Representation of undirected GM

Kayhan Batmanghelich

Review

Review: Directed Graphical Model

- Represent distribution of the form

Parents of X_i

$$p(X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i | \pi(X_i))$$

- Factorizes in terms of local conditional probabilities

- Each node has to maintain $p(X_i | \pi(X_i))$

- Each variable is Conditional Independent of its non-descendants given its parents

Parents of X_i

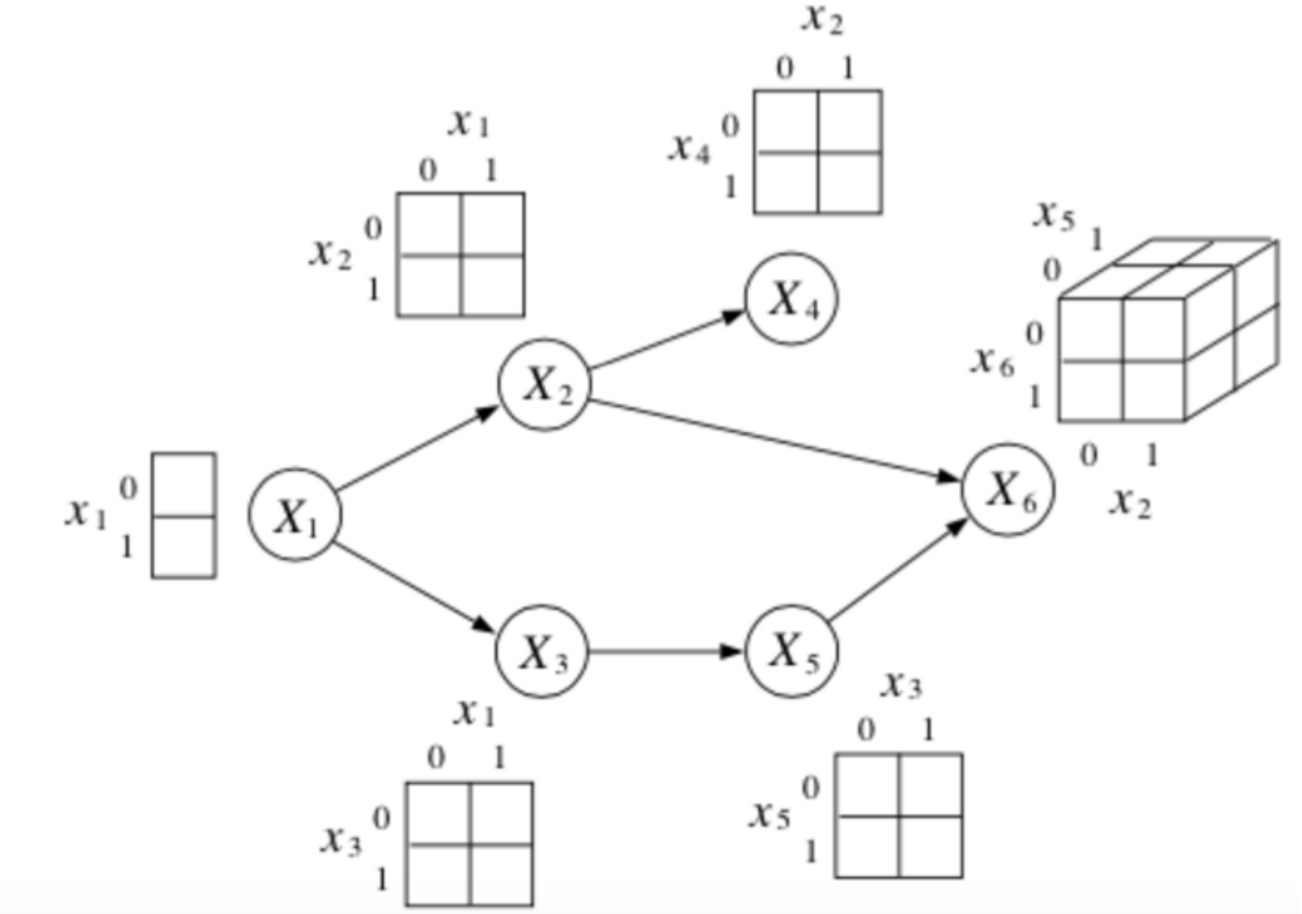
the nodes before X_i that are not its parents X_i

$$X_i \perp\!\!\!\perp \tilde{\pi}(X_i) | \pi(X_i)$$

- Such an ordering is a "topological" ordering (i.e., parents have lower numbers than their children)

Review: Directed Graphical Model

For discrete variables, each node stores a conditional probability table (CPT)

