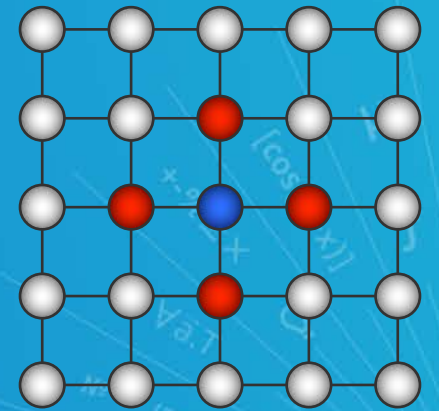


# Probabilistic Graphical Models

## Representation of undirected GM

Eric Xing

Lecture 2, January 15, 2020



**Reading: KF-chap4**



# Summary

- Last time
  - Conditional independence between two random variables
  - Conditional independence graph (CIG):
    - absence of an edge  $(X_1, X_2) \Rightarrow X_1 \perp X_2 | rest$
- Today
  - How can we read-off conditional independences from CIG?
  - What is the class of distributions represented by CIG?





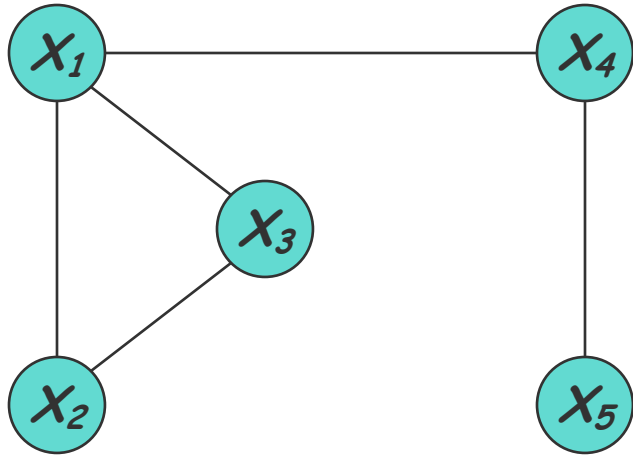
# Notation

- Variable, value and index
- Random variable
- Random vector
- Random matrix
- Parameters





# Undirected graphical models (UGM)



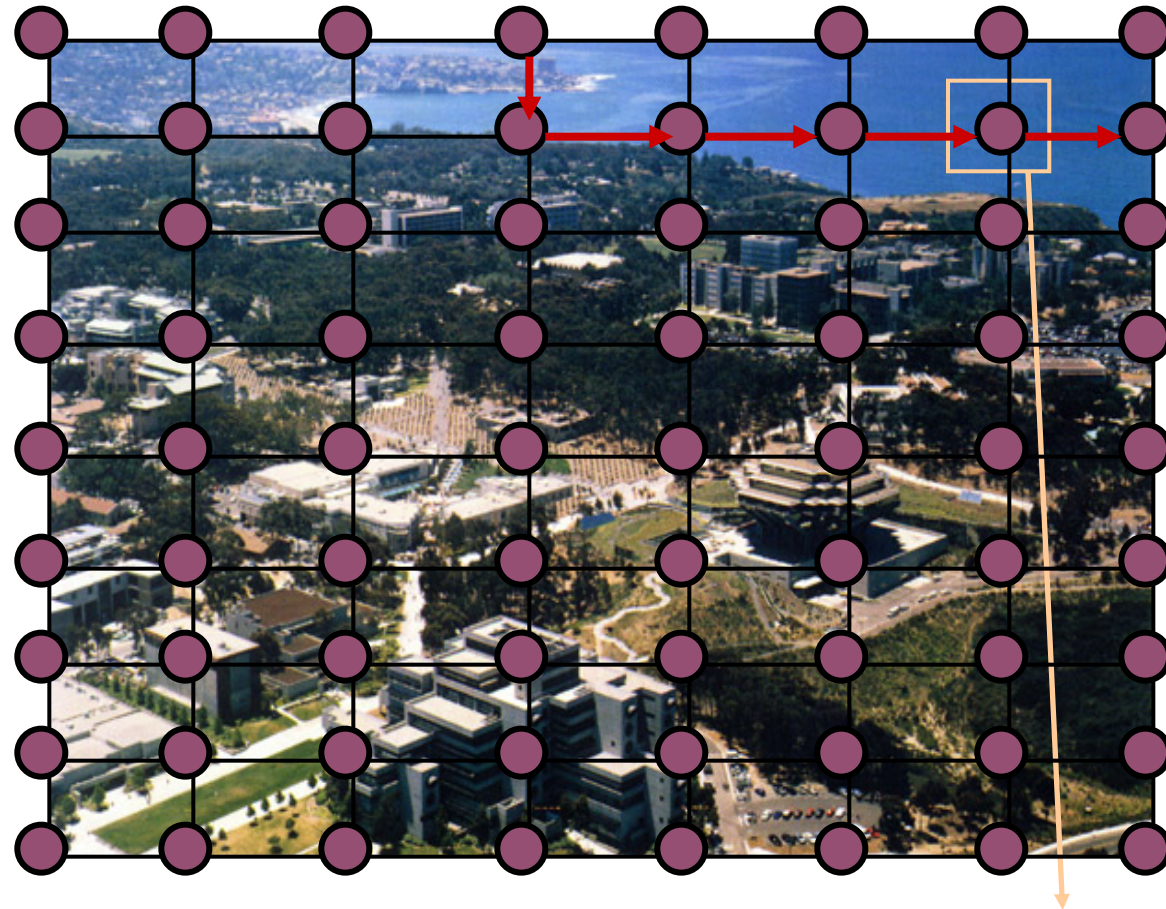
$$P(\mathbf{X}; \theta_G)$$

- ❑ Pairwise (non-causal) relationships
- ❑ Can write down model, and score specific configurations of the graph, but no explicit way to generate samples
- ❑ Contingency constrains on node configurations

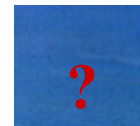




# A Canonical Example: understanding complex scene ...



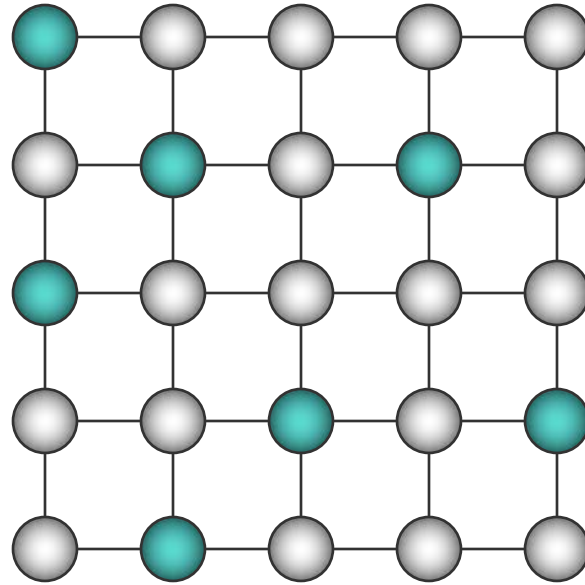
air or water ?





