Ch 8. Graphical Models

Pattern Recognition and Machine Learning,
C. M. Bishop, 2006.

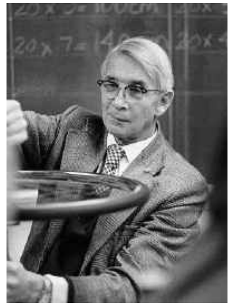
Summarized by
B.-H. Kim
Biointelligence Laboratory, Seoul National University
http://bi.snu.ac.kr/

Contents

- **8.3 Markov Random Field**
  - 8.3.1 Conditional independence properties
  - 8.3.2 Factorization properties
  - 8.3.3 Illustration: Image de-noising
  - 8.3.4 Relation to directed graphs
- **8.4 Inference in Graphical Models**
  - 8.4.3 Factor graphs

Markov Random Fields

Andrei Andreyevich Markov (1856 – 1922)

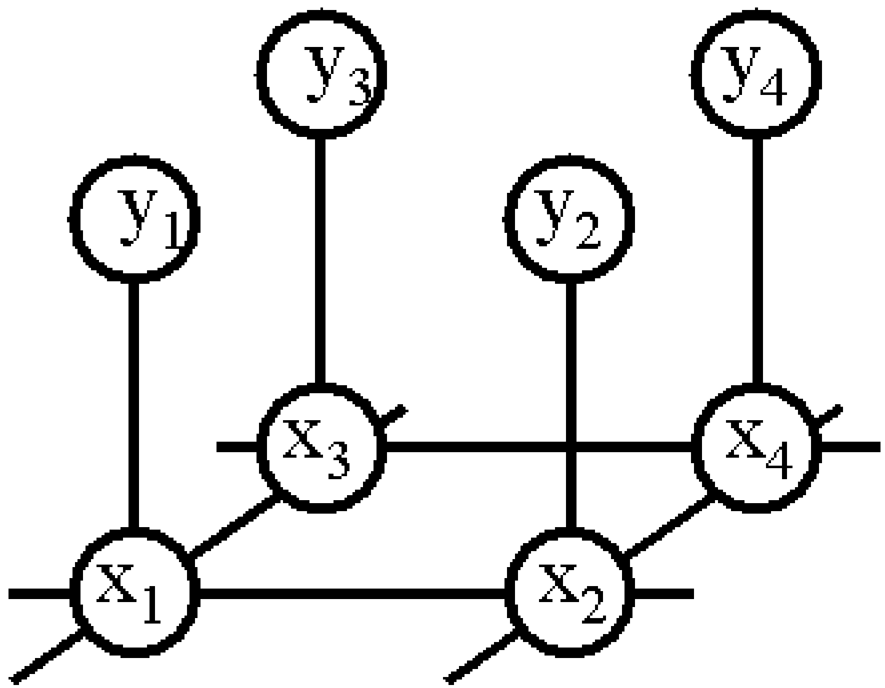


Ernst Ising (1900–1998)

- Various names
  - *Markov random field (MRF), Markov network, undirected graphical model*
- A set of random variables have a Markov property described by un undirected graph
- *Markov random field* was introduced as the general setting for the *Ising model*, which was originally motivated as the model for ferromagnetism
  - Formally, Markov random field is *n-dimensional random process defined on a discrete lattice*

Markov Random Fields
as Probability models for entire images

- Allows rich probabilistic models for images.
- But built in a local, modular way. Learn local relationships, get global effects out.



