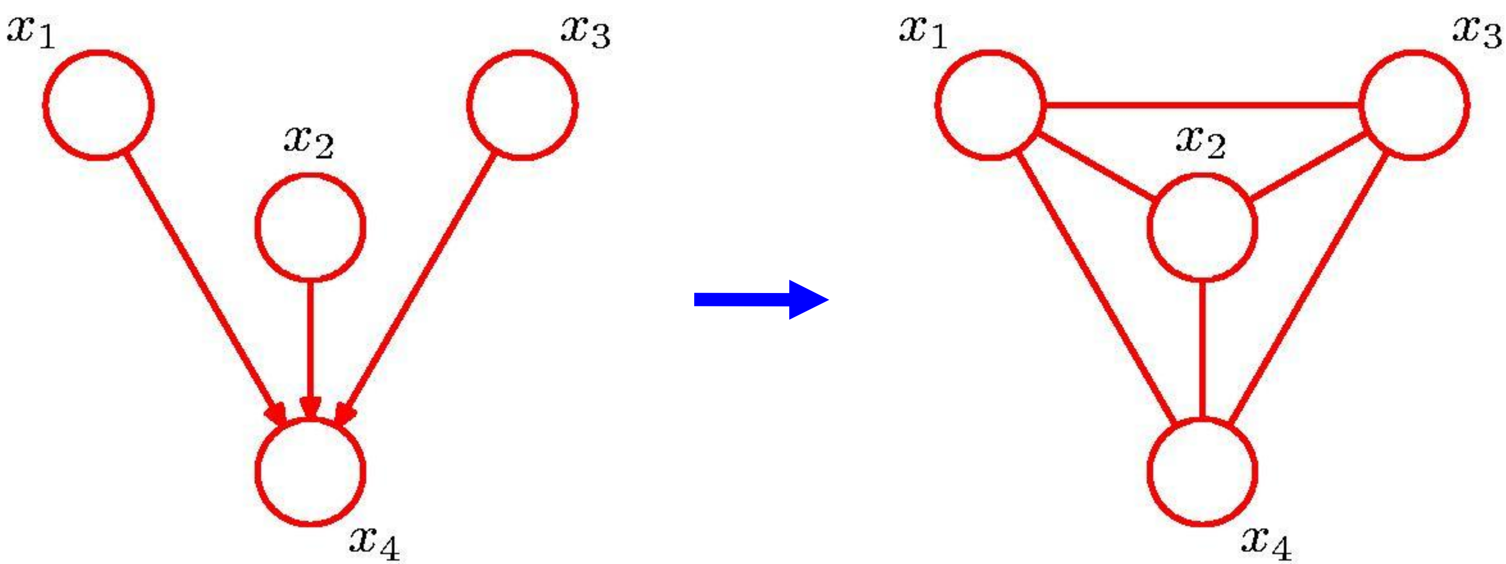
$P(x) = \frac{1}{Z} \prod_c \psi_c(x_c)$

$$p(\mathbf{x}) = \frac{1}{Z} \psi_{1,2}(x_1, x_2)\, \psi_{2,3}(x_2, x_3) \cdots \psi_{N-1,N}(x_{N-1}, x_N)$$

Converting Directed to Undirected Graphs (2)

