CS 8501 Advanced Topics in Machine Learning

Lecture 05: Undirected Graphical Models

Yangfeng Ji

Information and Language Processing Lab

Department of Computer Science

University of Virginia

https://yangfengji.net/

Introduction

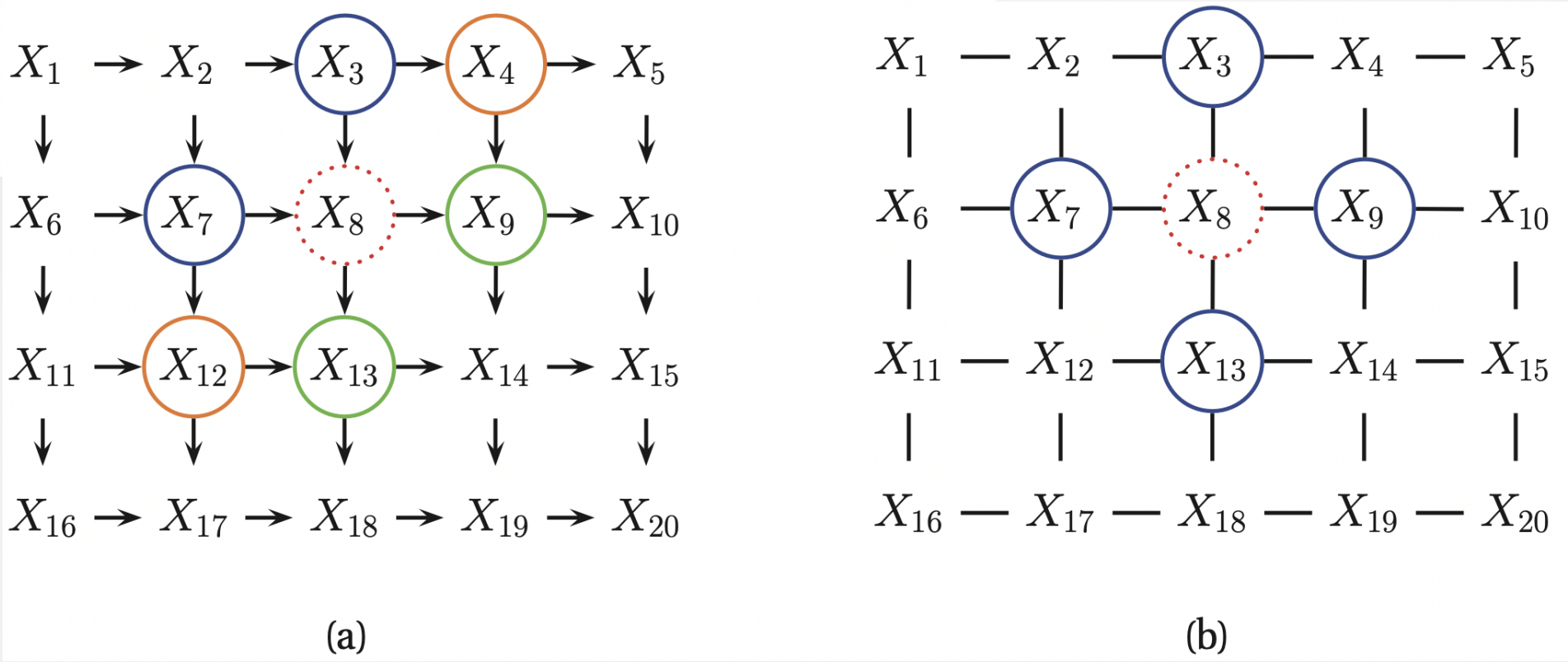
Other Names

Undirected graphical models (UGMs) also have some other names in the literature, e.g.,

- Markov random fields
- Markov networks

Markov Random Fields

MRFs are more natural to represent some data, for example, images



Conditional Independence Properties

Conditional Independence

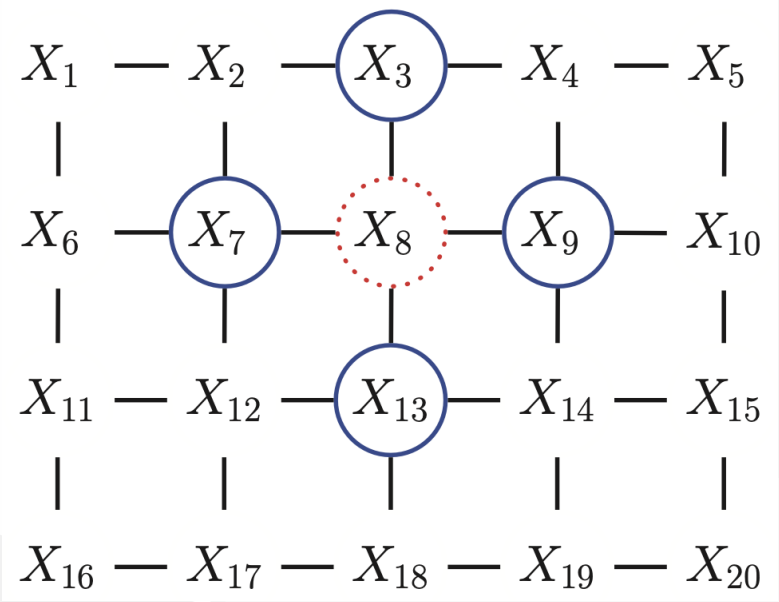
For sets of nodes A, B, and C

$$x_A \perp_G x_B \mid x_C$$

if and only if C separates A from B in the graph G

- by removing all the nodes in C and the connected edges, then see whether there is a path connecting a node in A with a node in B

