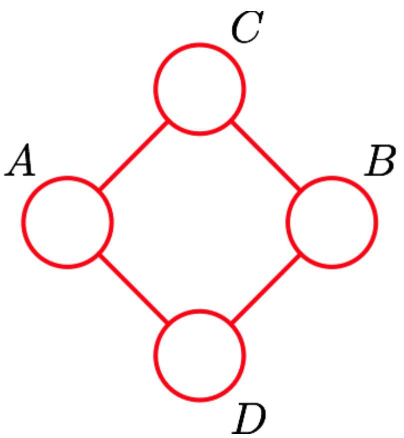
Graphical Models

Undirected Graphical Models

Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are useful for expressing soft constraints between random variables



- The joint distribution defined by the graph is given by the product of non-negative potential functions over the maximal cliques (connected subset of nodes).

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) \qquad \mathcal{Z} = \sum_{\mathbf{x}} \prod_C \phi_C(x_C)$$

where the normalizing constant $\mathcal{Z}$ is called a partition function.

- For example, the joint distribution factorizes:

$$p(A, B, C, D) = \frac{1}{\mathcal{Z}} \phi(A, C)\phi(C, B)\phi(B, D)\phi(A, D)$$

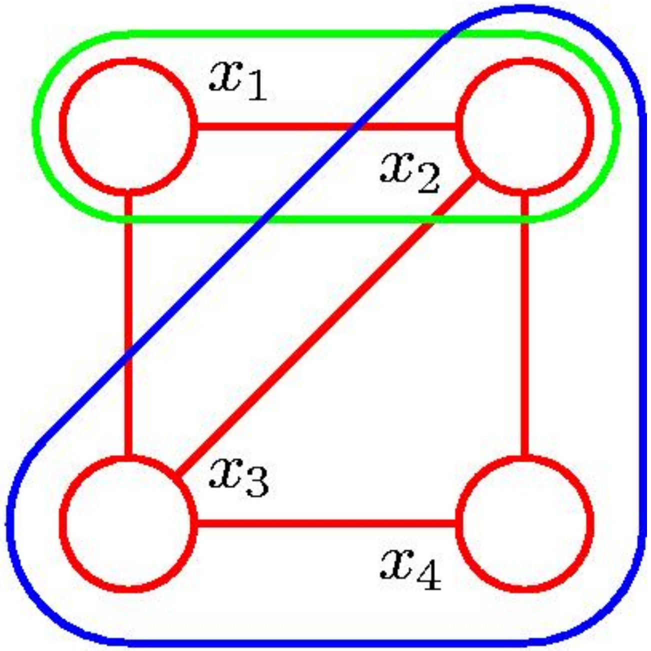
- Let us look at the definition of cliques.

Cliques

- The subsets that are used to define the potential functions are represented by maximal cliques in the undirected graph.

- Clique: a subset of nodes such that there exists a link between all pairs of nodes in a subset.

- Maximal Clique: a clique such that it is not possible to include any other nodes in the set without it ceasing to be a clique.



- This graph has 5 cliques:

$$\{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\},$$
$$\{x_4, x_2\}, \{x_1, x_3\}.$$

- Two maximal cliques:

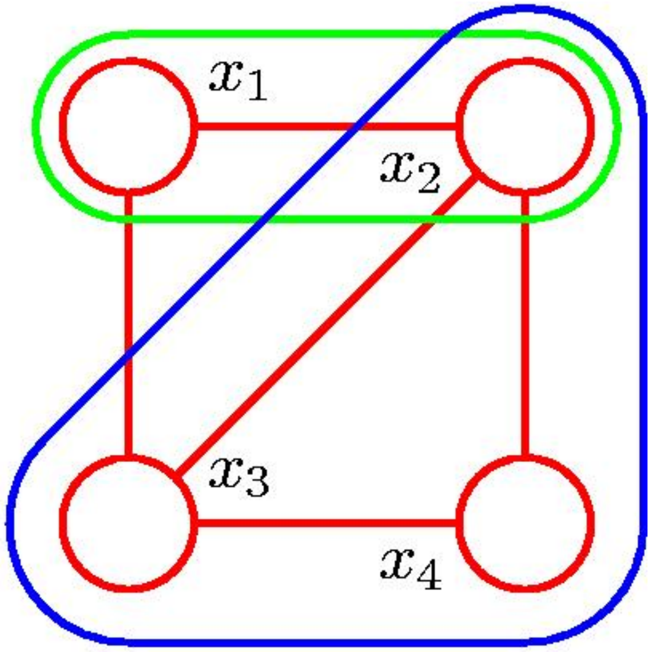
$$\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}.$$