# A Canonical Example

- The grid model

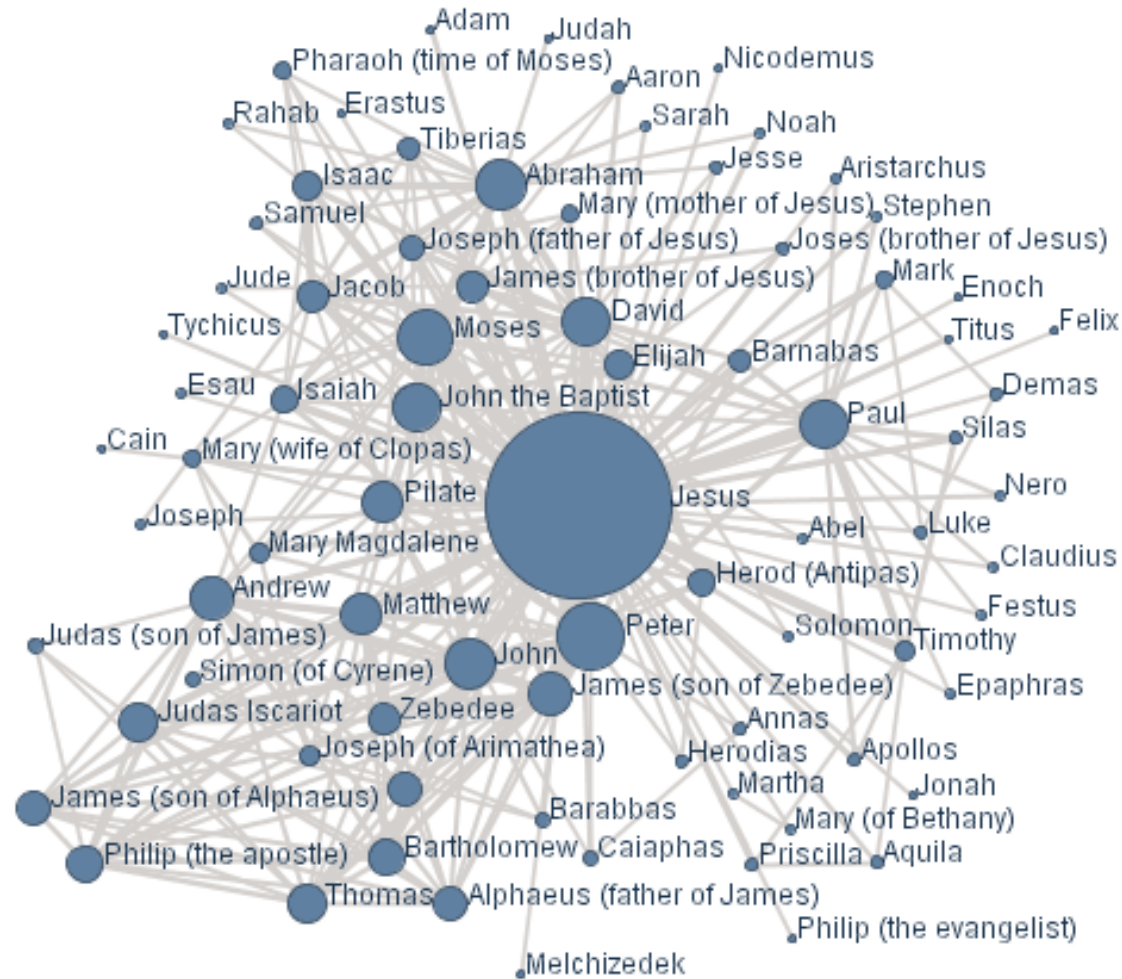


- Naturally arises in image processing, lattice physics, etc.
- Each node may represent a single "pixel", or an atom
  - The states of adjacent or nearby nodes are "coupled" due to pattern continuity or electro-magnetic force, etc.
  - Most likely joint-configurations usually correspond to a "low-energy" state