Example



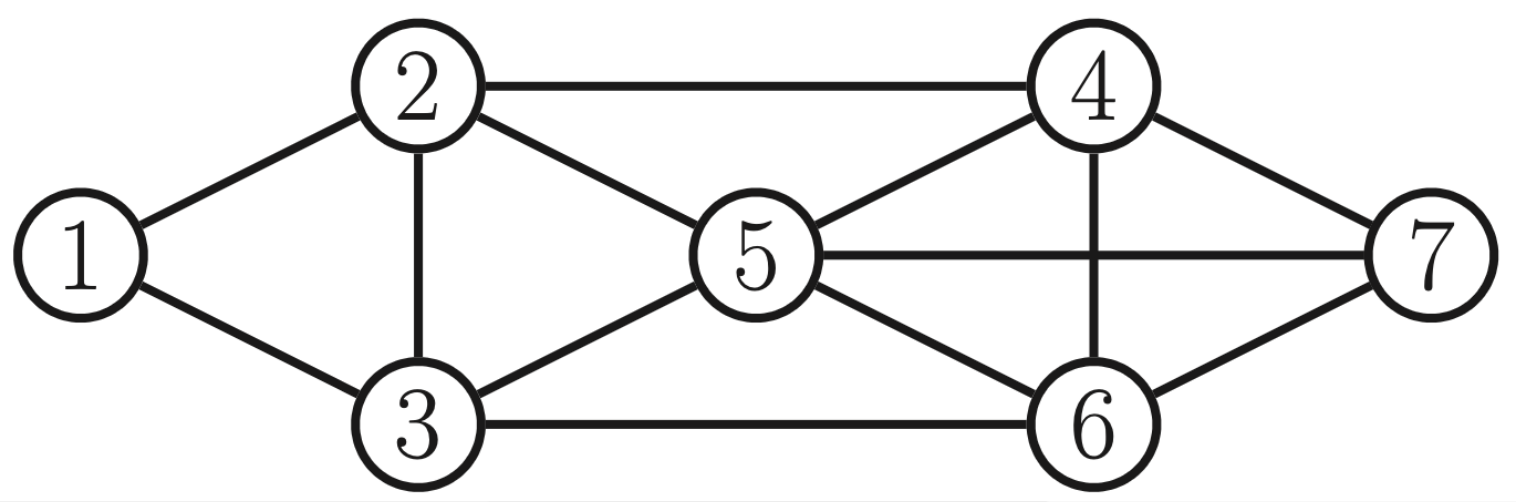
In this figure, if $A = \{8\}$ and $C = \{3, 7, 9, 13\}$ then

$$x_A \perp x_B \mid x_C$$

where $B \subseteq \mathcal{V} \backslash A \cup C$ is a subset of any other nodes

Conditional Independence: Pairwise

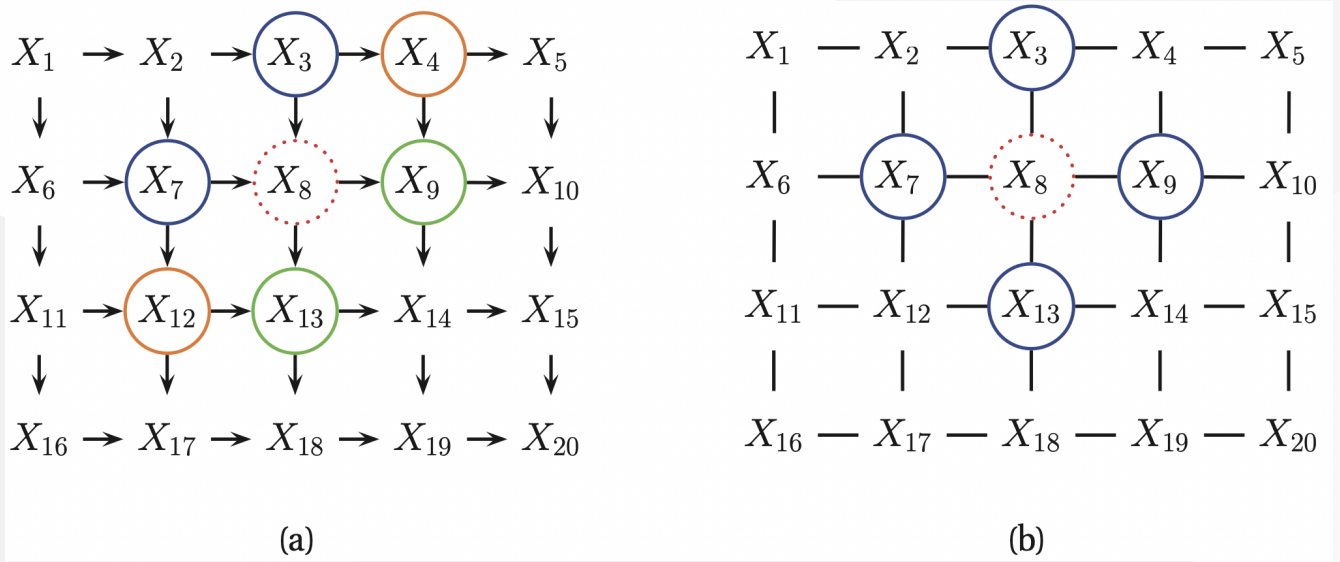
Two random variables are independent from each other, if the paths connected these two random variables are all blocked by observed variables