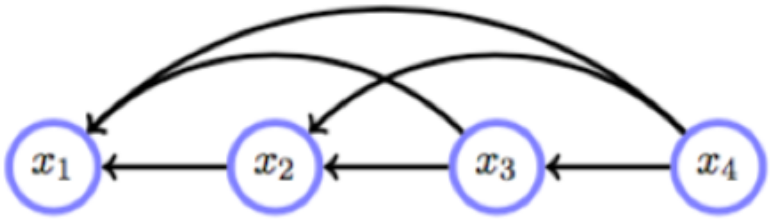


Review: independence properties of DAGs

- **Defn**: let $\mathcal{I}_l(\mathcal{G})$ be the set of local independence properties encoded by DAG $\mathcal{G}$, namely:

$$\mathcal{I}_\ell(\mathcal{G}) = \{X \perp\!\!\!\perp Z | dsep_{\mathcal{G}}(X; Z|Y)\}$$

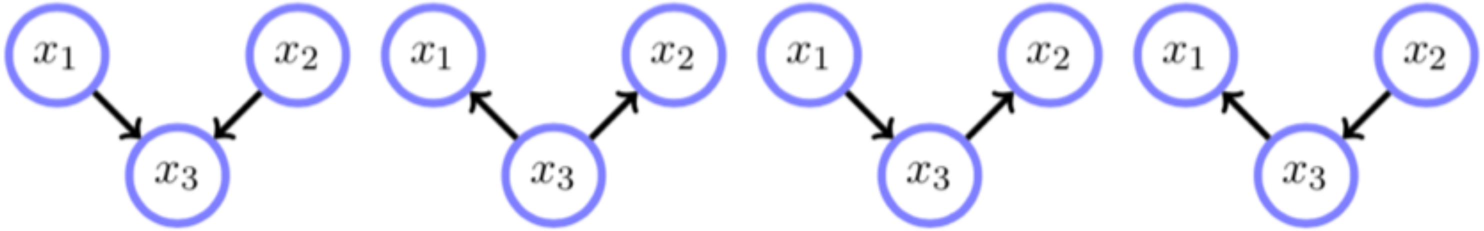
- **Defn**: A DAG $\mathcal{G}$ is an **I-map** (independence-map) of P if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$
- A fully connected DAG $\mathcal{G}$ is an I-map for any distribution, since $\mathcal{I}_l(\mathcal{G}) = \emptyset \subseteq \mathcal{I}(P)$ for any P.



$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1|x_2, \ldots, x_n)p(x_2, \ldots, x_n) \\ &= p(x_1|x_2, \ldots, x_n)p(x_2|x_3, \ldots, x_n)p(x_3, \ldots, x_n) \\ &= p(x_n) \prod_{i=1}^{n-1} p(x_i|x_{i+1}, \ldots, x_n) \end{aligned}$$

Review: I-equivalence

- Which graphs satisfy $\mathcal{I}(\mathcal{G}) = \{x_1 \perp\!\!\!\perp x_2 | x_3\}$?