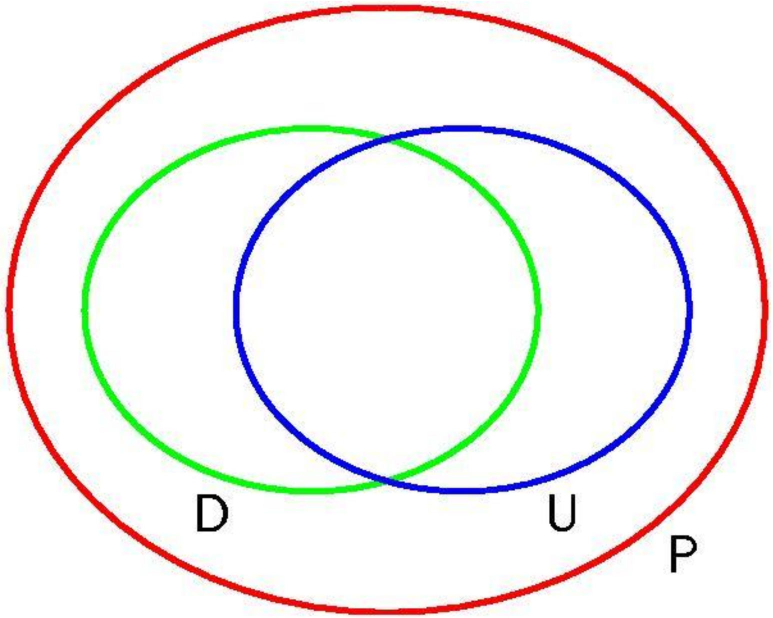
- Additional links

Moralizing: “Moral Graph”



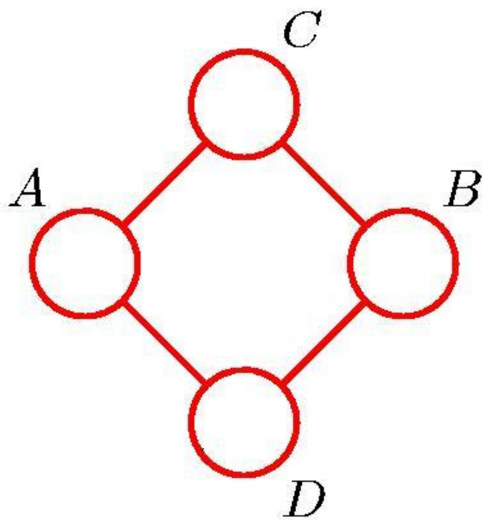
$$
\begin{aligned}
p(\mathbf{x}) &= p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3) \\
&= \frac{1}{Z}\psi_A(x_1,x_2,x_3)\psi_B(x_2,x_3,x_4)\psi_C(x_1,x_2,x_4)
\end{aligned}
$$

Directed vs. Undirected Graphs (1)



Directed vs. Undirected Graphs (2)

E.g., Markov Network, but cannot be represented by Bayesian Network



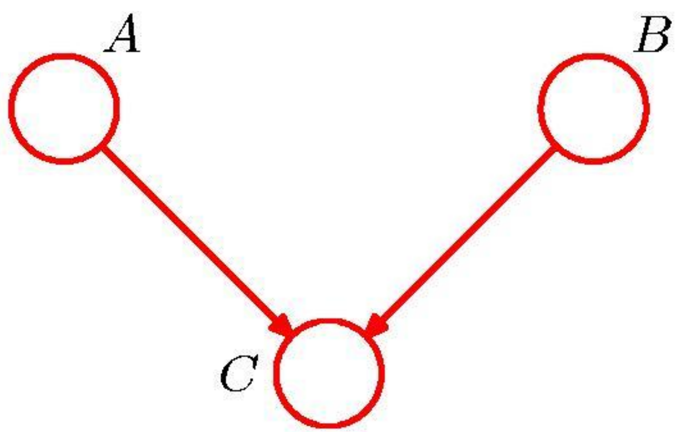
$$A \not\perp\!\!\!\perp B \mid \emptyset$$

$$A \perp\!\!\!\perp B \mid C \cup D$$

$$C \perp\!\!\!\perp D \mid A \cup B$$

Directed vs. Undirected Graphs (2)

E.g., Bayesian Network, but cannot be represented by Markov Network



$$A \perp\!\!\!\perp B \mid \emptyset$$

$$A \not\!\perp\!\!\!\perp B \mid C$$

Next

- Inference in graphical models
- Mixture models and the EM algorithm