$$X_1 \perp\!\!\!\perp X_7 | \text{rest}$$

Markov Blanket

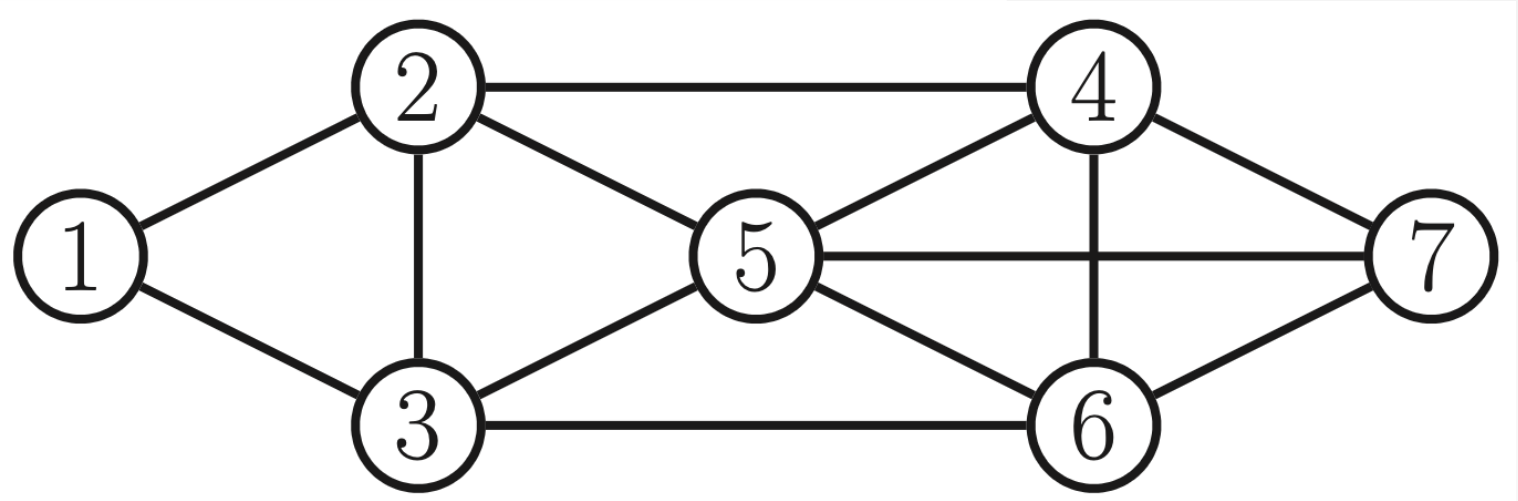
Markov blanket: the set of nodes $\mathrm{mb}(t)$ that renders a node t conditionally independent of all other nodes in the graph.



- (a): $\mathrm{mb}(18) = \{3, 4, 7, 9, 12, 13\}$
- (b): $\mathrm{mb}(18) = \{3, 7, 9, 13\}$

Conditional Independence: Local

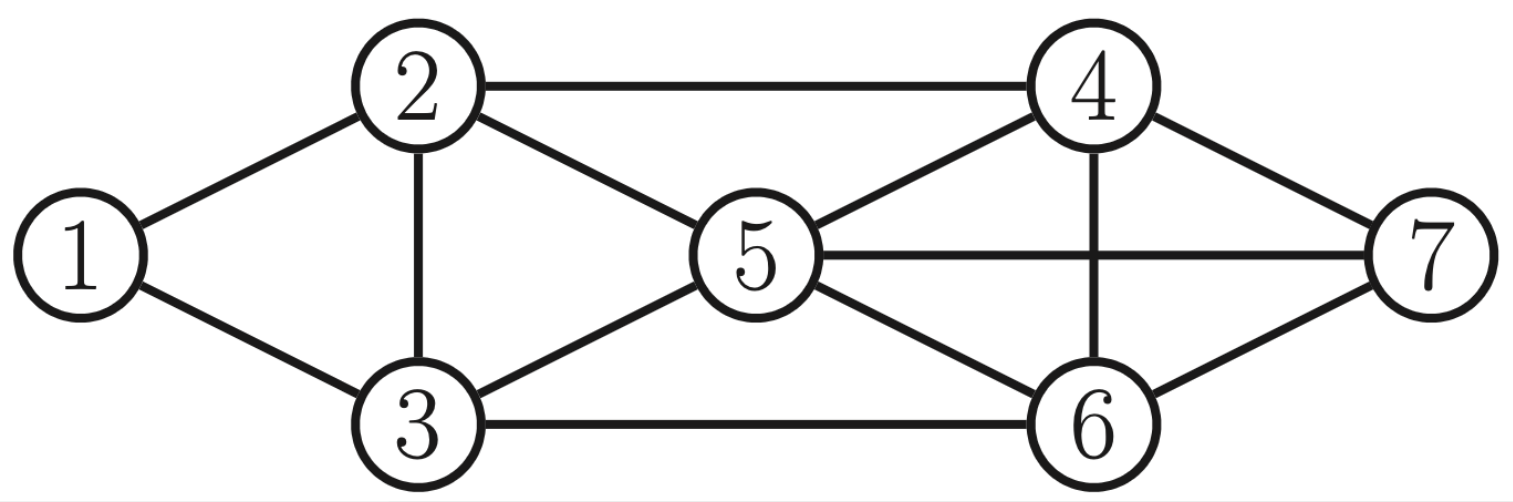
$$X_1 \perp\!\!\!\perp \text{rest} | \text{mb}(X_1)$$