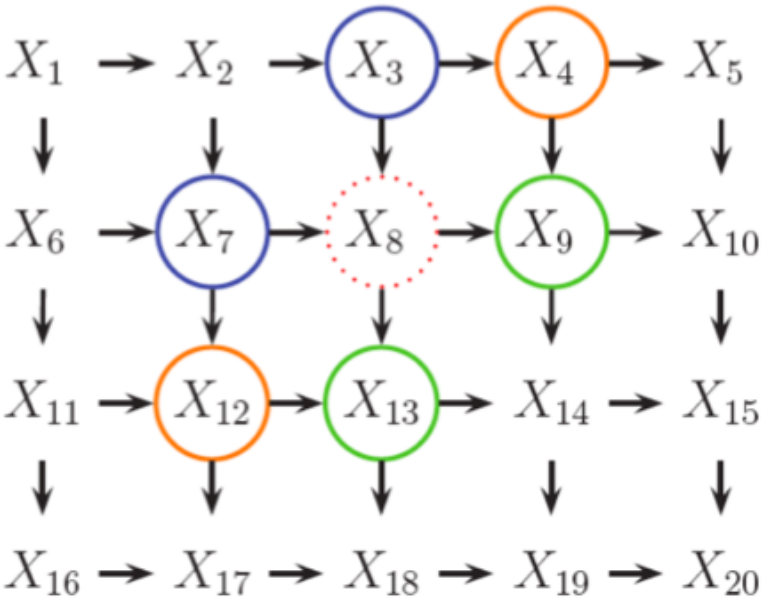


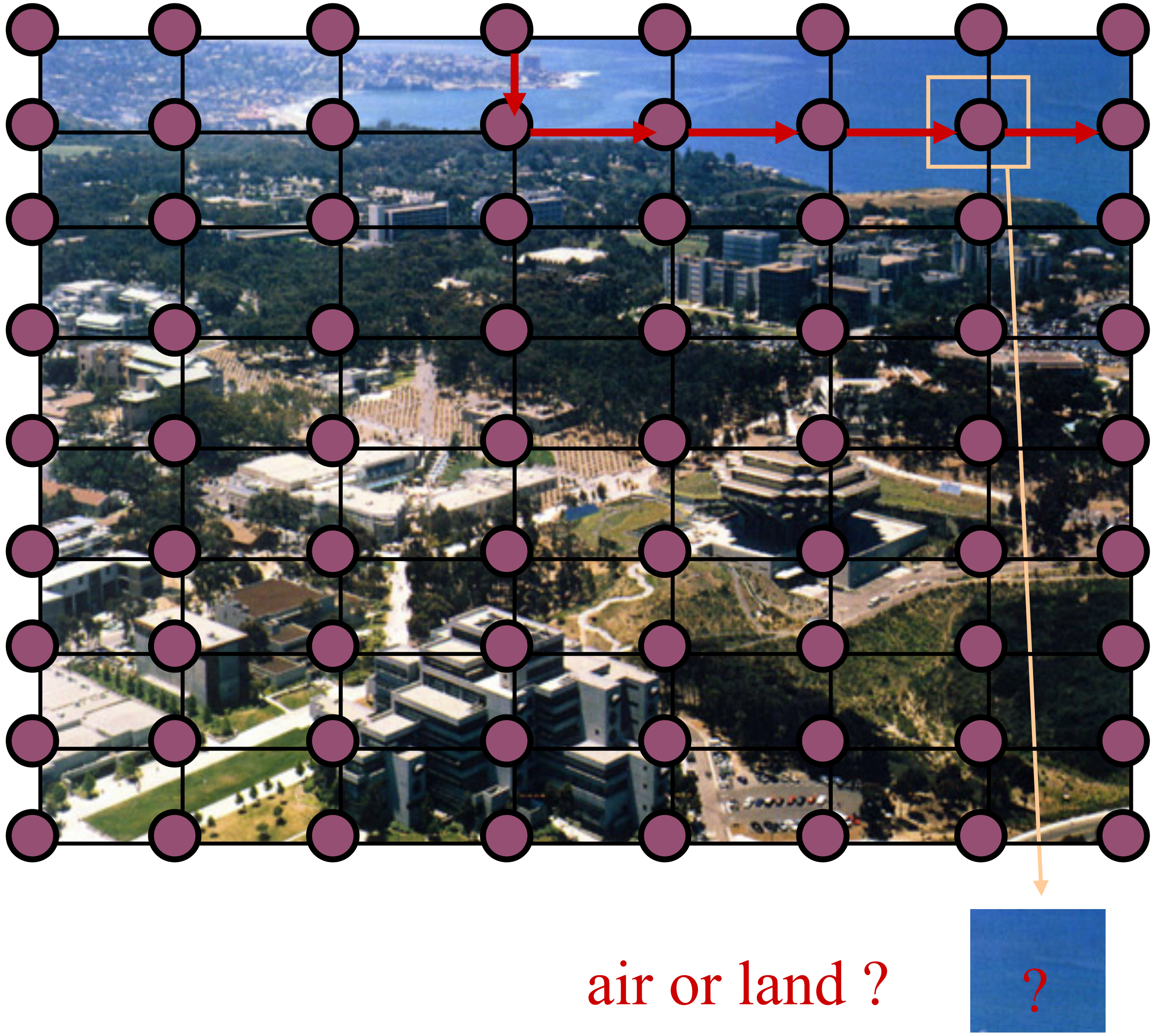
(a) (b) (c) (d)

- **Defn :** The *skeleton* of a Bayesian network graph G over $\boldsymbol{V}$ is an undirected graph over $\boldsymbol{V}$ that contains an edge $\{X, Y\}$ for every edge (X, Y) in G.

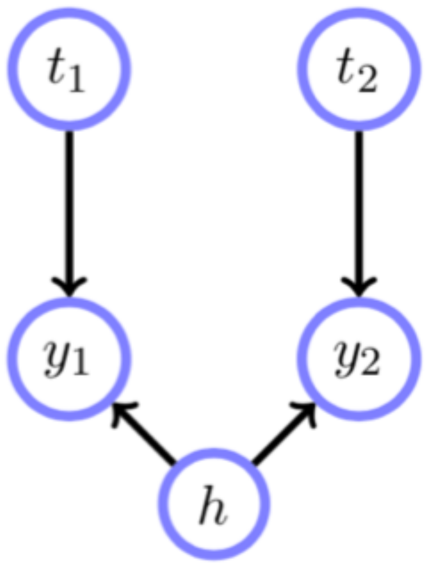
Why Undirected GM?

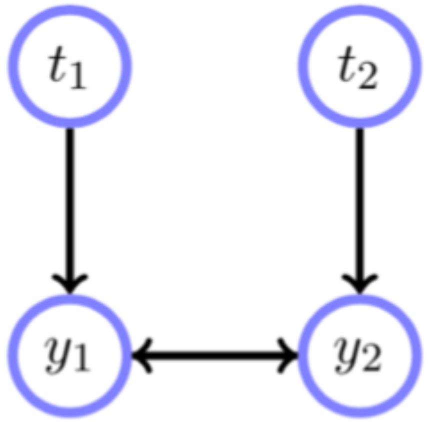
DGM is not always a good choice...