# Social networks



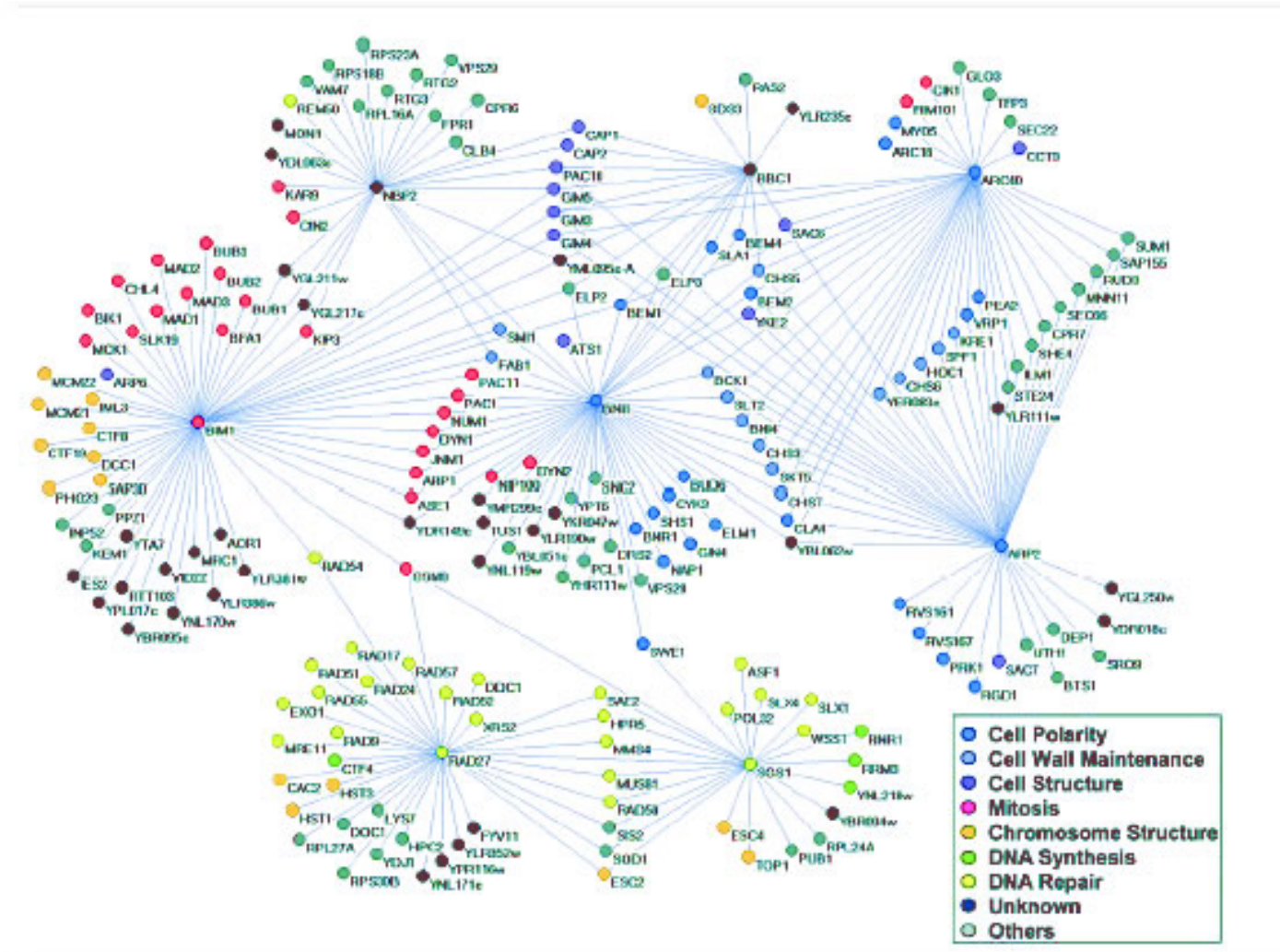
## The New Testament Social Networks







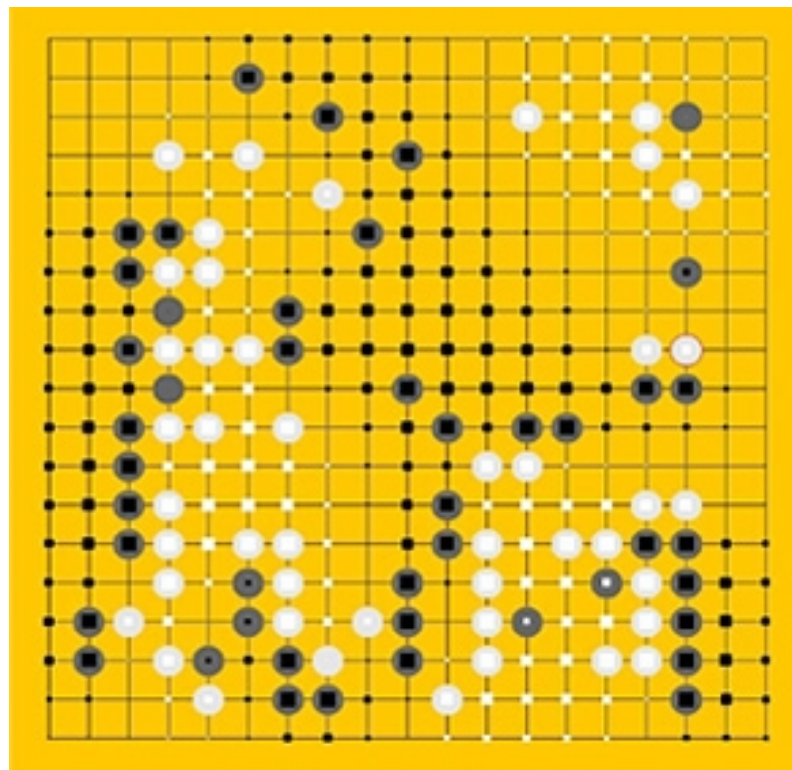
# Protein interaction networks







# Modeling Go

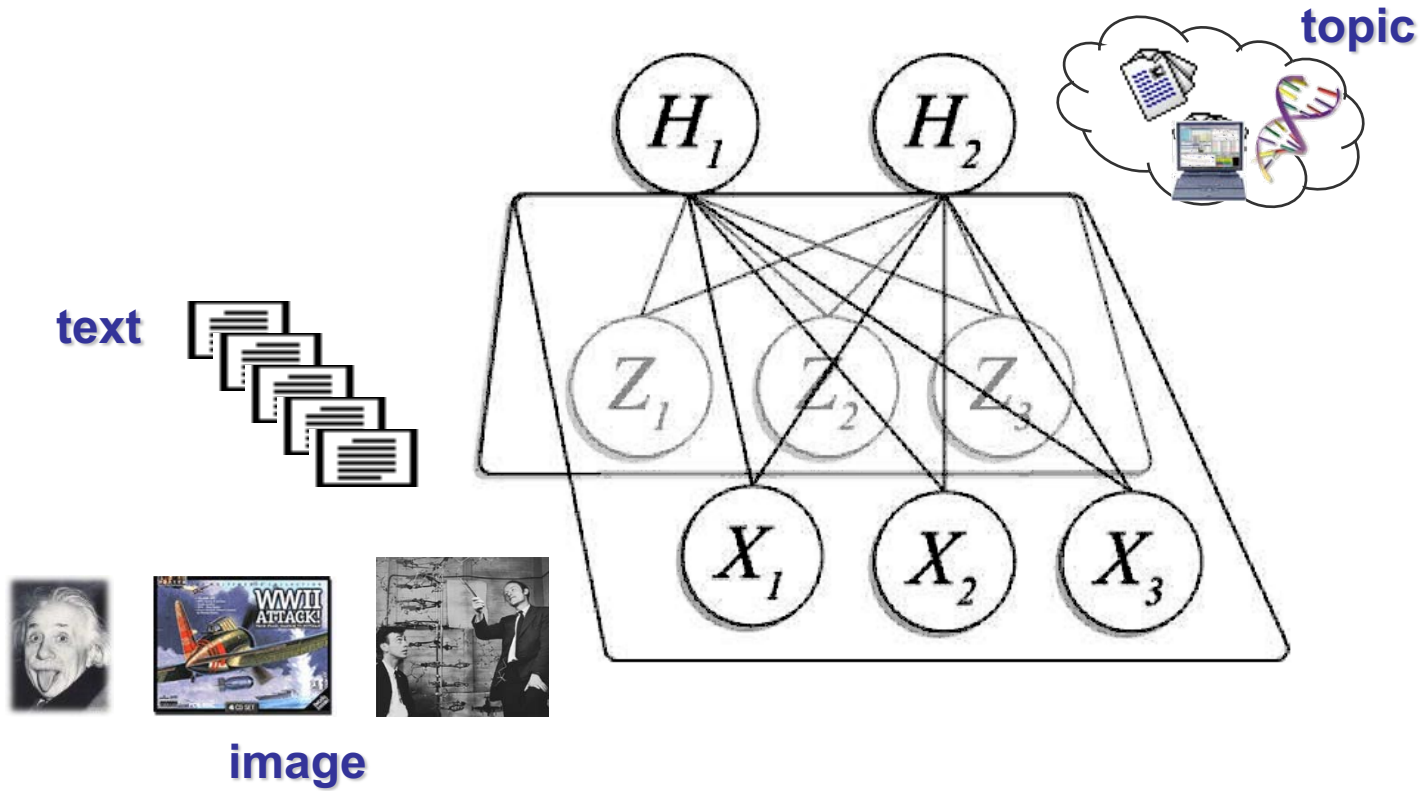


This is the middle position of a Go game.  
Overlaid is the estimate for the probability of becoming black or white for every intersection.  
Large squares mean the probability is higher.





# Information retrieval





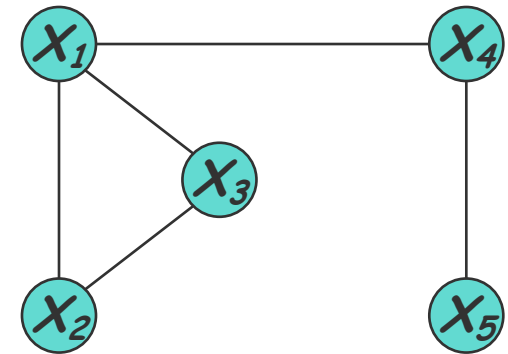
# Representation

- Defn: an **undirected graphical model** represents a distribution  $P(X_1, \dots, X_n)$  defined by an undirected graph  $H$ , and a set of positive **potential functions**  $\psi_c$  associated with the cliques of  $H$ , s.t.

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

where  $Z$  is known as the partition function:

$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$



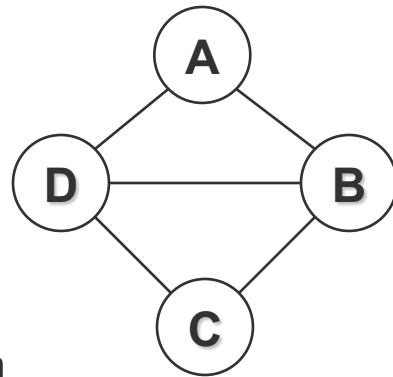
- Also known as **Markov Random Fields**, **Markov networks** ...
- The **potential function** can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.





# I. Quantitative Specification: Cliques

- For  $G=\{V,E\}$ , a complete subgraph (clique) is a subgraph  $G'=\{V'\subseteq V,E'\subseteq E\}$  such that nodes in  $V'$  are fully interconnected
- A (maximal) clique is a complete subgraph s.t. any **superset**  $V''\supset V'$  is not complete.
- A sub-clique is a not-necessarily-maximal clique.



- Example:
  - max-cliques =  $\{A,B,D\}, \{B,C,D\}$ ,
  - sub-cliques =  $\{A,B\}, \{C,D\}, \dots \rightarrow$  all edges and singletons
- Why Cliques: a basic unit that “all dependencies” are possible, and not to be missed





# Gibbs Distribution and Clique Potential

- Defn: an **undirected graphical model** represents a distribution  $P(X_1, \dots, X_n)$  defined by an undirected graph  $H$ , and a **set** of positive **potential functions**  $\psi_c$  associated with cliques of  $H$ , s.t.

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c) \quad (\text{A Gibbs distribution})$$

where  $Z$  is known as the partition function:

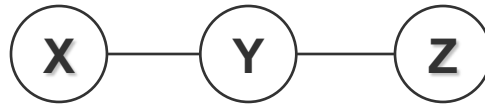
$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

- Also known as **Markov Random Fields**, **Markov networks** ...
- The **potential function** can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.





# Interpretation of Clique Potentials



- The model implies  $X \perp Z \mid Y$ . This independence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x | y)p(z | y)$$

- We can write this as:  
$$p(x, y, z) = p(x, y)p(z | y)$$
  
$$p(x, y, z) = p(x | y)p(z, y)$$
 , but

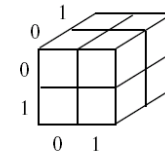
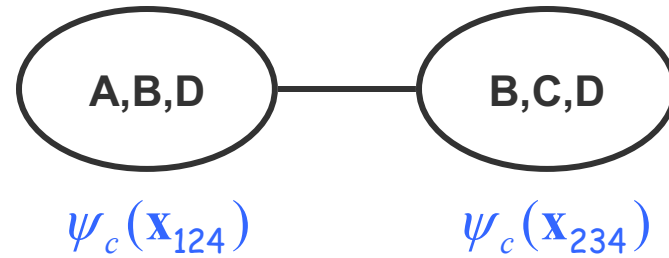
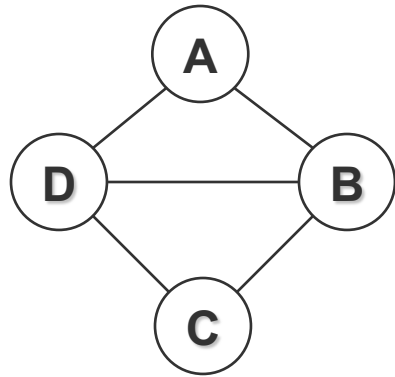
- **cannot** have all potentials be **marginals**
- **cannot** have all potentials be **conditionals**
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.







# Example UGM – using max cliques



$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

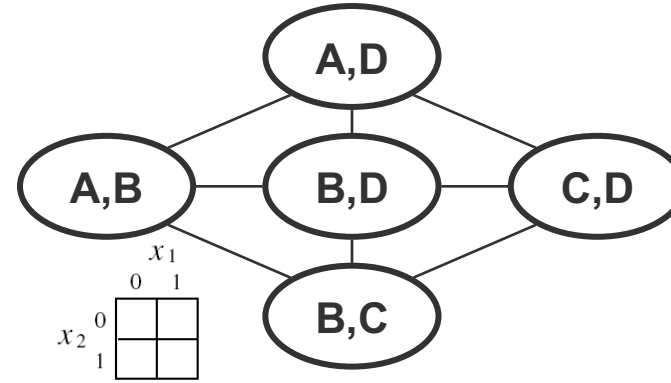
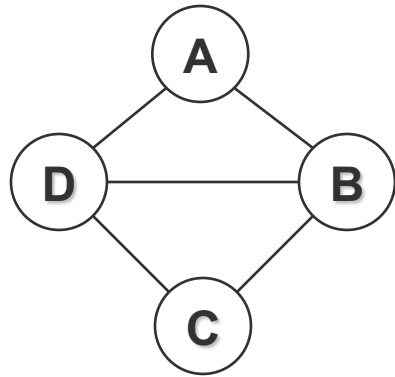
$$Z = \sum_{x_1, x_2, x_3, x_4} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

- For discrete nodes, we can represent  $P(X_{1:4})$  as two 3D tables instead of one 4D table





# Example UGM – using subcliques



$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{ij} \psi_{ij}(x_{ij})$$

$$= \frac{1}{Z} \psi_{12}(x_{12}) \psi_{14}(x_{14}) \psi_{23}(x_{23}) \psi_{24}(x_{24}) \psi_{34}(x_{34})$$