Conditional Independence: Global

X_A and X_B are independent, if there is no path from between A and B given C

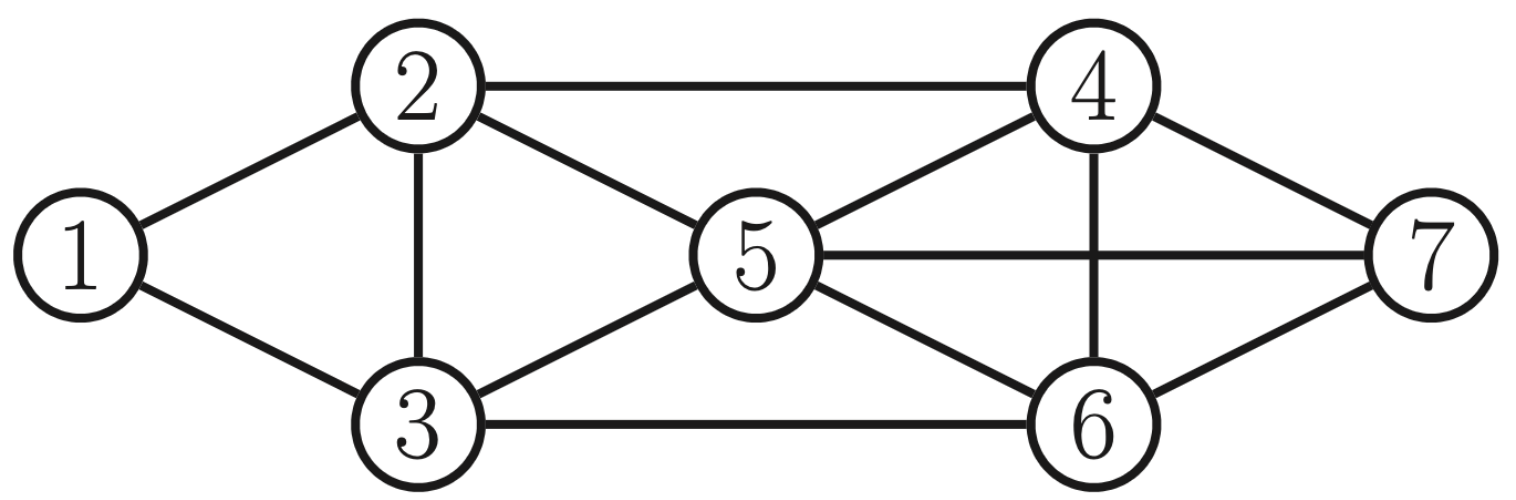
$$\{X_1, X_2\} \perp\!\!\!\perp \{X_6, X_7\} | \{X_3, X_4, X_5\}$$



Conditional Independence

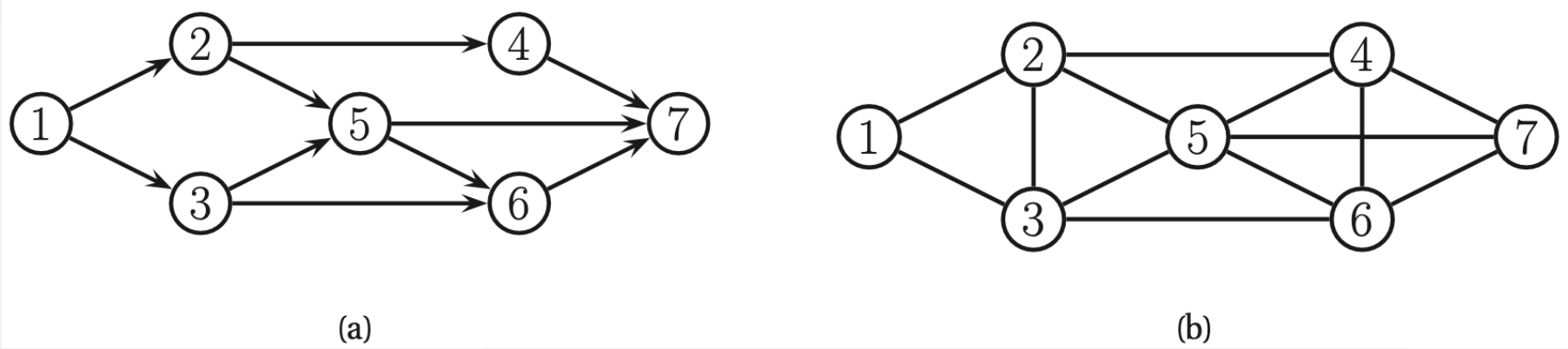
There are three types of conditional independence