Using Cliques to Represent Subsets

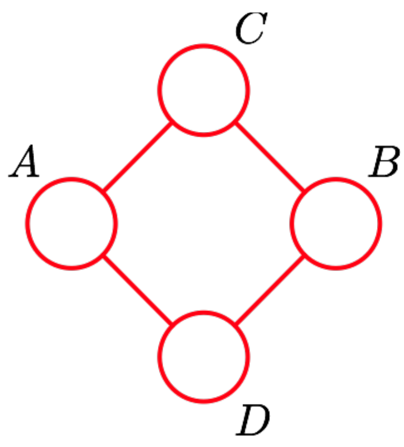
• If the potential functions only involve two nodes, an undirected graph has a nice representation.

• If the potential functions involve more than two nodes, using a different factor graph representation is much more useful.

• For now, let us consider only potential functions that are defined over two nodes.



Markov Random Fields (MRFs)



$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C)$$

- Each potential function is a mapping from the joint configurations of random variables in a clique to non-negative real numbers.

- The choice of potential functions is not restricted to having specific probabilistic interpretations.

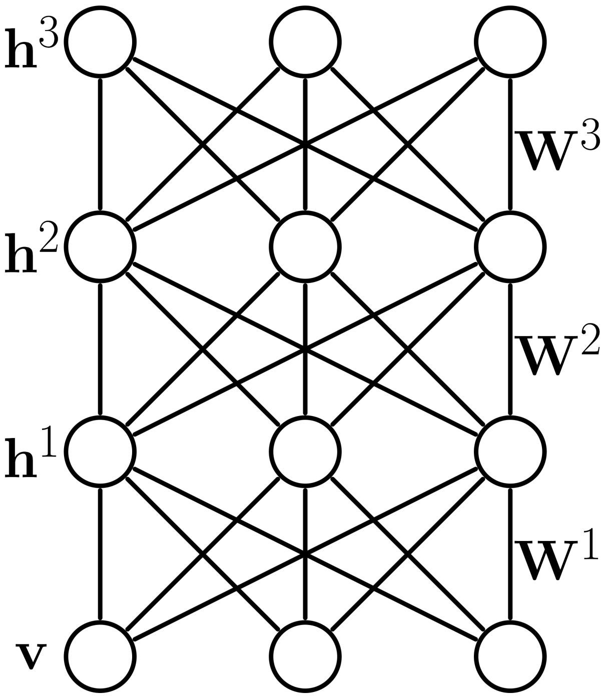
Potential functions are often represented as exponentials:

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) = \frac{1}{\mathcal{Z}} \exp(-\sum_C E(x_c)) = \underbrace{\frac{1}{\mathcal{Z}} \exp(-E(\mathbf{x}))}_{\text{Boltzmann distribution}}$$

where E(x) is called an energy function.

MRFs with Hidden Variables

For many interesting real-world problems, we need to introduce hidden or latent variables.



- Our random variables will contain both visible and hidden variables x=(v,h).