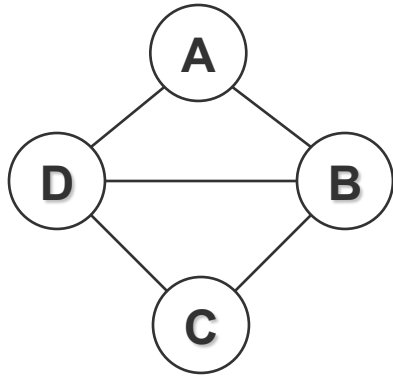
$$Z = \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(x_{ij})$$

- We can represent  $P(X_{1:4})$  as 5 2D tables instead of one 4D table
- Pair MRFs, a popular and simple special case





# Example UGM – canonical representation



$$\begin{aligned} P(x_1, x_2, x_3, x_4) \\ &= \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234}) \\ &\quad \times \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34}) \\ &\quad \times \psi_1(x_1) \psi_2(x_2) \psi_3(x_3) \psi_4(x_4) \end{aligned}$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \begin{aligned} &\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234}) \\ &\times \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34}) \\ &\times \psi_1(x_1) \psi_2(x_2) \psi_3(x_3) \psi_4(x_4) \end{aligned}$$

- Most general, subsume P' and P'' as special cases

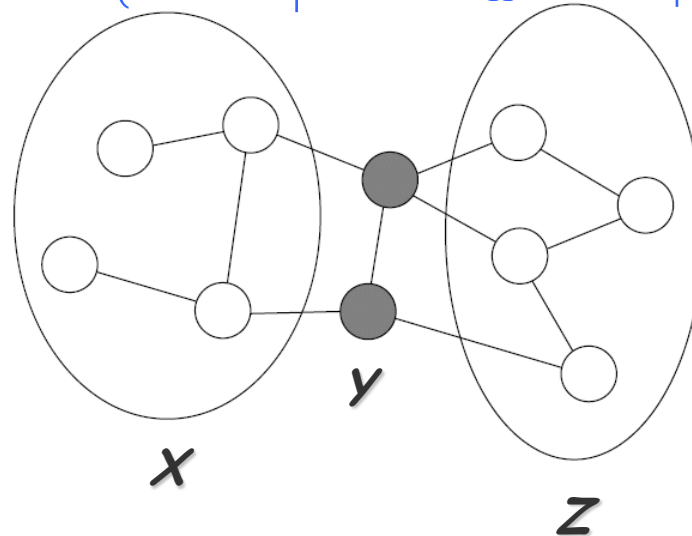




## II: Independence properties:

- Now let us ask what kinds of distributions, **in terms of the set of independence relationships between variables**, can be represented by undirected graphs (ignoring the details of the particular parameterization).
- Defn: the global Markov properties of a UG  $H$  are

$$I(H) = \{X \perp Z | Y : \text{sep}_H(X; Z | Y)\}$$





# I-maps

- Defn : Let  $P$  be a distribution over  $X$ . We define  $I(P)$  to be the set of independence assertions of the form  $(X \perp Y \mid Z)$  that hold in  $P$  (no matter how we set the parameter-values).
- Defn : Let  $K$  be *any graph object* associated with a set of independencies  $I(K)$ . We say that  $K$  is an *I-map* for a set of independencies  $I$ , if  $I(K) \subseteq I$ .
- We now say that  $G$  is an I-map for  $P$  if  $G$  is an I-map for  $I(P)$ , where we use  $I(G)$  as the set of independencies associated.





# Facts about I-map

- For  $G$  to be an I-map of  $P$ , it is necessary that  $G$  does not mislead us regarding independencies in  $P$ :

any independence that  $G$  asserts must also hold in  $P$ . Conversely,  $P$  may have additional independencies that are not reflected in  $G$

- (Perhaps unintuitive) Example:
  - Two variables:  $(X, Y)$
  - $P(X, Y) = P(X) P(Y)$
  - $G =$  full graph
  - Is  $G$  an I-map of  $P$ ?

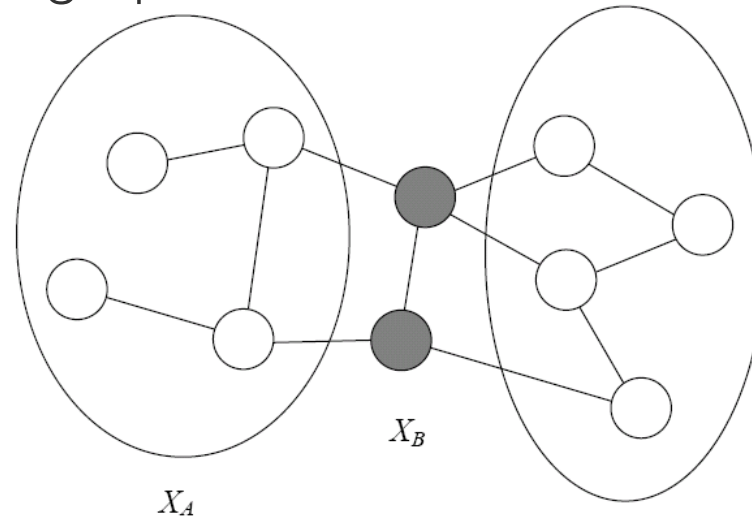






# Global Markov Independencies

- Let  $H$  be an undirected graph:



- $B$  *separates*  $A$  and  $C$  if every path from a node in  $A$  to a node in  $C$  passes through a node in  $B$ :

$$\text{sep}_H(A; C|B)$$

- A probability distribution satisfies the *global Markov property* if for any disjoint  $A, B, C$ , such that  $B$  separates  $A$  and  $C$ ,  $A$  is independent of  $C$  given  $B$ :

$$I(H) = \{A \perp C | B : \text{sep}_H(A; C|B)\}$$





# Local Markov independencies

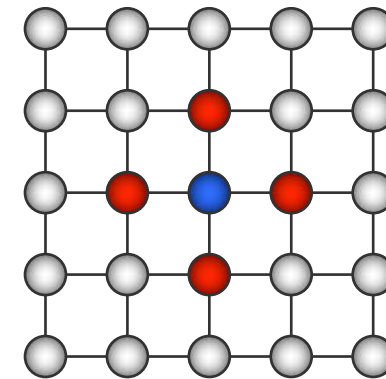
- For each node  $X_i \in \mathbf{V}$ , there is *unique Markov blanket* of  $X_i$ , denoted  $MB_{X_i}$ , which is the set of neighbors of  $X_i$  in the graph (those that share an edge with  $X_i$ )

- Defn:

The *local Markov independencies* associated with H is:

$$I_d(H): \{X_i \perp \mathbf{V} - \{X_i\} - MB_{X_i} \mid MB_{X_i} : \forall i\},$$

In other words,  $X_i$  is independent of the rest of the nodes in the graph given its immediate neighbors





# Soundness and completeness of global Markov property

- Defn: An UG  $H$  is an I-map for a distribution  $P$  if  $\mathcal{I}(H) \subseteq \mathcal{I}(P)$ , i.e.,  $P$  entails  $\mathcal{I}(H)$ .
- Defn:  $P$  is a **Gibbs distribution** over  $H$  if it can be represented as

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$$

- Thm (soundness): If  $P$  is a Gibbs distribution over  $H$ , then  $H$  is an I-map of  $P$ .
- Thm (completeness): If  $\neg \text{sep}_H(X; Z | Y)$ , then  $X \perp_P Z | Y$  in **some**  $P$  that factorizes over  $H$ .



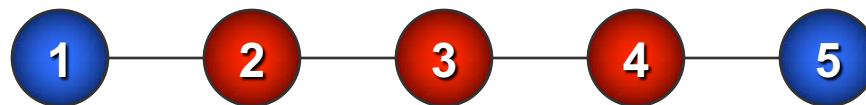


# Other Markov properties

- For undirected graphs, we defined I-maps in terms of global Markov properties, and will now derive local independence.
- For directed graphs, we defined I-maps in terms of local Markov properties, and derived global independence.
- Defn: The *pairwise Markov independencies* associated with UG  $H = (V, E)$  are

$$I_p(H) = \{X \perp Y | V \setminus \{X, Y\} : \{X, Y\} \notin E\}$$

- e.g.,  $X_1 \perp X_5 | \{X_2, X_3, X_4\}$





# Hammersley-Clifford Theorem

- If arbitrary potentials are utilized in the following product formula for probabilities,

$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

then the family of probability distributions obtained is exactly that set which **respects** the *qualitative specification* (the conditional independence relations) described earlier

- Thm : Let  $P$  be a **positive** distribution over  $\mathbf{V}$ , and  $H$  a Markov network graph over  $\mathbf{V}$ . If  $H$  is an I-map for  $P$ , then  $P$  is a **Gibbs distribution over  $H$** .