DGM is not always a good choice…

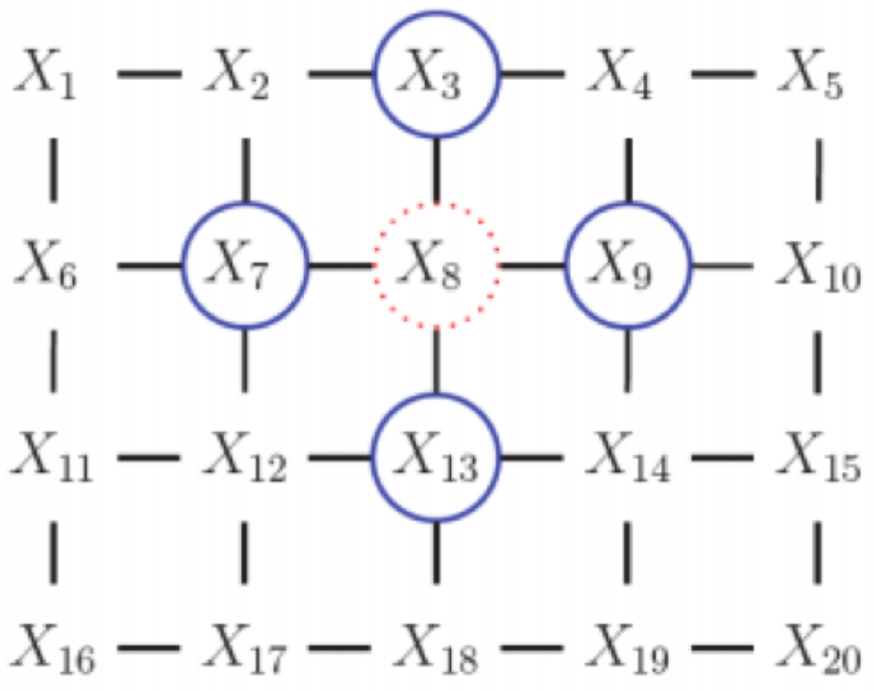




What if we cannot observe h ?

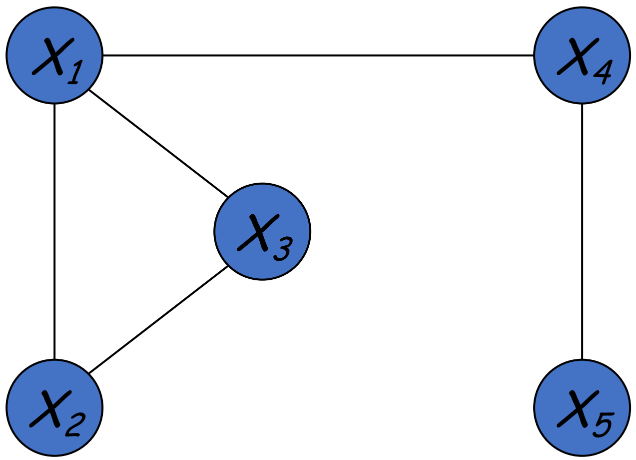
Undirected Graphical Models (UGM)

- As in DGM, the nodes in the graph represent the variables
- Edges represent probabilistic interaction between neighboring variables
- Parametrization?
  - In DGM we used CPD (conditional probabilities) to represent distribution of a node given others
  - For undirected graphs, we use a more symmetric parameterization that captures the affinities between related variables.
  - Differences:
    - Pairwise (non-causal) relationships
    - No explicit way to generate samples



What is UGM?