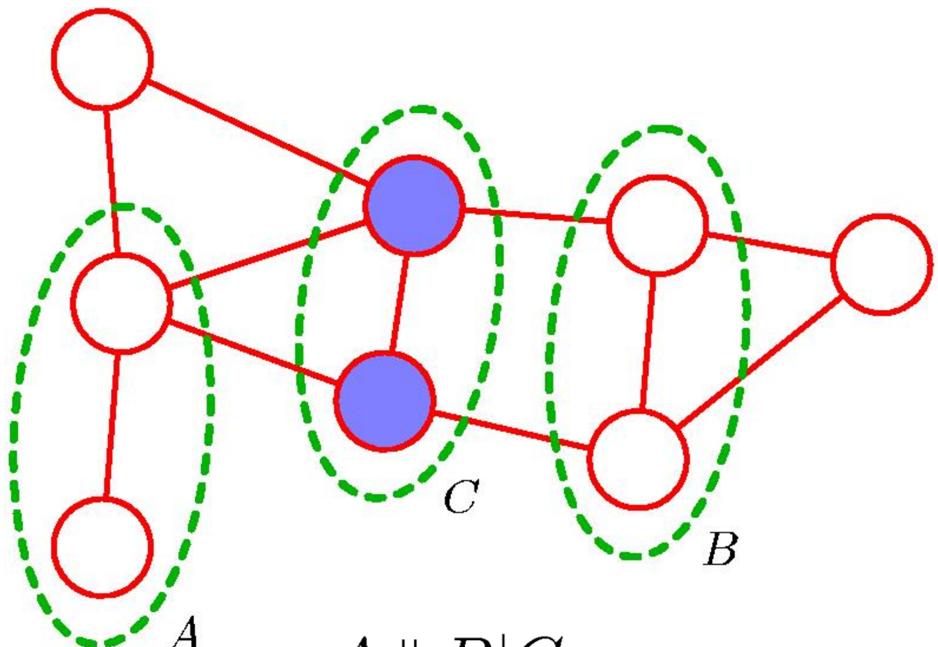
$$p(\mathbf{v}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$$

- In general, computing both partition function and summation over hidden variables will be intractable, except for special cases.

- Parameter learning becomes a very challenging task.

Conditional Independence

- Conditional Independence is easier compared to directed models:

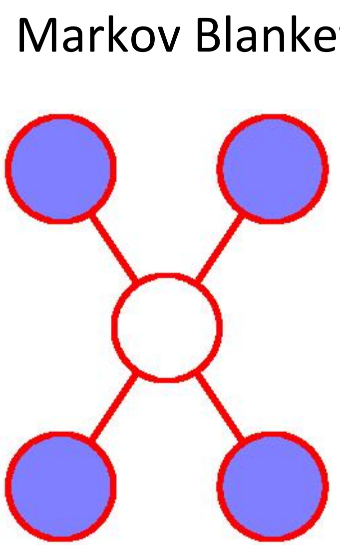


$$A \perp\!\!\!\perp B \mid C$$

- Observation blocks a node.
- Two sets of nodes are conditionally independent if the observations block all paths between them.

Markov Blanket

• The Markov blanket of a node is simply all of the directly connected nodes.

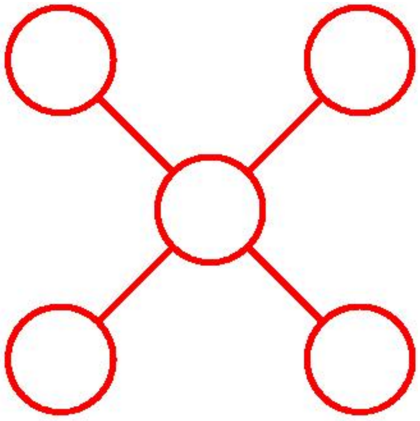


• This is simpler than in directed models, since there is no explaining away.

• The conditional distribution of x_i conditioned on all the variables in the graph is dependent only on the variables in the Markov blanket.

Conditional Independence and Factorization

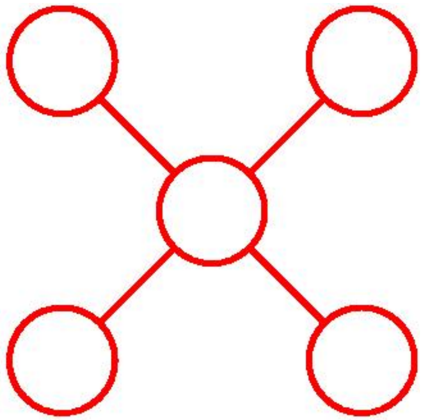
- Consider two sets of distributions:
  - The set of distributions consistent with the conditional independence relationships defined by the undirected graph.
  - The set of distributions consistent with the factorization defined by potential functions on maximal cliques of the graph.
- The Hammersley-Clifford theorem states that these two sets of distributions are the same.



$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \phi_C(x_C)$$

Interpreting Potentials

• In contrast to directed graphs, the potential functions **do not have a specific probabilistic interpretation**.