# Perfect maps

- Defn: A Markov network  $H$  is a perfect map for  $P$  if for any  $X; Y; Z$  we have that

$$\text{sep}_H(X; Z | Y) \Leftrightarrow P \models (X \perp Z | Y)$$

- Thm: not every distribution has a perfect map as UGM.
  - See next lecture for proof.







# Exponential Form

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential  $\psi_c(\mathbf{x}_c)$  in an unconstrained form using a real-value "energy" function  $\phi_c(\mathbf{x}_c)$ :

$$\psi_c(\mathbf{x}_c) = \exp\{-\phi_c(\mathbf{x}_c)\}$$

For convenience, we will call  $\phi_c(\mathbf{x}_c)$  a potential when no confusion arises from the context.

- This gives the joint a nice additive structure

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{-\sum_{c \in C} \phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z} \exp\{-H(\mathbf{x})\}$$

where the sum in the exponent is called the "free energy":

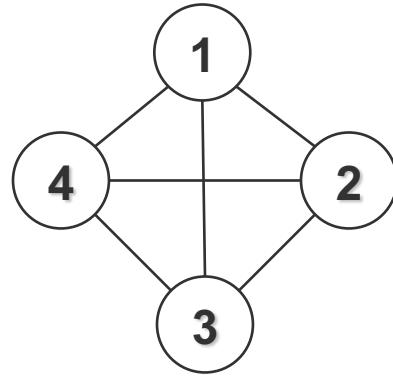
$$H(\mathbf{x}) = \sum_{c \in C} \phi_c(\mathbf{x}_c)$$

- In physics, this is called the "Boltzmann distribution".
- In statistics, this is called a log-linear model.





# Example: Boltzmann machines



- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for  $x_i \in \{-1,+1\}$  or  $x_i \in \{0,1\}$ ) is called a Boltzmann machine

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= \frac{1}{Z} \exp \left\{ \sum_{ij} \phi_{ij}(x_i, x_j) \right\} \\ &= \frac{1}{Z} \exp \left\{ \sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i + C \right\} \end{aligned}$$

- Hence the overall energy function has the form:

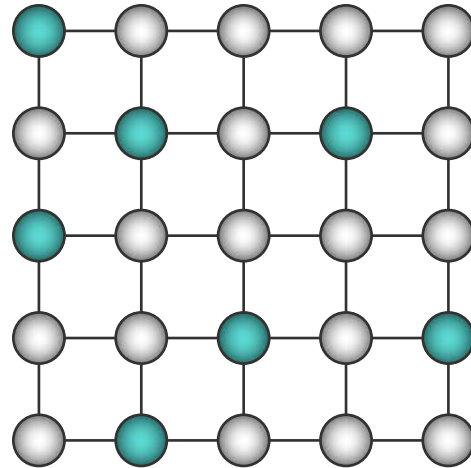
$$H(x) = \sum_{ij} (x_i - \mu) \Theta_{ij} (x_j - \mu) = (x - \mu)^T \Theta (x - \mu)$$





# Ising models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.



$$p(X) = \frac{1}{Z} \exp \left\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \right\}$$

- Same as sparse Boltzmann machine, where  $\theta_{ij} \neq 0$  iff  $i, j$  are neighbors.
  - e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.
- Potts model**: multi-state Ising model.

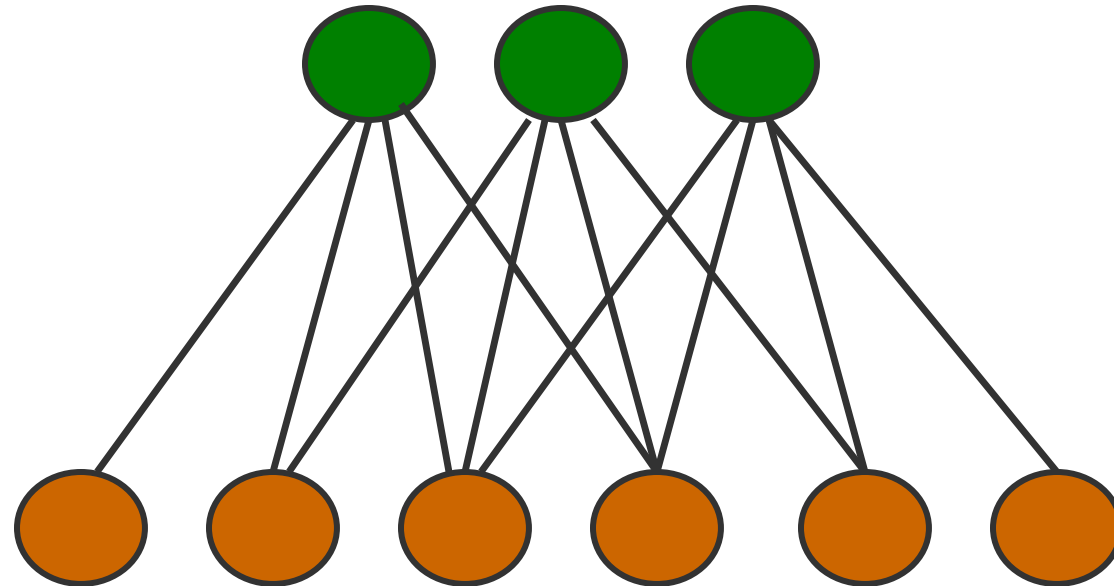




# Restricted Boltzmann Machines

hidden units

visible units



$$p(x, h | \theta) = \exp \left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\theta) \right\}$$



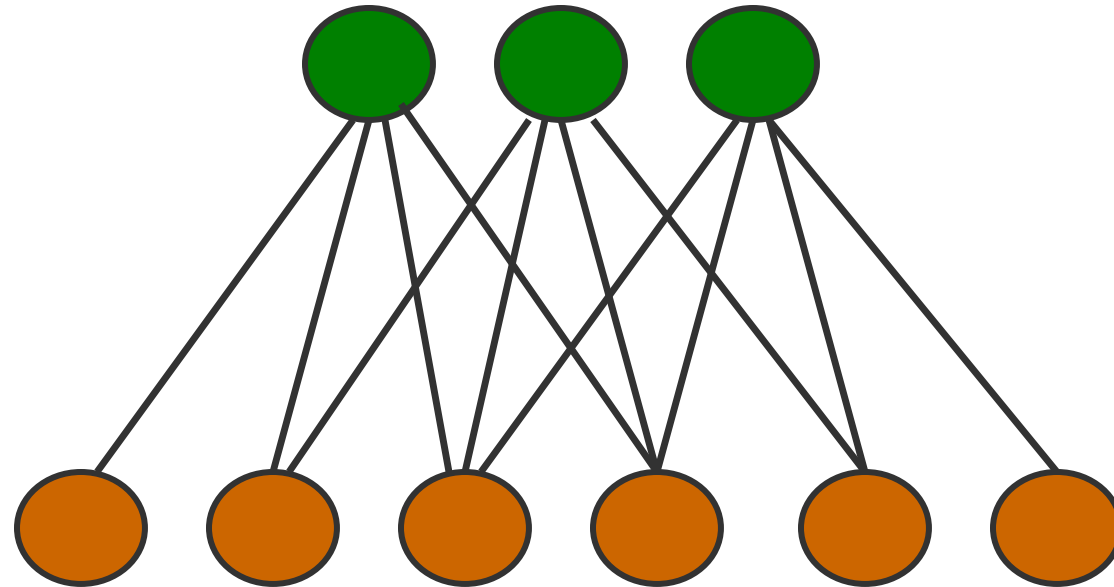


# Restricted Boltzmann Machines

## The Harmonium (Smolensky –'86)

hidden units

visible units



History:

Smolensky ('86), Proposed the architecture.

Freund & Haussler ('92), The “Combination Machine” (binary), learning with projection pursuit.

Hinton ('02), The “Restricted Boltzman Machine” (binary), learning with contrastive divergence.

Marks & Movellan ('02), Diffusion Networks (Gaussian).