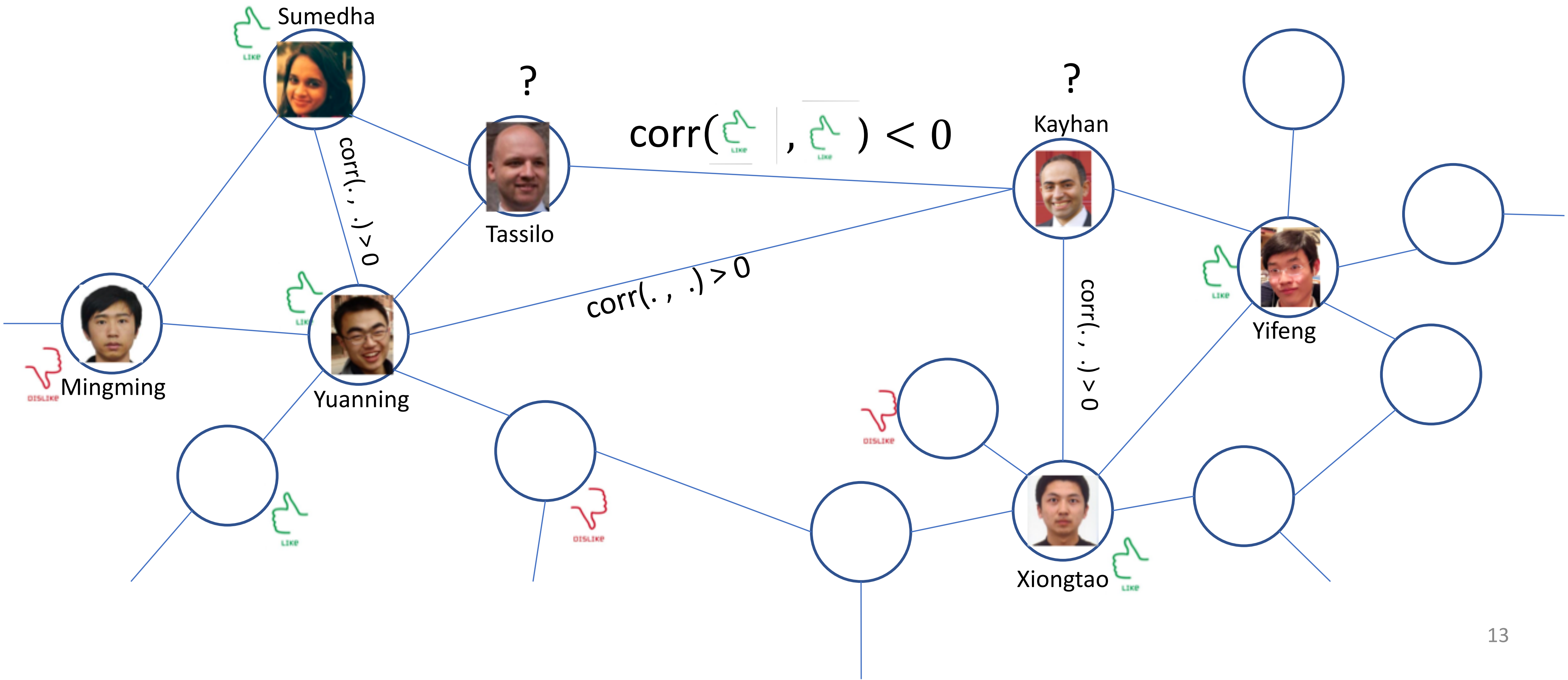
Undirected graphical models (UGM)



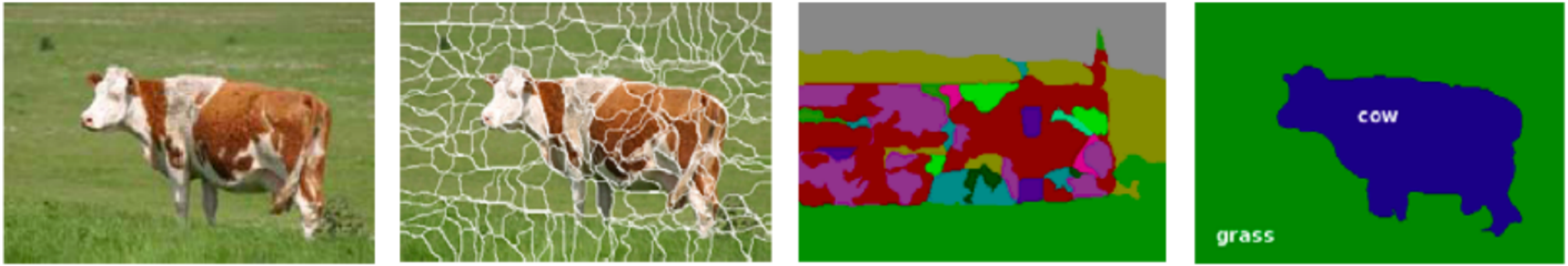
- Pairwise (non-causal) relationships
- Can write down model, and score specific configurations of the graph, but no explicit way to generate samples
- Contingency constrains on node configurations