MRF nodes as pixels

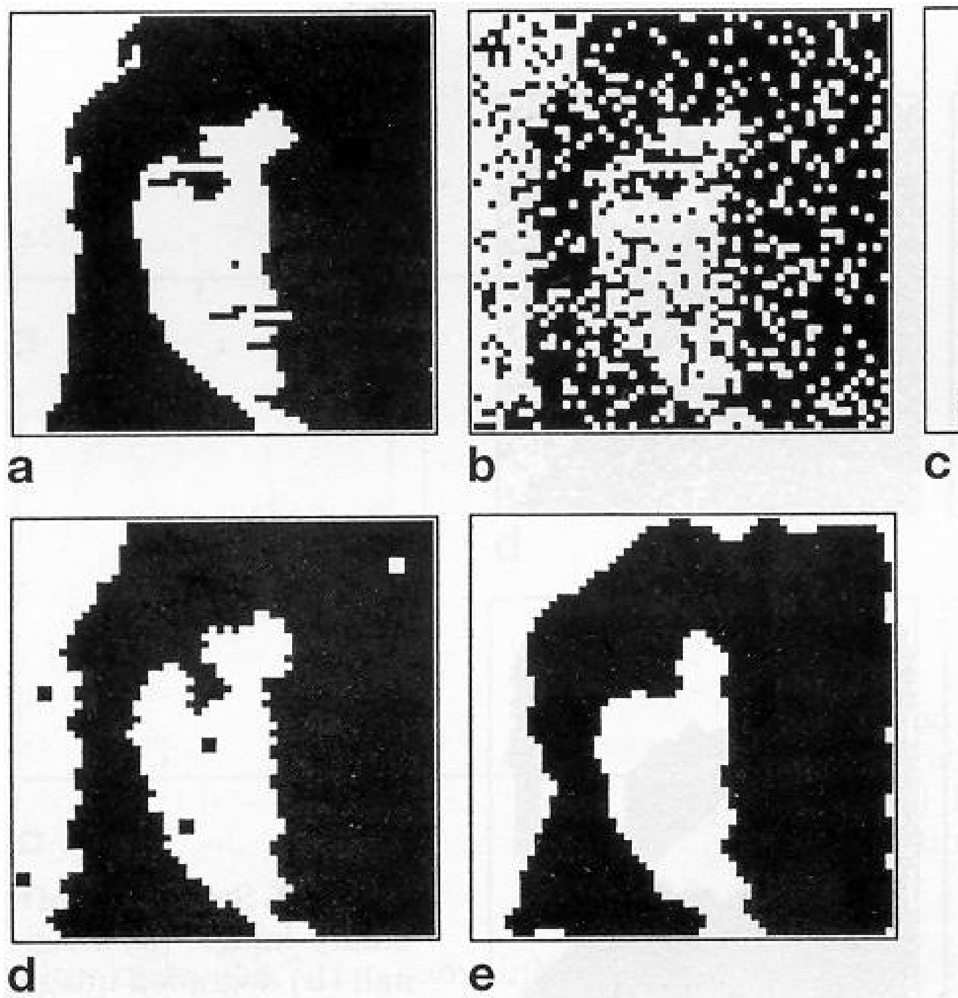