Welling, Hinton, Osindero ('02), “Product of Student-T Distributions” (super-Gaussian)



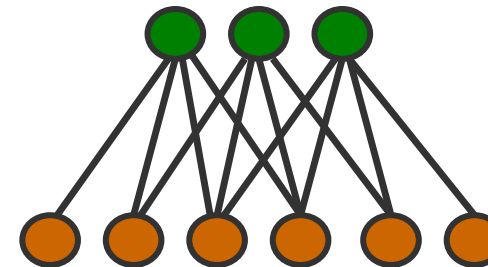
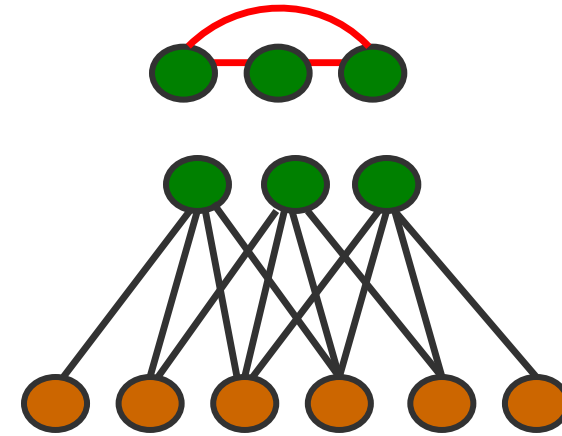


# Properties of RBM

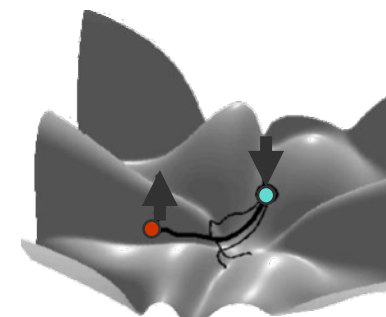
- Factors are marginally *dependent*.
- Factors are conditionally *independent* given observations on the visible nodes.

$$P(\ell | \mathbf{w}) = \prod_i P(\ell_i | \mathbf{w})$$

- Iterative Gibbs sampling.
- Learning with contrastive divergence



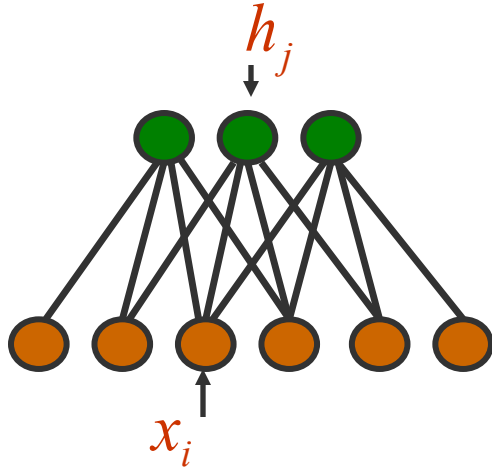
$$h \sim p(h | x)$$
$$x \sim p(x | h)$$



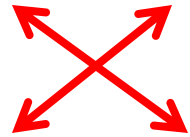




# A Constructive Definition



$$p_{\text{ind}}(\mathbf{h}) \propto \prod_j \exp\{ \theta_j g_j(h_j) \}$$



how do we couple them?

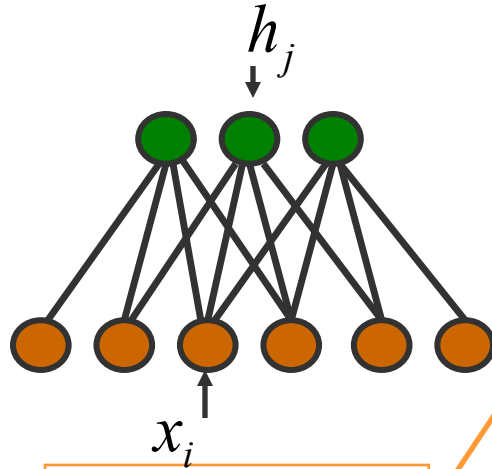
$$p_{\text{ind}}(\mathbf{x}) \propto \prod_i \exp\{ \theta_i f_i(x_i) \}$$

$$p(x, h | \theta) = \exp\left\{ \sum_i \bar{\theta}_i \vec{f}_i(x_i) + \sum_j \vec{\lambda}_j \vec{g}_j(h_j) + \sum_{i,j} \vec{f}_i^T(x_i) \mathbf{W}_{i,j} \vec{g}_j(h_j) \right\}$$





# A Constructive Definition



coupling in the log-domain with shifted parameters

vector of local sufficient statistics (features)

$$p(\mathbf{x} | \mathbf{h}) = \prod_i p(x_i | \mathbf{h}),$$

$$p(x_i | \mathbf{h}) = \exp\left\{ \sum_a \hat{\theta}_{ia} f_{ia}(x_i) + A_i(\{\hat{\theta}_{ia}\}) \right\}$$

$$\hat{\theta}_{ia} = \theta_{ia} + \sum_{jb} W_{ia}^{jb} g_{jb}(h_j) = \theta_{ia} + \sum_j \vec{W}_{ia}^j \vec{g}_j(h_j)$$

$$p(\mathbf{h} | \mathbf{x}) = \prod_j p(h_j | \mathbf{x})$$

$$p(h_j | \mathbf{x}) = \exp\left\{ \sum_b \hat{\lambda}_{jb} g_{jb}(h_j) + B_j(\{\hat{\lambda}_{jb}\}) \right\}$$

$$\hat{\lambda}_{jb} = \lambda_{jb} + \sum_{ia} W_{ia}^{jb} f_{ia}(x_i) = \lambda_{jb} + \sum_i \vec{W}_i^{jb} \vec{f}_i(x_i)$$

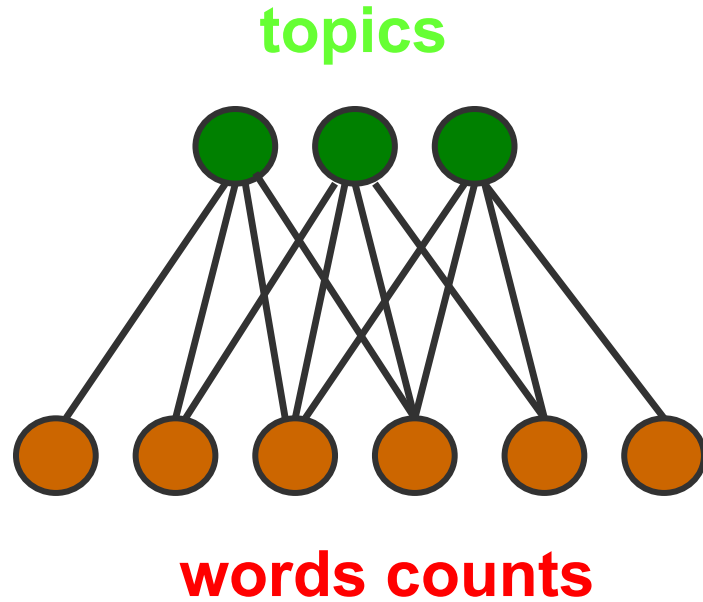
They map to the RBM random field:

$$p(x, h | \theta) = \exp\left\{ \sum_i \bar{\theta}_i \vec{f}_i(x_i) + \sum_j \bar{\lambda}_j \vec{g}_j(h_j) + \sum_{i,j} \vec{f}_i^T(x_i) \mathbf{W}_{i,j} \vec{g}_j(h_j) \right\}$$





# An RBM for Text Modeling



$h_j = 3$ : topic  $j$  has strength 3

$$h_j \in \mathbf{R}, \quad \langle h_j \rangle = \sum_i W_{i,j} x_i$$

$x_i = n$ : word  $i$  has count  $n$

$$x_i \in \mathbf{I}$$

$$p(\mathbf{h} | \mathbf{x}) = \prod_j \text{Normal}_{h_j} \left[ \sum_i \vec{W}_{ij} \vec{x}_i, 1 \right]$$

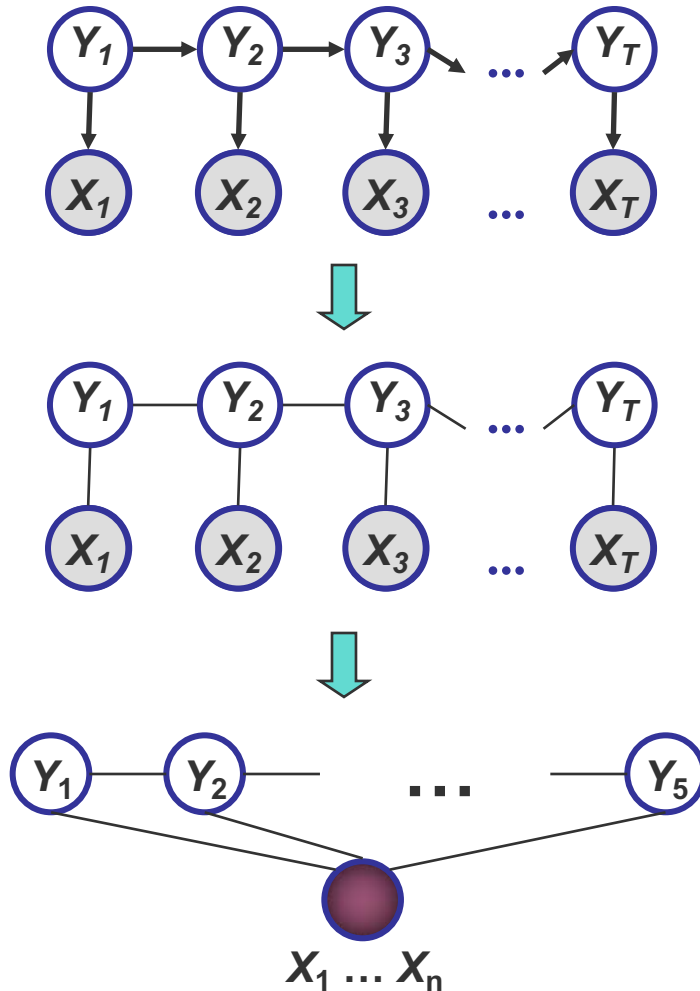
$$p(\mathbf{x} | \mathbf{h}) = \prod_i \text{Bi}_{x_i} \left[ N, \frac{\exp(\alpha_j + \sum_j W_{ij} h_j)}{1 + \exp(\alpha_j + \sum_j W_{ij} h_j)} \right]$$

$$\Rightarrow p(\mathbf{x}) \propto \exp \left\{ \left( \sum_i \alpha_i x_i - \log \Gamma(x_i) - \log \Gamma(N - x_i) \right) + \frac{1}{2} \sum_j \left( \sum_i W_{i,j} x_i \right)^2 \right\}$$