- Pairwise independence
- Local independence
- Global independence



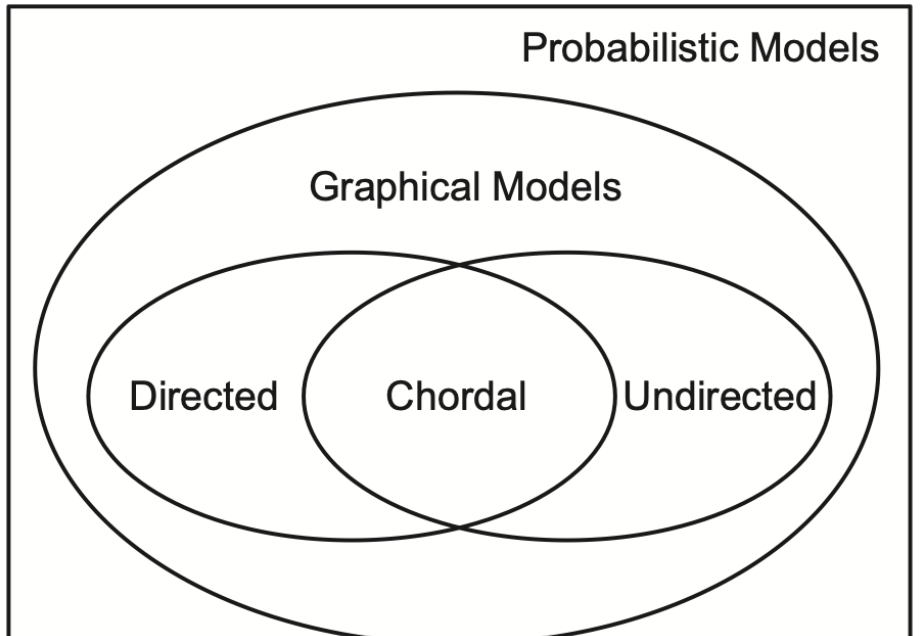
Moralization

- **Moralization**: the process of converting a directed graph to a undirected graph
- To avoid introducing incorrect **conditional independence**, the moralization process needs to add some extra edges.



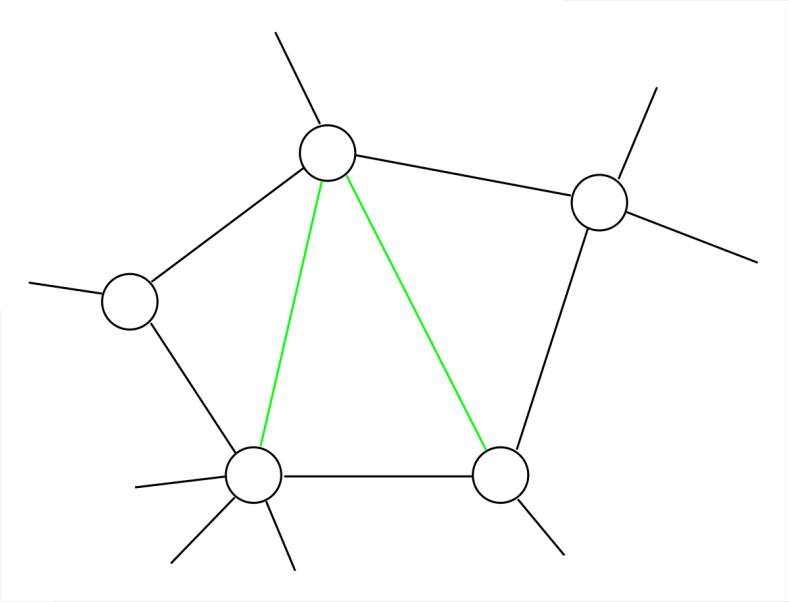
Comparison

- **I-map**: G is an I-map of a distribution p, if $I(G) \subseteq I(p)$
- **Perfect map**: if $I(G) = I(p)$
- Directed graphs and undirected graphs are perfect maps for different sets of distribution, unless the graph is a chordal graph



Chordal Graphs

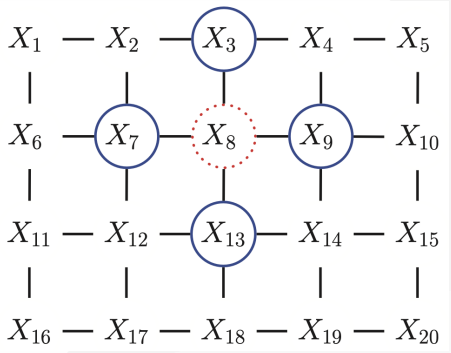
A chordal graph is a simple graph in which every graph cycle of length four and greater has a cycle chord.



Some Examples MRFs

Ising Models

Consider the image example with binary pixel values, where x_m and x_n are two neighbor pixels



$$E(x; J, H) = -\left[\frac{1}{2}\sum_{(m,n)\in G} J_{mn}x_m x_n + \sum_n H x_n\right]$$

Unlike directed graphical models that can specify conditional independence for any two given adjacent random variables, formulation on undirected graphical models mostly focuses on "interaction"

Ising Models: From Energy Function to Probabilitic Model

- Energy function

$$E(x; J, H) = -\Big[\frac{1}{2}\sum_{(m,n)\in G} J_{mn}x_m x_n + \sum_n H x_n\Big]$$

- The sign of J indicates whether we want to encourage x_m and x_n to have the same value
- The sign of H indicates what value we want each individual x_n to have
- Probabilistic formulation

$$p(x; \beta, J, H) \propto \exp[-\beta E(x; J, H)]$$

Ising Models: Partition Function

$$p(x;\beta,J,H) = \frac{1}{Z(\beta,J,H)}\exp[-\beta E(x;J,H)]$$

where

$$Z(\beta,J,H) = \sum_x exp[-\beta E(x;J,H)]$$

is the partition function.