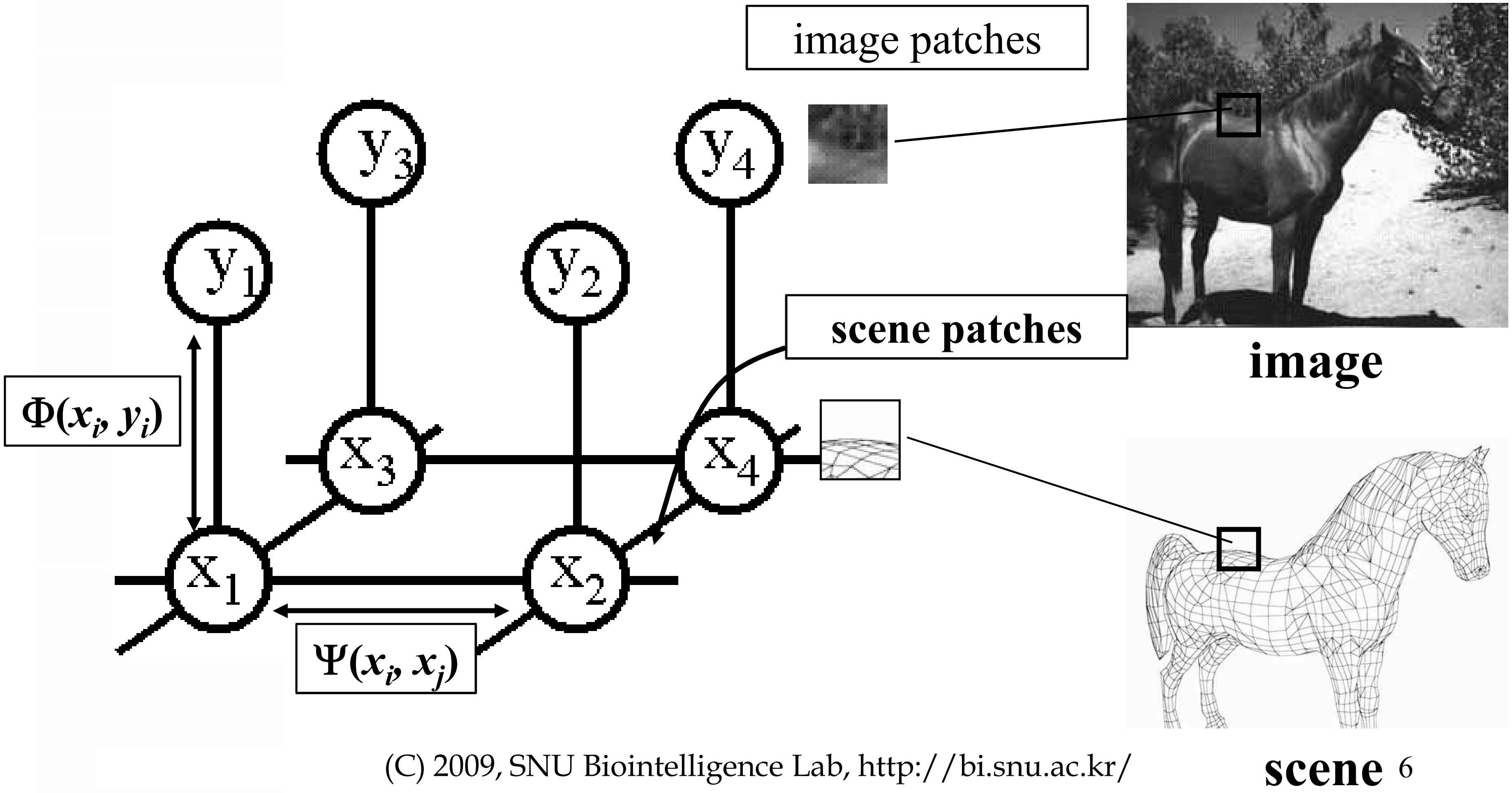