Social networks

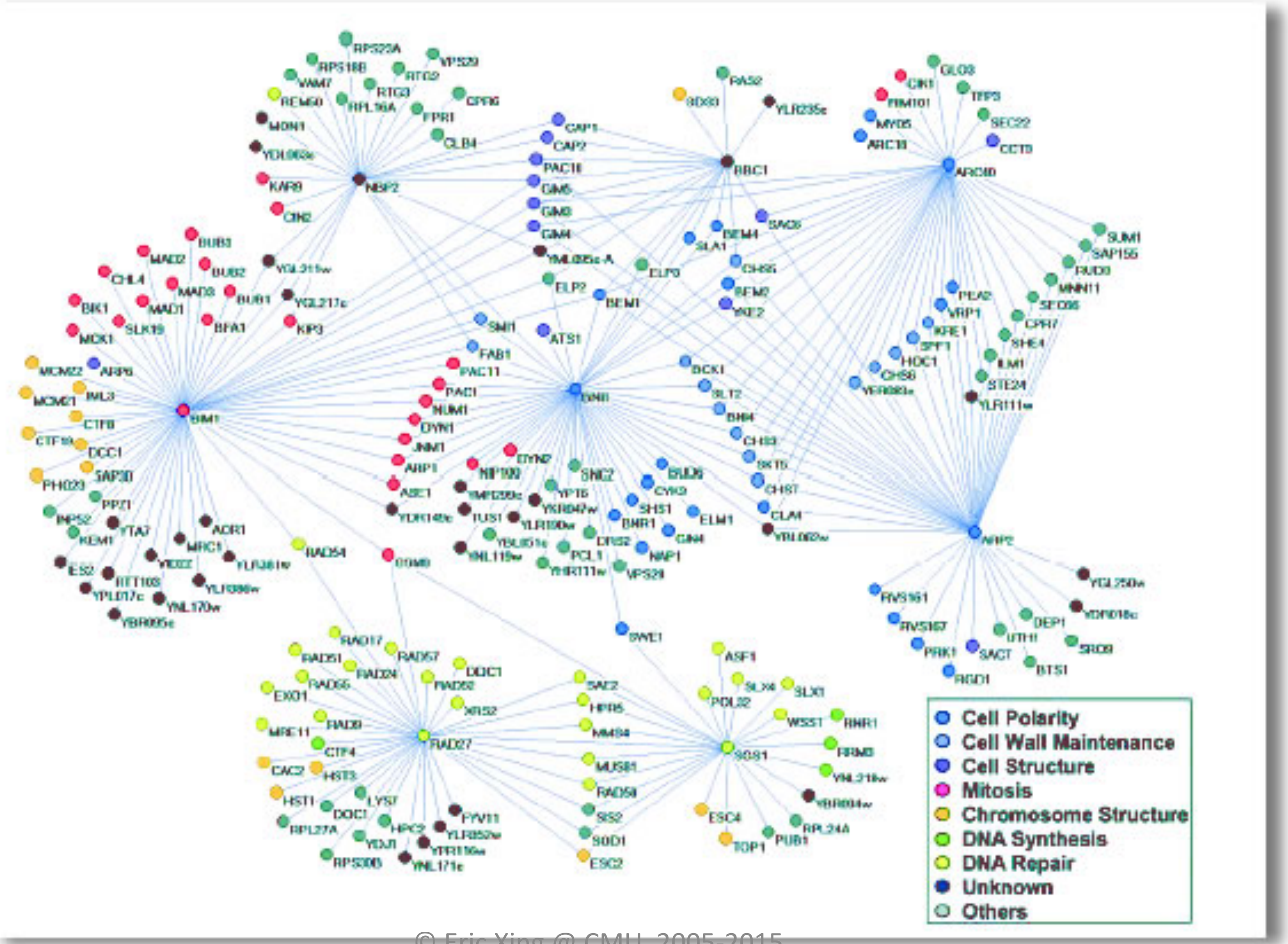
Did you like HW0?
Links represent correlation between members.



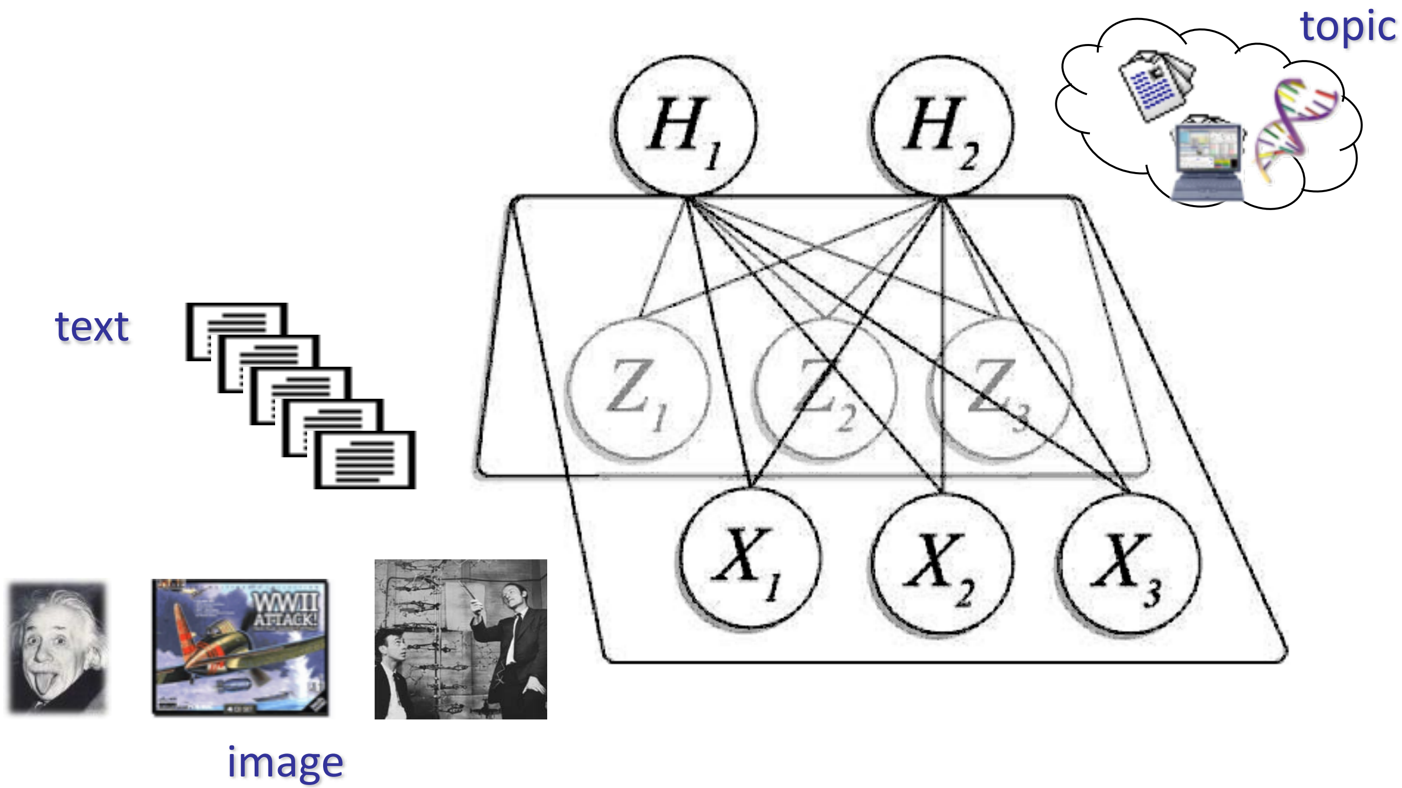
A Canonical Example: understanding complex scene …