# Conditional Random Fields



- Discriminative

$$p_{\theta}(y | x) = \frac{1}{Z(\theta, x)} \exp \left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Doesn't assume that features are independent
- When labeling  $X_i$  future observations are taken into account





# Conditional Models

- Conditional probability  $P(\text{label sequence } \mathbf{y} \mid \text{observation sequence } \mathbf{x})$  rather than joint probability  $P(\mathbf{y}, \mathbf{x})$ 
  - Specify the probability of possible label sequences given an observation sequence
- Allow arbitrary, non-independent features on the observation sequence  $\mathbf{X}$
- The probability of a transition between labels may depend on **past** and **future** observations
- Relax strong independence assumptions in generative models



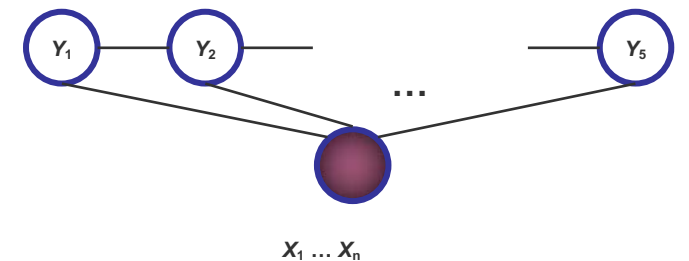


# Conditional Distribution

- If the graph  $G = (V, E)$  of  $\mathbf{Y}$  is a tree, the conditional distribution over the label sequence  $\mathbf{Y} = \mathbf{y}$ , given  $\mathbf{X} = \mathbf{x}$ , by the Hammersley Clifford theorem of random fields is:

$$p_{\theta}(\mathbf{y} | \mathbf{x}) \propto \exp \left( \sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, \mathbf{x}) \right)$$

- $\mathbf{x}$  is a data sequence
- $\mathbf{y}$  is a label sequence
- $v$  is a vertex from vertex set  $V$  = set of label random variables
- $e$  is an edge from edge set  $E$  over  $V$
- $f_k$  and  $g_k$  are given and fixed.  $g_k$  is a Boolean vertex feature;  $f_k$  is a Boolean edge feature
- $k$  is the number of features
- $\theta = (\lambda_1, \lambda_2, \dots, \lambda_n; \mu_1, \mu_2, \dots, \mu_n)$ ;  $\lambda_k$  and  $\mu_k$  are parameters to be estimated
- $\mathbf{y}|_e$  is the set of components of  $\mathbf{y}$  defined by edge  $e$
- $\mathbf{y}|_v$  is the set of components of  $\mathbf{y}$  defined by vertex  $v$





## Conditional Distribution (cont'd)

- CRFs use the observation-dependent normalization  $Z(\mathbf{x})$  for the conditional distributions:

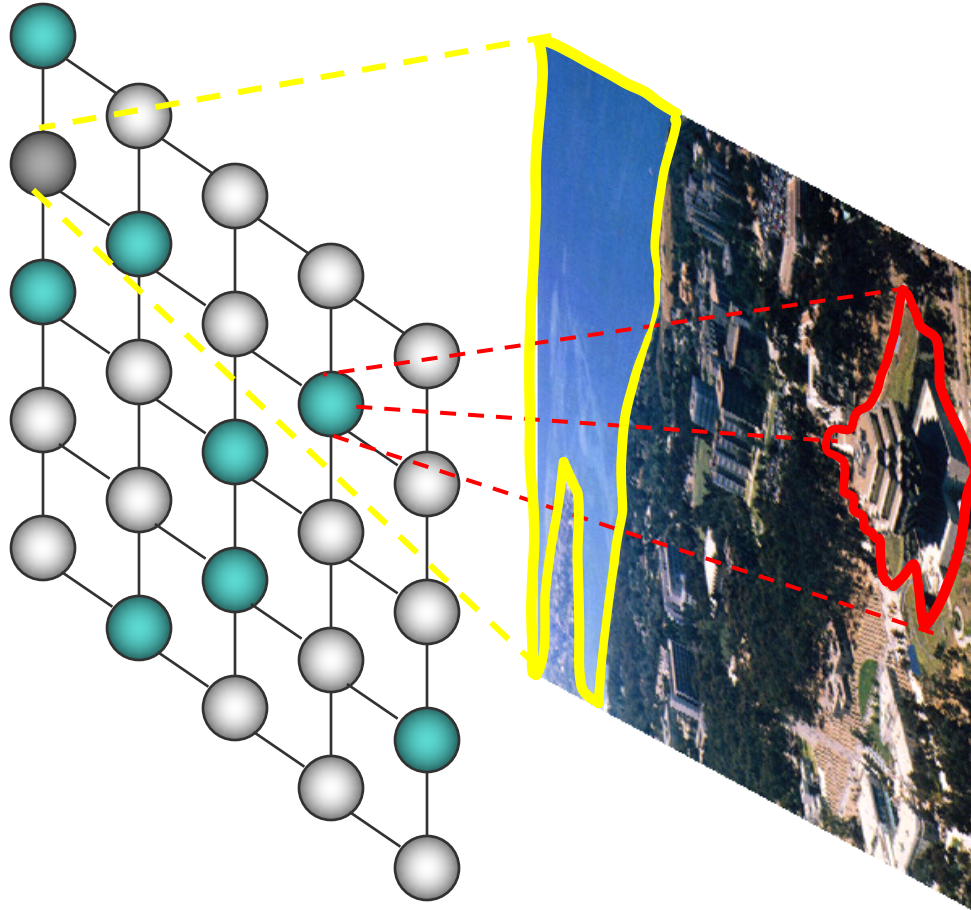
$$p_{\theta}(y | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left( \sum_{e \in E, k} \lambda_k f_k(e, y | e, \mathbf{x}) + \sum_{v \in V, k} \mu_k g_k(v, y | v, \mathbf{x}) \right)$$

- $Z(\mathbf{x})$  is a normalization over the data sequence  $\mathbf{x}$





# Conditional Random Fields



$$p_{\theta}(y | x) = \frac{1}{Z(\theta, x)} \exp \left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Allow arbitrary dependencies on input
- Clique dependencies on labels
- Use approximate inference for general graphs

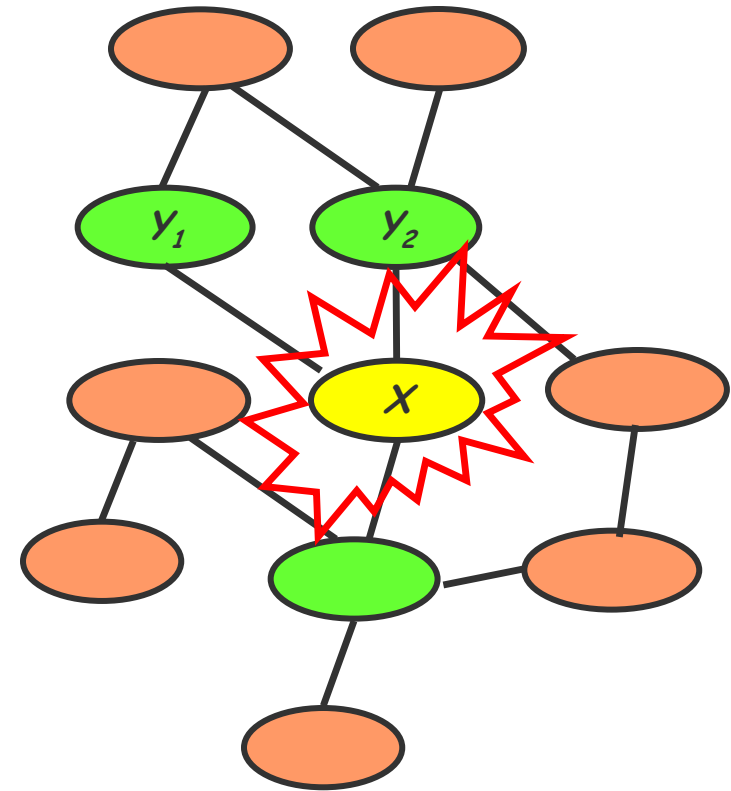






# Summary: Conditional Independence Semantics in an MRF

- Structure: an *undirected graph*
  - Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**
  - Local contingency functions (**potentials**) and the **cliques in the graph** completely determine the **joint dist.**
  - Give **correlations** between variables, but no explicit way to generate samples