Fig. 2.3. Smoothing with the wrong prior. (**a**) Original, (**b**) degraded image, (**c**) MAP estimate $\beta = 1$, (**d**) MAP estimate $\beta = 0.3$, (**e**) median filter

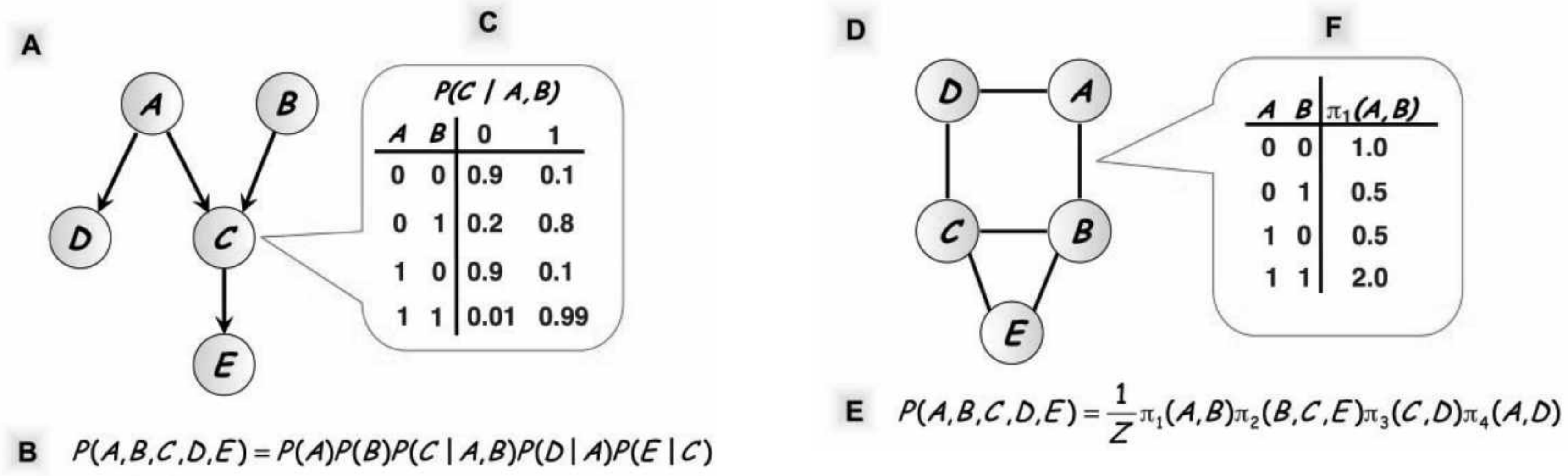
Winkler, 1995, p. 32

MRF nodes as patches



Directed graph vs. undirected graph

- Both graphical models
  - ♦ Specify a **factorization** (how to express the joint distribution)
  - ♦ Define a set of **conditional independence** properties



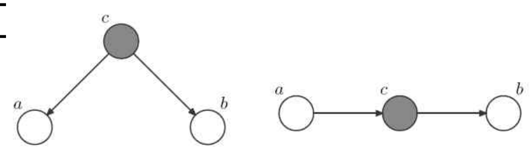
Parent - child
Local conditional distribution

Maximal clique
Potential function

• **Chain graphs: graphs that include both directed and undirected links**

8.3.1 Conditional independence properties

- In directed graphs

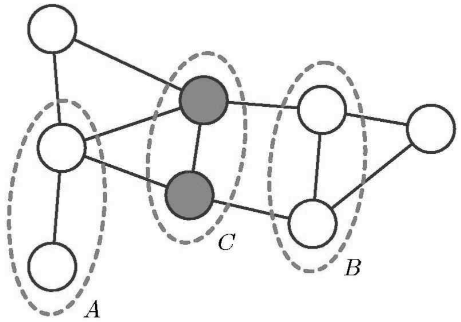


Shaded circle: evidence, i.e. observed variables

  - ‘D-separation’ test: if the paths connecting two sets of nodes are ‘blocked’
  - Subtle case: ‘head-to-head’ nodes

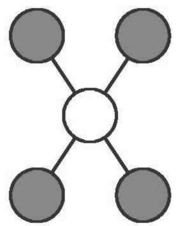
- In undirected graphs

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



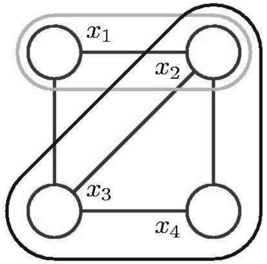
  - Simple graph separation (simpler than in directed graphs)
  - Checking all the paths btw A and B
    - if all the paths are blocked by C or not
    - After removing C, if there is any path remaining

- Markov blanket for an undirected graph



8.3.2 Factorization properties

- A maximal clique
  - ♦ Clique: a subset of the nodes in a graph s.t. there exists a link btw all pairs of nodes in the subset



- Functions of the maximal cliques become the factors in the decomposition of the joint distribution
  - ♦ Potential function $\psi_C(\mathbf{x}_C)$

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C). \qquad Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C)$$ Partition function (normalization constant)