Protein interaction networks



Information retrieval



Undirected graphical models (UGM)

Defn (also called Markov Network): For a set of variables $\mathcal{X} = \{x_1, \cdots, x_n\}$ a Markov network is defined as a product of potentials on subsets of the variables $\mathcal{X}_c \subseteq \mathcal{X}$

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

Normalizer to ensure it is a p is a probability

This is called potential ≥ 0 (this does not have to be probability)

Maximal clique

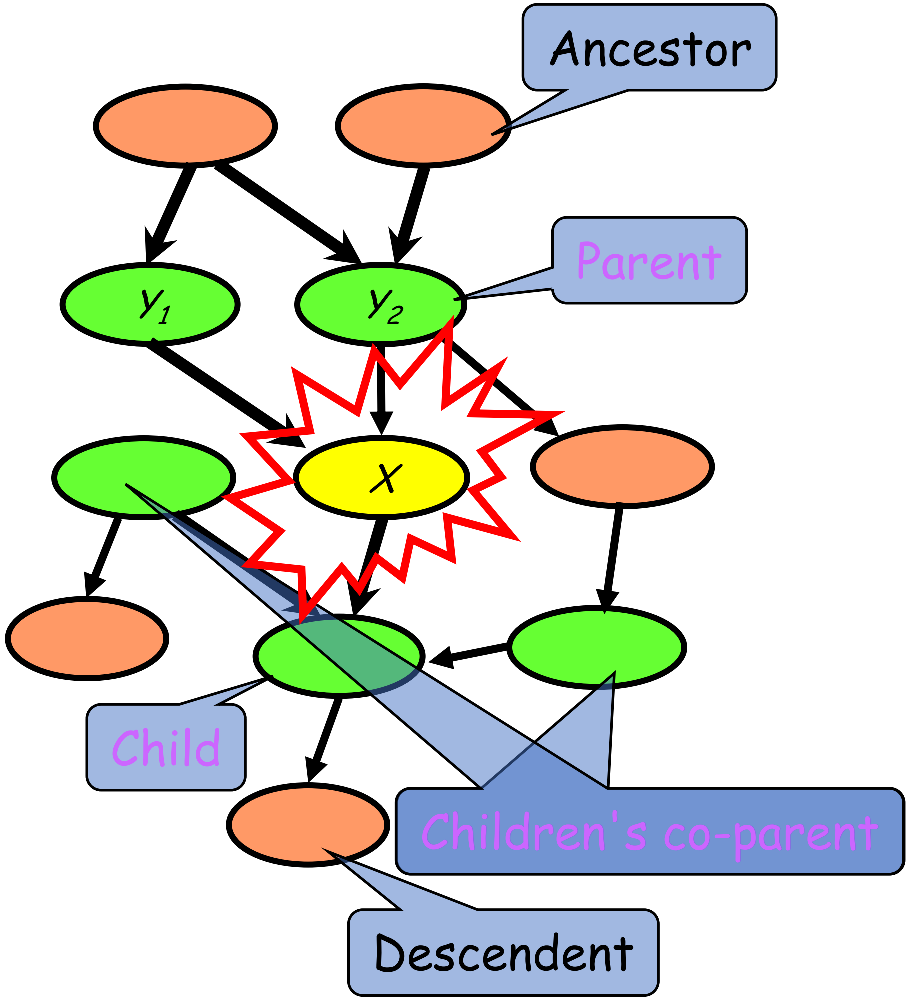
Def: A maximal clique is a clique that cannot be extended by including one more adjacent vertex, meaning it is not a subset of a larger clique.