Boltzmann Distribution

Boltzmann distribution (also called **Gibbs distribution**) is a distribution can be formulated as

$$p(x) \propto \exp[-E(x;\theta)]$$

where $E(x;\theta)$ is an energy function and θ represents the parameter of this function.

Generalization of Ising Models

- Potts models
- Hopfield networks
- Boltzmann machines
- Restricted Boltzmann machines

Potts Models

Potts models: by extending x_n from binary random variable to $x_n \in \{1, 2, \ldots, K\}$, and J is in a matrix form as

$$J = [J_{ij}]$$

and J_{ij} indicates the interaction between $x_m = i$ and $x_n = j$

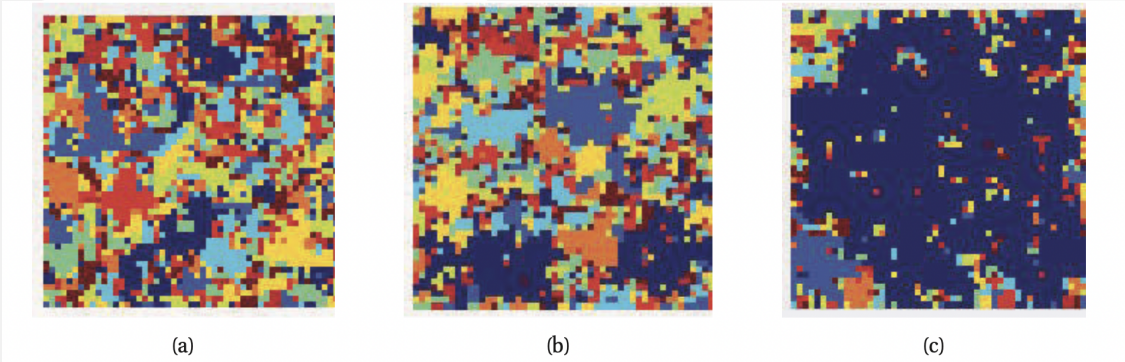
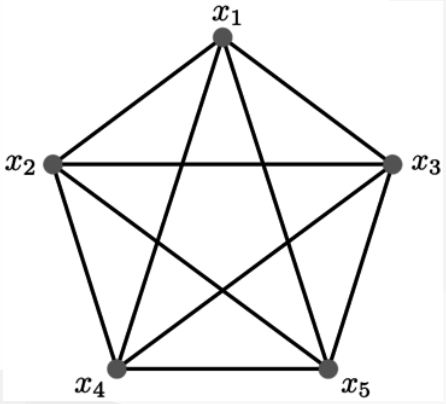


Figure 19.8 Visualizing a sample from a 10-state Potts model of size 128×128 for different association strengths: (a) $J = 1.42$, (b) $J = 1.44$, (c) $J = 1.46$. The regions are labeled according to size: blue is largest, red is smallest. Used with kind permission of Erik Sudderth. See `gibbsDemoIsing` for Matlab code to produce a similar plot for the Ising model.

Hopfield Networks

Hopfield networks: by extending Ising models to a fully-connected graph, with $J_{mn} = J_{nm}$, each x_n can still be binary



$$E(x; J, H) = -\frac{1}{2}\sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n$$

- Hopfield networks can also defined on the image cases
- Essentially, it introduce more dependence on the graph

Hopfield Networks (II)

Because of the dependence between any two nodes, the correlation can act as some kind of memory to constrain the values between nodes. For example