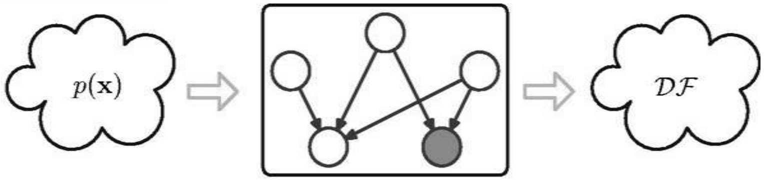
• Potential functions are not restricted to marginal or conditional distributions

• Normalization constant: major limitation of undirected graph.
But we can overcome when we focus on local conditional distribution

8.3.2 Factorization properties

- Considering formal connection btw **conditional independence** and **factorization**
  - Restriction: $\psi_C(\mathbf{x}_C)$ should be strictly positive

- Hammersley-Clifford theorem



(a graphical model as a filter)

$\mathcal{UI}$ Set of distributions that are consistent with the set of conditional independence from the graph

$\mathcal{UF}$ Set of distributions that can be expressed as a factorization of the form (8.39)

$\mathcal{UI}$ and $\mathcal{UF}$ are identical

- Expressing potential functions in exponential form

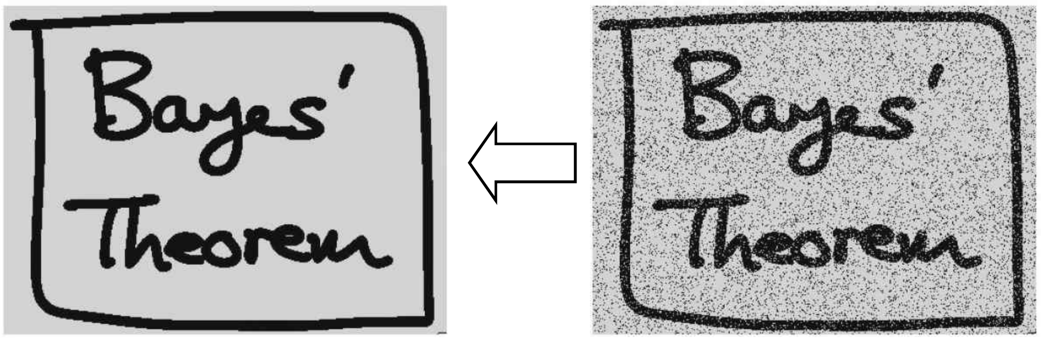
$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}$$

$E(\mathbf{x}_C)$: energy function

Boltzmann distribution

8.3.3 Illustration: Image de-noising (1)

- Setting
  - Image as a set of 'binary pixel values' {-1, +1}
  - In the observed noisy image $y_i \in \{-1, +1\}$
  - In the unknown noise-free image $x_i \in \{-1, +1\}$
  - Noise: randomly flipping the sign of pixels with some small probability
- Goal: to recover the original noise-free image

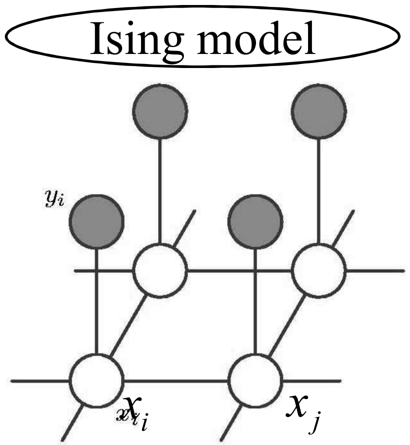


(Original image) (noisy-image: 10% noise)

8.3.3 Illustration: Image de-noising (2)

- Prior knowledge (when the noise level is small)
  - Strong correlation between x_i and y_i
  - Strong correlation between neighboring pixels x_i and x_j

- Corresponding Markov random field
  - A simple energy function for the cliques
    - form $\{x_i, y_i\}$: $-\eta x_i y_i$
    - form $\{x_i, x_j\}$: $-\beta x_i x_j$
  - Bias (preference of one particular sign) : hx_i
  - The complete energy function for the model / joint distribution :