$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_C \phi_C(x_C) = \frac{1}{\mathcal{Z}} \exp(-\sum_C E(x_c))$$

• This gives us greater flexibility in choosing the potential functions.

• We can view the potential function as expressing which configuration of the local variables are preferred to others.

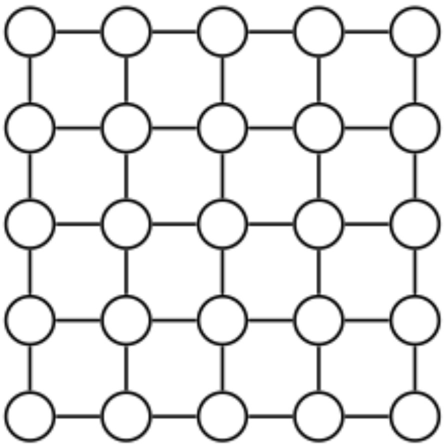
• Global configurations with relatively high probabilities are those that find a good balance in satisfying the (possibly conflicting) influences of the clique potentials.

• So far we did not specify the nature of random variables, discrete or continuous.

Discrete MRFs

- MRFs with all discrete variables are widely used in many applications.

- MRFs with binary variables are sometimes called Ising models in statistical mechanics, and Boltzmann machines in machine learning



- Denoting the binary valued variable at node j by $x_j \in \{0, 1\}$, the Ising model for the joint probabilities is given by:

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big(\sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \Big)$$

- The conditional distribution is given by logistic:

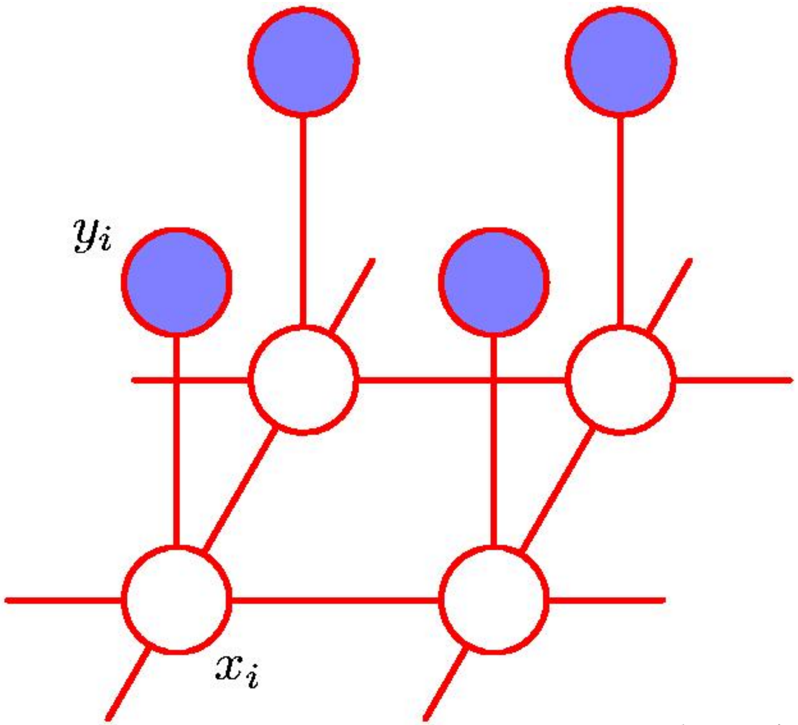
$$P_\theta(x_i = 1 | \mathbf{x}_{-i}) = \frac{1}{1 + \exp(-\theta_i - \sum_{ij \in E} x_j \theta_{ij})},$$

where x_{-i} denotes all nodes except for i.

Hence the parameter θ_{ij} measures the dependence of x_i on x_j, conditional on the other nodes.

Example: Image Denoising

- Let us look at the example of noise removal from a binary image.
- Let the observed noisy image be described by an array of binary pixel values: $y_j \in \{-1, +1\}$, i=1,…,D.



- We take a noise-free image $x_j \in \{-1, +1\}$, and randomly flip the sign of pixels with some small probability.