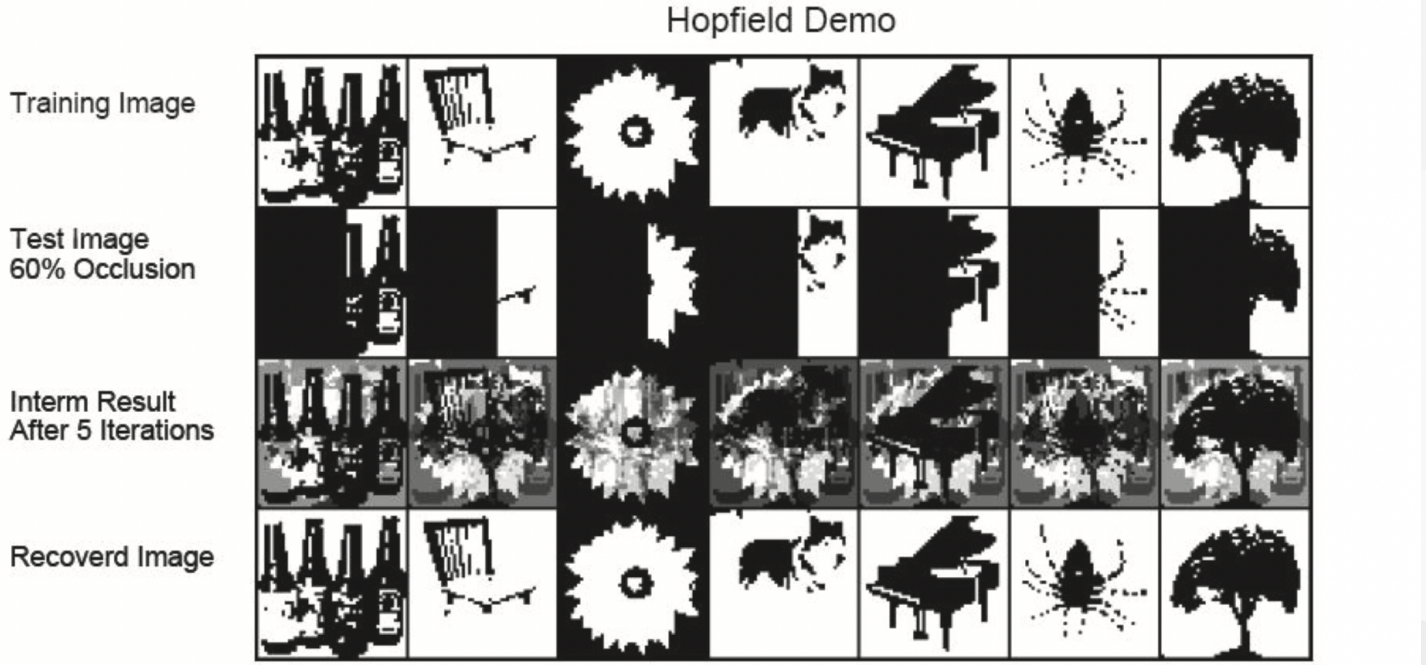
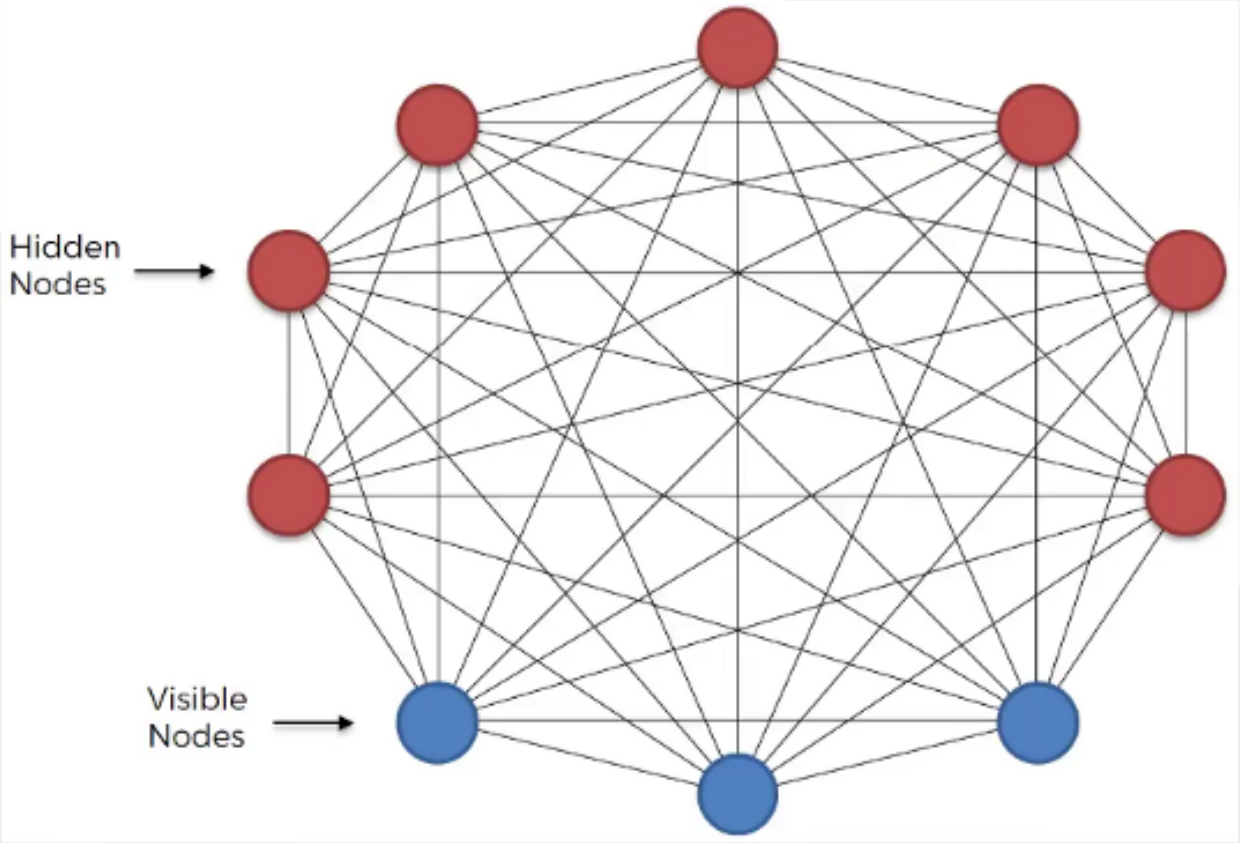


Figure 19.7 Examples of how an associative memory can reconstruct images. These are binary images of size 50×50 pixels. Top: training images. Row 2: partially visible test images. Row 3: estimate after 5 iterations. Bottom: final state estimate. Based on Figure 2.1 of Hertz et al. (1991). Figure generated by