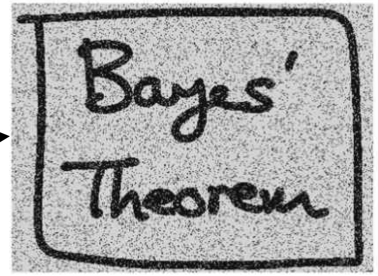


$$E(\mathbf{x},\mathbf{y}) = h\sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i \qquad p(\mathbf{x},\mathbf{y}) = \frac{1}{Z}\exp\{-E(\mathbf{x},\mathbf{y})\}.$$

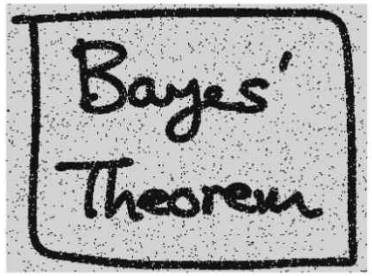
8.3.3 Illustration: Image de-noising (3)

- Image restoration results
  - Iterated conditional modes (ICM)
    - Coordinate-wise gradient ascent
    - Initialization: $x_i = y_i$ for all I
    - Take one node, evaluate the total energy, change the state of the node if it results in lower energy
    - Repeat till some stopping criterion is satisfied
  - Graph-cut algorithm
    - Guaranteed to find the global maximum in Ising model

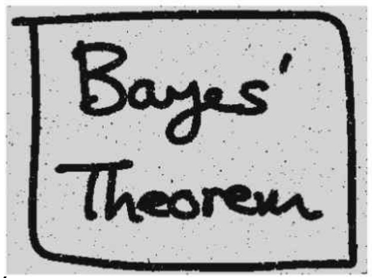
original



10% noise



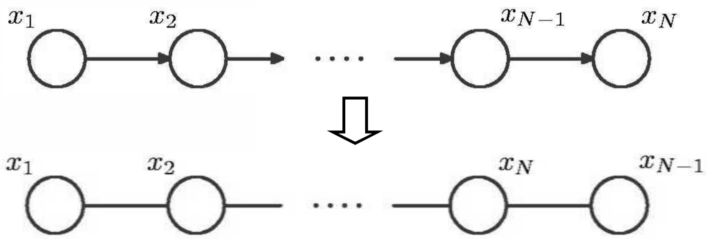
Restored by ICM



Restored by graph-cut

8.3.4 Relation to directed graphs (1)

- Converting a directed graph to un undirected graph
  - ♦ Case 1: straight line



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

$$\Downarrow$$

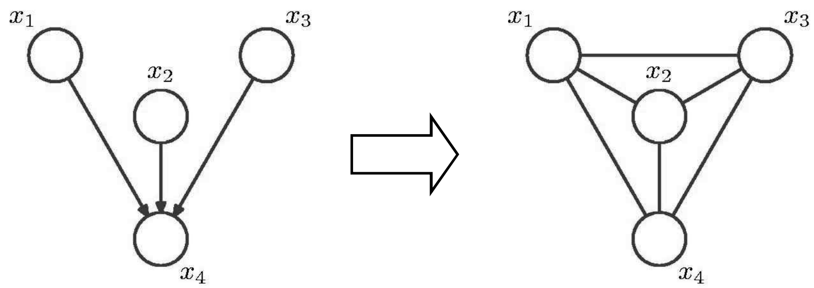
$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1,x_2)\psi_{2,3}(x_2,x_3)\cdots\psi_{N-1,N}(x_{N-1},x_N).$$

$$\begin{aligned}
\psi_{1,2}(x_1,x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2,x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(x_{N-1},x_N) &= p(x_N|x_{N-1})
\end{aligned}$$

  - ♦ In this case, the partition function Z=1

8.3.4 Relation to directed graphs (2)

- Converting a directed graph to un undirected graph
  - Case 2: general case. **Moralization,** 'marrying the parents'
    - Add additional undirected links btw all pairs of parents
    - Drop the arrows



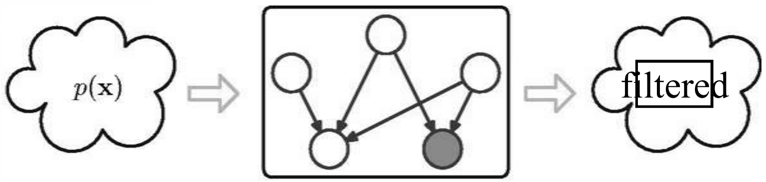
Usage example:
Exact inference algorithm
Ex) junction tree alg.