Bias term

Neighboring pixels are likely to have the same sign

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

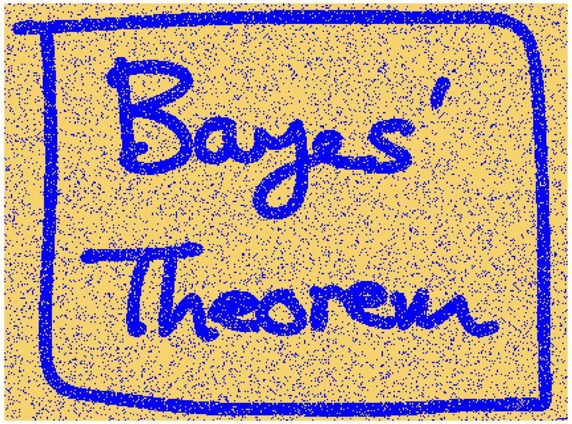
Noisy and clean pixels are likely to have the same sign

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

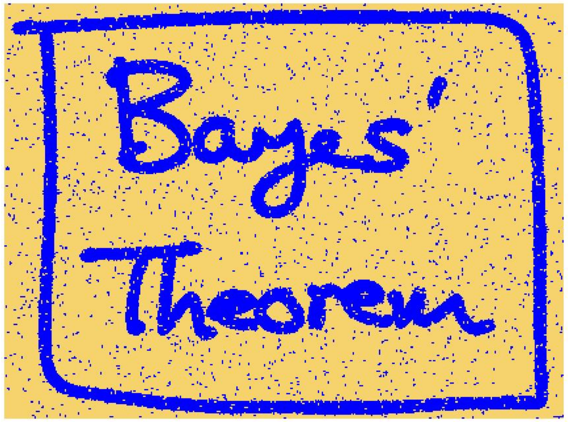
Iterated Conditional Modes

- Iterated conditional modes: coordinate-wise gradient descent.
- Visit the unobserved nodes sequentially and set each x to whichever of its two values has the lowest energy.
  - This only requires us to look at the Markov blanket, i.e. the connected nodes.
  - Markov blanket of a node is simply all of the directly connected nodes.

Original Image



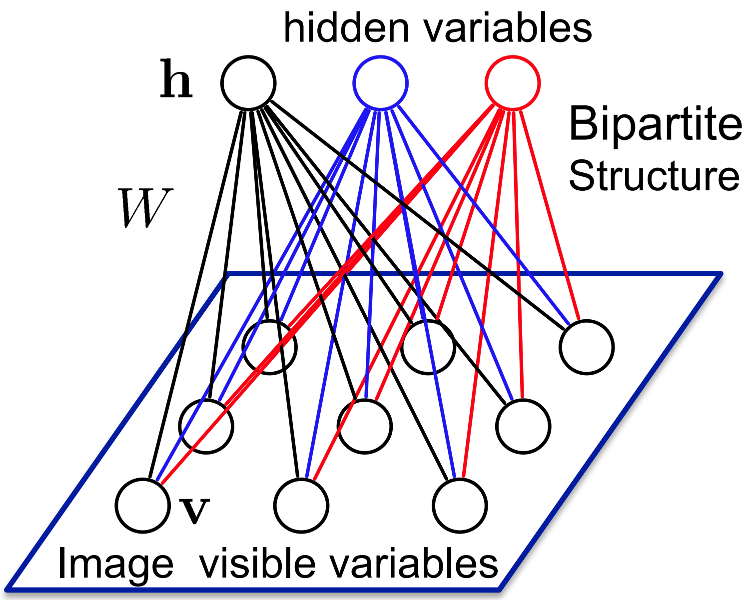
Noisy Image



ICM

Restricted Boltzmann Machines

- For many real-world problems, we need to introduce hidden variables.
- Our random variables will contain visible and hidden variables x=(v,h).



Stochastic binary visible variables $\mathbf{v} \in \{0,1\}^D$ are connected to stochastic binary hidden variables $\mathbf{h} \in \{0,1\}^F$.