Boltzmann Machines

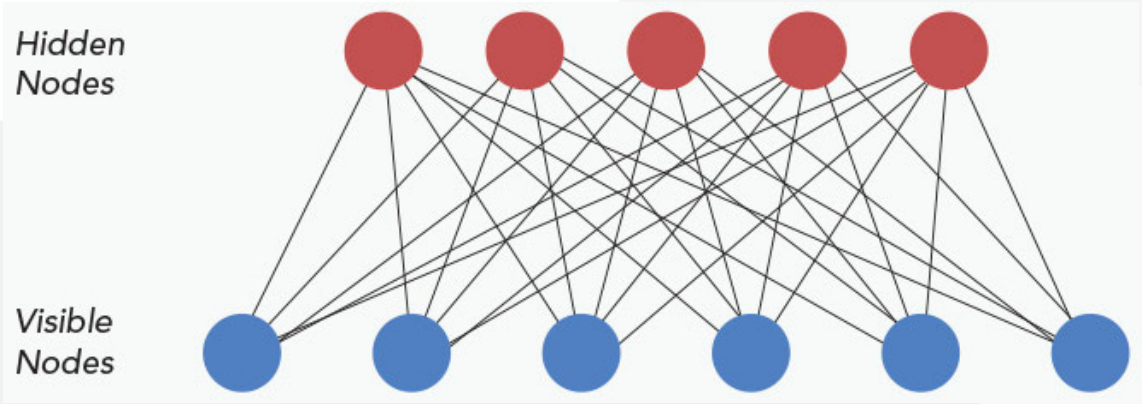
Boltzmann Machines is a generalization of the Hopfield networks with latent variables



Restricted Boltzmann machine

The energy function of the RBM

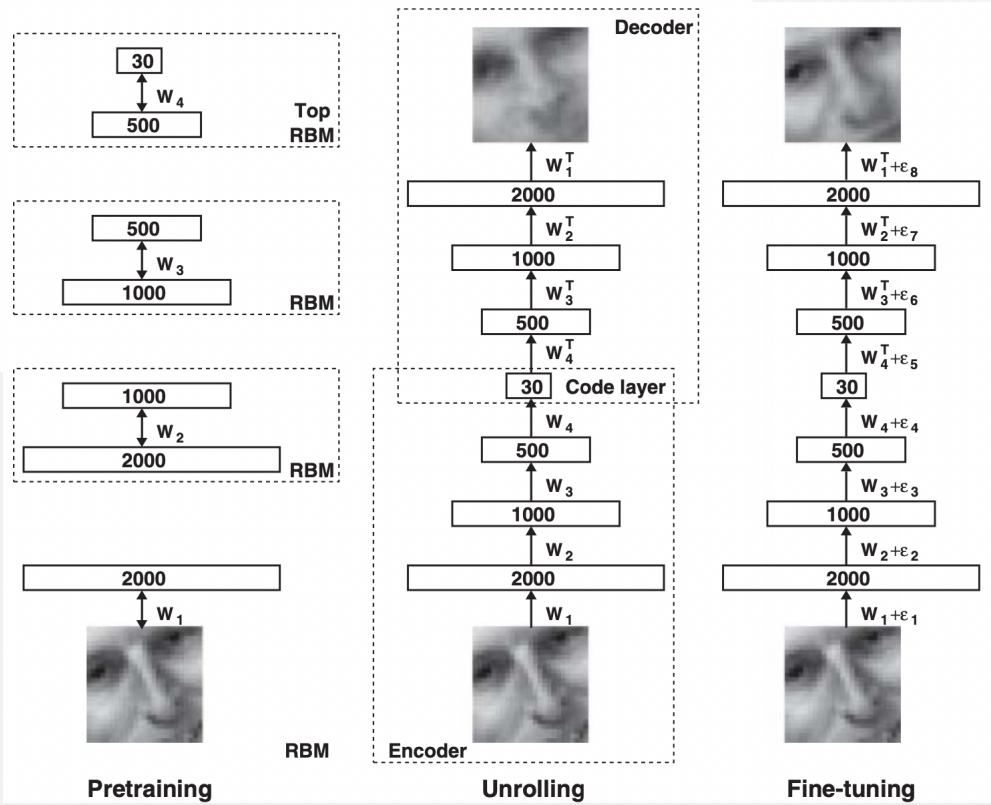
$$E(x, z; J, U, H) = -\sum_{m,n} J_{mn} x_m z_n - \sum_m u_m x_m - \sum_n h_n z_n$$



This architecture provides the possibility of building multi-layer hidden variables.

RBMs for Pre-training

From [Hinton et al., 2006; Science]



This work marks the beginning of deep learning

Parameterization of MRFs

- Representing the joint distribution for a UGM is less natural than for a DGM

Hammersley-Clifford Theorem

A positive distribution $p(y) > 0$ satisfies the CI properties of an undirected graph G iff p can be represented as a product of factors, one per maximal clique, i.e.,

$$p(y \mid \theta) = \frac{1}{Z(\theta)} \prod_{c \in \mathcal{C}} \psi_c(y_c \mid \theta_c)$$

where $\mathcal{C}$ is the set of all the (maximal) cliques of G, and $Z(\theta)$ is the partition function given by

$$Z(\theta) = \sum_x \prod_{c \in \mathcal{C}} \psi_c(y_c \mid \theta_c)$$

Thank You!