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3).$$

  - Result in the *moral graph*
    - Fully connected -> no conditional independence properties, in contrast to the original directed graph
  - We should add the fewest extra links to retain the maximum number of independence properties

8.3.4 Relation to directed graphs (3)

- Directed and undirected graphs can express different conditional independence properties



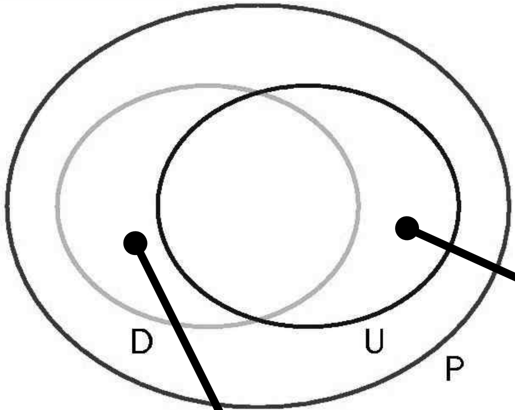
specific view:
graphical model as a filter (map)

⇨ ***D map*** Ex) completely disconnected graph is a trivial D map for any distribution

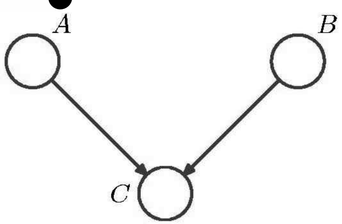
⇦ ***I map*** Ex) fully connected graph is a trivial I map for any distribution

⇔ ***Perfect map = both I&D map***

8.3.4 Relation to directed graphs (4)

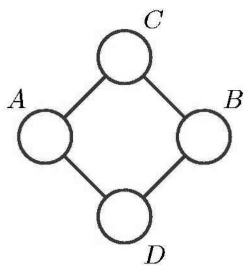


- D: the set of distributions that can be represented as a perfect map using a directed graph
- U: ~ using a undirected graph



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

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



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

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

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

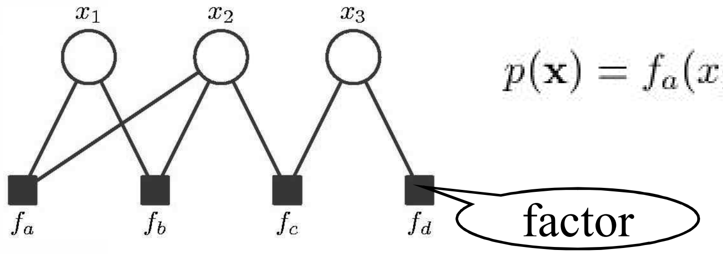
8.4.3 Factor graphs (1)

- Factor graphs
  - Introducing additional nodes for the factors themselves
  - Explicit decomposition /factorization
  - Joint distribution in the form of a product of factors $p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$
  - Factors in directed/undirected graphs

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

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

  - example

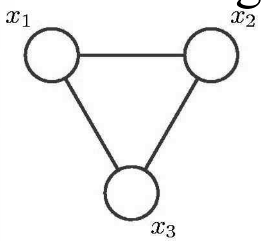


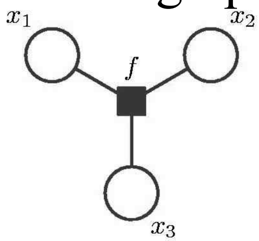
$$p(\mathbf{x}) = f_a(x_1,x_2)f_b(x_1,x_2)f_c(x_2,x_3)f_d(x_3)$$