The energy of the joint configuration:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{ij} W_{ij} v_i h_j - \sum_i b_i v_i - \sum_j a_j h_j$$

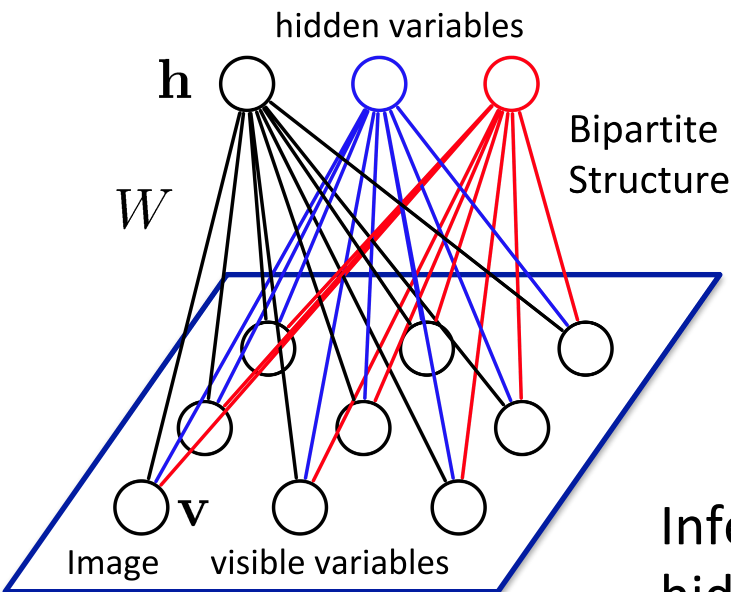
$\theta = \{W, a, b\}$ model parameters.

Probability of the joint configuration is given by the Boltzmann distribution:

$$P_\theta(\mathbf{v},\mathbf{h}) = \frac{1}{\mathcal{Z}(\theta)} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big) = \underbrace{\frac{1}{\mathcal{Z}(\theta)}}_{\text{partition function}} \prod_{ij} \underbrace{e^{W_{ij} v_i h_j}}_{\text{potential functions}} \prod_i e^{b_i v_i} \prod_j e^{a_j h_j}$$

$$\mathcal{Z}(\theta) = \sum_{\mathbf{h},\mathbf{v}} \exp\big(-E(\mathbf{v},\mathbf{h};\theta)\big)$$

Restricted Boltzmann Machines



Restricted: No interaction between hidden variables

Inferring the distribution over the hidden variables is easy:

$$\underbrace{P(\mathbf{h}|\mathbf{v}) = \prod_j P(h_j|\mathbf{v})}_{\text{Factorizes: Easy to compute}} \qquad P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp(-\sum_i W_{ij} v_i - a_j)}$$

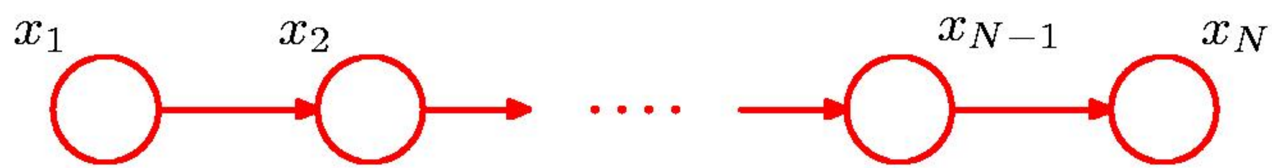
Similarly:

$$P(\mathbf{v}|\mathbf{h}) = \prod_i P(v_i|\mathbf{h}) \qquad P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp(-\sum_j W_{ij} h_j - b_i)}$$

Markov random fields, Boltzmann machines, log-linear models.

Relation to Directed Graphs

- Let us try to convert directed graph into an undirected graph:



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

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

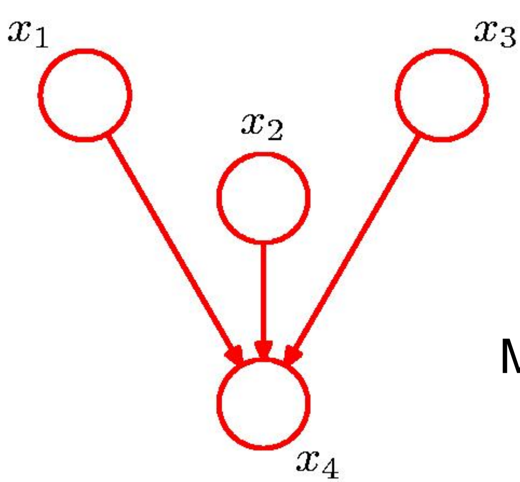
Directed vs. Undirected

• Directed Graphs can be more precise about independencies than undirected graphs.