# Summary

- Undirected graphical models capture “relatedness”, “coupling”, “co-occurrence”, “synergism”, etc. between entities
  - Local and global independence properties identifiable via graph separation criteria
  - Defined on clique potentials
- Can be used to define either joint or conditional distributions
- Generally intractable to compute likelihood due to presence of “partition function”
  - Therefore not only inference, but also likelihood-based learning is difficult in general
- Important special cases:
  - Ising models
  - RBM
  - CRF
- Learning GM structures:
  - the Chow-Liu Algorithm





# Supplementary:



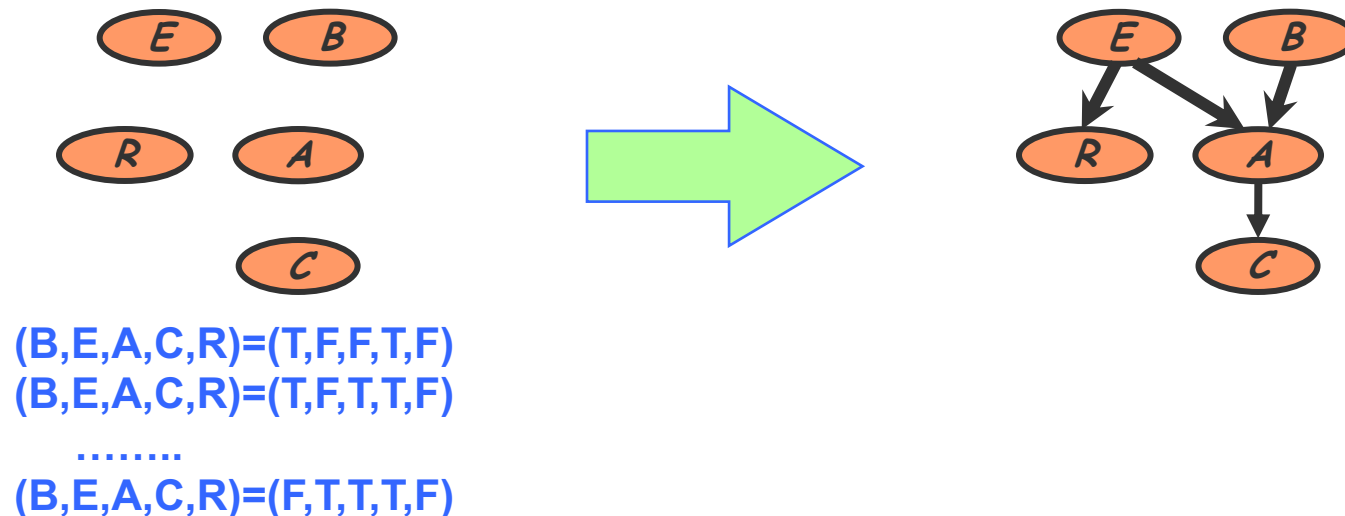


# Where is the graph structure come from?

## The goal:

- Given set of independent samples (*assignments* of random variables), find the *best* (the most likely?) graphical model topology

ML Structural Learning for completely observed GMs





# Information Theoretic Interpretation of ML

$$\begin{aligned}\ell(\theta_G, G; D) &= \log p(D | \theta_G, G) \\ &= \log \prod_n \left( \prod_i p(x_{n,i} | \mathbf{x}_{n,\pi_i(G)}, \theta_{i|\pi_i(G)}) \right) \\ &= \sum_i \left( \sum_n \log p(x_{n,i} | \mathbf{x}_{n,\pi_i(G)}, \theta_{i|\pi_i(G)}) \right) \\ &= M \sum_i \left( \sum_{x_i, \mathbf{x}_{\pi_i(G)}} \frac{\text{count}(x_i, \mathbf{x}_{\pi_i(G)})}{M} \log p(x_i | \mathbf{x}_{\pi_i(G)}, \theta_{i|\pi_i(G)}) \right) \\ &= M \sum_i \left( \sum_{x_i, \mathbf{x}_{\pi_i(G)}} \hat{p}(x_i, \mathbf{x}_{\pi_i(G)}) \log p(x_i | \mathbf{x}_{\pi_i(G)}, \theta_{i|\pi_i(G)}) \right)\end{aligned}$$

**From sum over data points to sum over count of variable states**





# Information Theoretic Interpretation of ML (con'd)

$$\begin{aligned}\ell(\theta_G, G; D) &= \log \hat{p}(D | \theta_G, G) \\ &= M \sum_i \left( \sum_{x_i, \mathbf{x}_{\pi_i(G)}} \hat{p}(x_i, \mathbf{x}_{\pi_i(G)}) \log \hat{p}(x_i | \mathbf{x}_{\pi_i(G)}, \theta_{i|\pi_i(G)}) \right) \\ &= M \sum_i \left( \sum_{x_i, \mathbf{x}_{\pi_i(G)}} \hat{p}(x_i, \mathbf{x}_{\pi_i(G)}) \log \frac{\hat{p}(x_i, \mathbf{x}_{\pi_i(G)}, \theta_{i|\pi_i(G)}) \hat{p}(x_i)}{\hat{p}(\mathbf{x}_{\pi_i(G)}) \hat{p}(x_i)} \right) \\ &= M \sum_i \left( \sum_{x_i, \mathbf{x}_{\pi_i(G)}} \hat{p}(x_i, \mathbf{x}_{\pi_i(G)}) \log \frac{\hat{p}(x_i, \mathbf{x}_{\pi_i(G)}, \theta_{i|\pi_i(G)})}{\hat{p}(\mathbf{x}_{\pi_i(G)}) \hat{p}(x_i)} \right) - M \sum_i \left( \sum_{x_i} \hat{p}(x_i) \log \hat{p}(x_i) \right) \\ &= M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)}) - M \sum_i \hat{H}(x_i)\end{aligned}$$

**Decomposable score and a function of the graph structure**





# Structural Search

- How many graphs over  $n$  nodes?  $O(2^{n^2})$
- How many trees over  $n$  nodes?  $O(n!)$
- But it turns out that we can find exact solution of an optimal tree (under MLE)!
  - Trick: in a tree each node has only one parent!
  - Chow-liu algorithm





# Chow-Liu tree learning algorithm

- Objection function:

$$\begin{aligned} \ell(\theta_G, G; D) &= \log \hat{p}(D | \theta_G, G) \\ &= M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)}) - M \sum_i \hat{H}(x_i) \end{aligned} \Rightarrow \boxed{C(G) = M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)})}$$

- Chow-Liu:

- For each pair of variable  $x_i$  and  $x_j$ 
  - Compute empirical distribution:
  - Compute mutual information:
- Define a graph with node  $x_1, \dots, x_n$ 
  - Edge (i,j) gets weight

$$\hat{p}(X_i, X_j) = \frac{\text{count}(x_i, x_j)}{M}$$

$$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{p}(x_i, x_j) \log \frac{\hat{p}(x_i, x_j)}{\hat{p}(x_i) \hat{p}(x_j)}$$

$$\hat{I}(X_i, X_j)$$







# Chow-Liu algorithm (con'd)

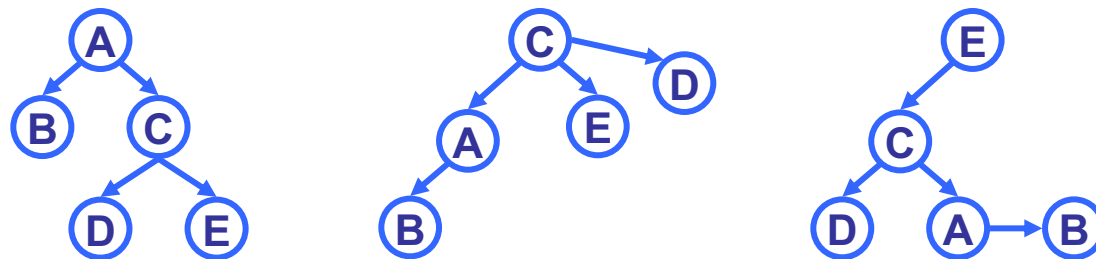
- Objection function:

$$\begin{aligned} \ell(\theta_G, G; D) &= \log \hat{p}(D | \theta_G, G) \\ &= M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)}) - M \sum_i \hat{H}(x_i) \end{aligned} \Rightarrow \boxed{C(G) = M \sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)})}$$

- Chow-Liu:

Optimal tree BN

- Compute maximum weight spanning tree
- Direction in BN: pick any node as root, do breadth-first-search to define directions
- I-equivalence:



$$C(G) = I(A, B) + I(A, C) + I(C, D) + I(C, E)$$





# Structure Learning for general graphs

- Theorem:
  - The problem of learning a BN structure with at most  $d$  parents is NP-hard for any (fixed)  $d \geq 2$
- Most structure learning approaches use heuristics
  - Exploit score decomposition
  - Two heuristics that exploit decomposition in different ways
    - Greedy search through space of node-orders
    - Local search of graph structures