(Factor graphs are *bipartite*)

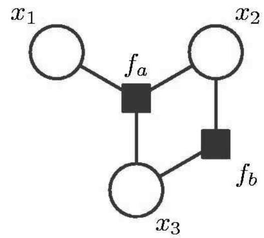
8.4.3 Factor graphs (2)

- Conversion
  - An undirected graph => factor graph



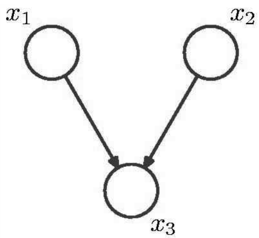


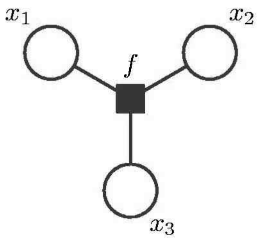
$$f(x_1, x_2, x_3) = \psi(x_1, x_2, x_3)$$



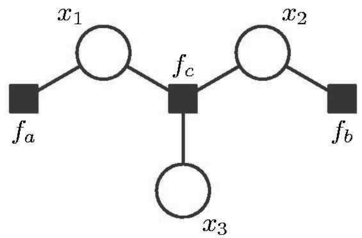
$$f_a(x_1, x_2, x_3) f_b(x_1, x_2) = \psi(x_1, x_2, x_3)$$

  - A directed graph => factor graph





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

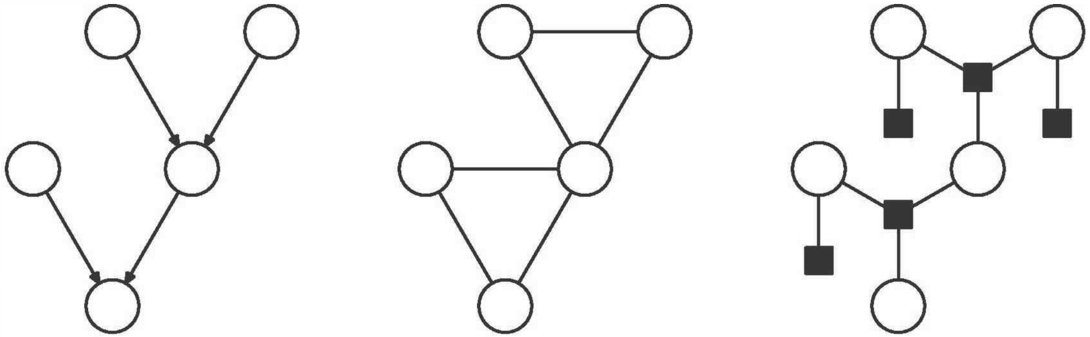


$$f_a(x_1) = p(x_1), \; f_b(x_2) = p(x_2)$$
$$f_c(x_1, x_2, x_3) = p(x_3|x_1, x_2)$$

  - There can be multiple factor graphs all of which correspond to the same undirected/directed graph

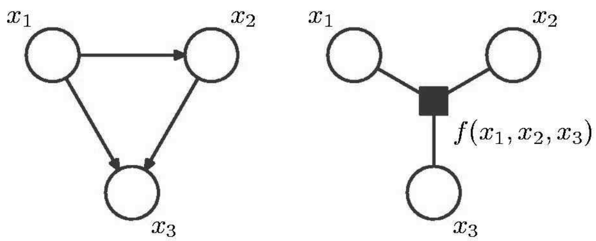
8.4.3 Factor graphs (3)

- Converting directed/undirected tree to a factor graph
  - The result is again a tree (no loops, one and only one path connecting any two nodes)
- In the case of a directed polytree
  - To undirected: results in loops due to the moralization step
  - To factor graphs: we can avoid loops



8.4.3 Factor graphs (4)

- Local cycles in a directed graph can be removed on conversion to a factor graph



- Factor graphs are more specific about the precise form of the factorization