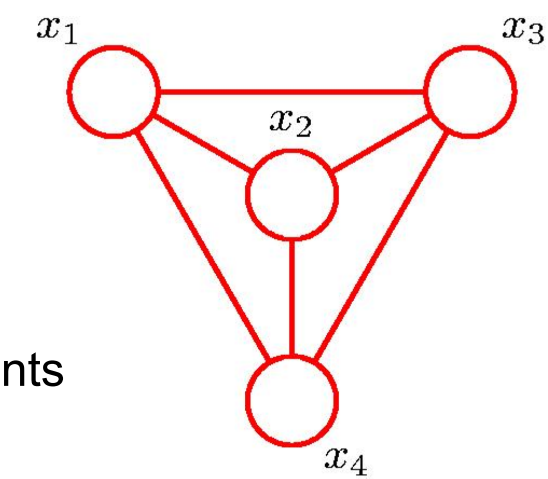
$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$

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

need 4th
order clique



Moralize: Marry the parents



• All the parents of x_4 can interact to determine the distribution over x_4.

• The directed graph represents independencies that the undirected graph cannot model.

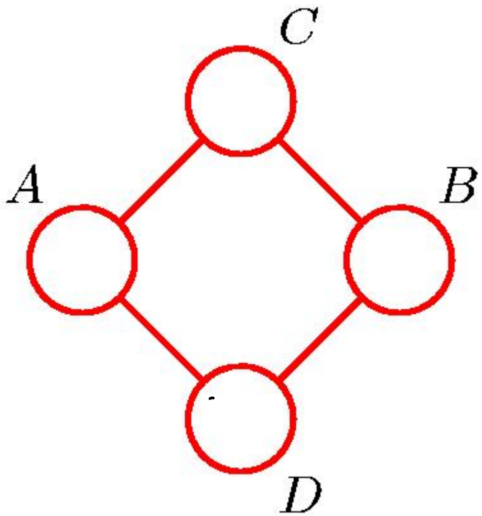
• To represent the high-order interaction in the directed graph, the undirected graph needs a fourth-order clique.

• This fully connected graph exhibits no conditional independence properties

Undirected vs. Directed

• Undirected Graphs can be more precise about independencies than directed graphs

• There is no directed graph over four variables that represents the same set of conditional independence properties.



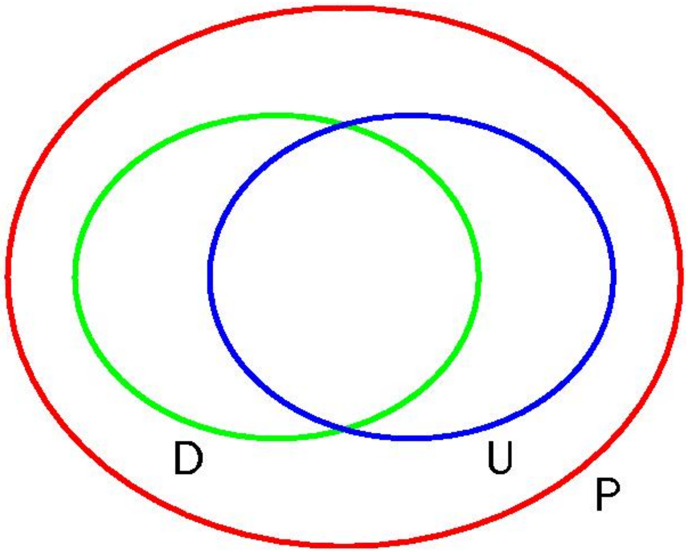
$$A \not\perp\!\!\!\perp B \mid \emptyset$$

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

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

Directed vs. Undirected

• If every conditional independence property of the distribution is reflected in the graph and vice versa, then the graph is a perfect map for that distribution.



• Venn diagram:

- The set of all distributions P over a given set of random variables.
- The set of distributions D that can be represented as a perfect map using directed graph.
- The set of distributions U that can be represented as a perfect map using undirected graph.

• We can extend the framework to graphs that include both directed and undirected graphs.