Independence

Remember the Markov Blanket for BN

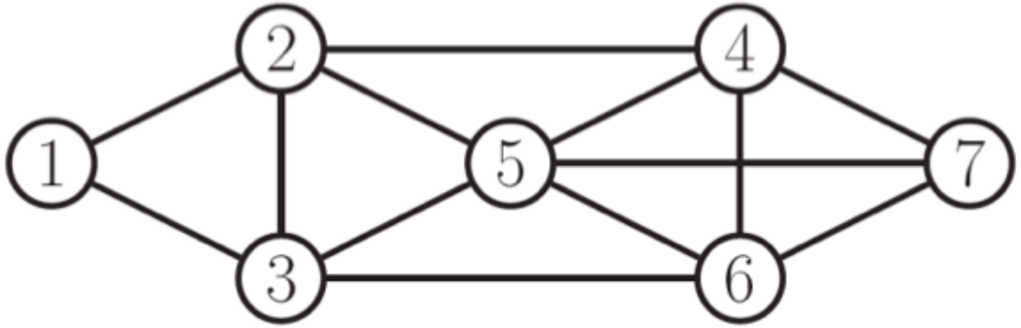
Structure: *DAG*

- Meaning: a node is conditionally independent of every other node in the network outside its Markov blanket



About Conditional Independence

Global Markov Property: $X_A \perp\!\!\!\perp X_B | X_C$ if and only if C separates A from B (there is no path connecting them)



$$\{1,2\} \perp \{6,7\} | \{3,4,5\}$$

Markov Blanket (local property) is the set of nodes that renders a node t conditionally independent of all the other nodes in the graph

$$t \perp\!\!\!\perp \mathcal{V} - mb(t) - \{t\} | mb(t)$$

All nodes in the graph

Markov Blanket

$$mb(5) = \{2,3,4,6,7\}$$

Example of Dependencies

Pairwise: $1 \perp 7 | \text{rest}$

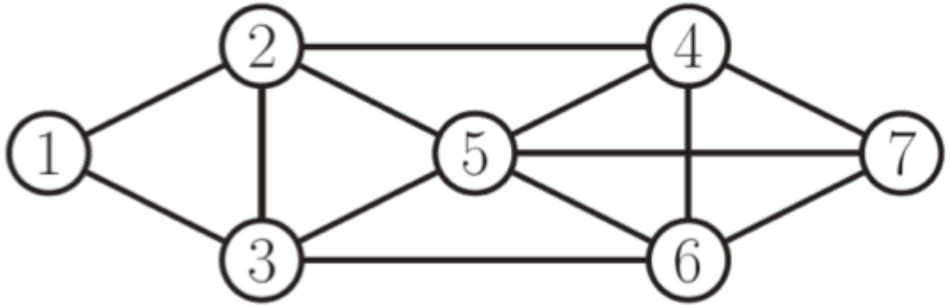
Local: $1 \perp \text{rest} | 2, 3$

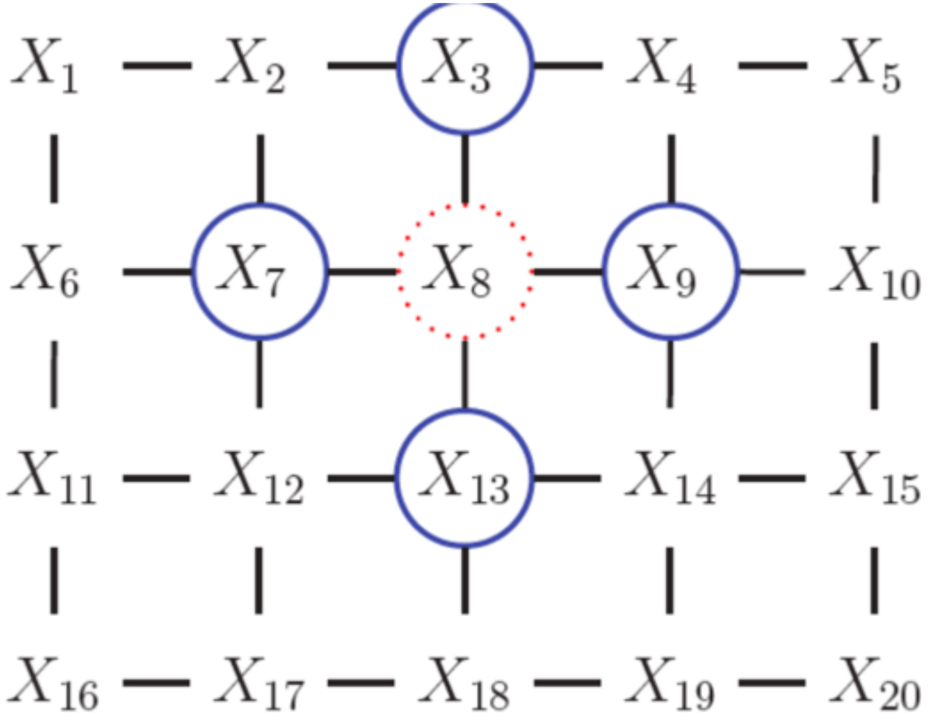
Global: $1, 2 \perp 6, 7 | 3, 4, 5$

$1 \perp 7 | \text{rest}?$, $1 \perp 20 | \text{rest}?$, $1 \perp 2 | \text{rest}?$

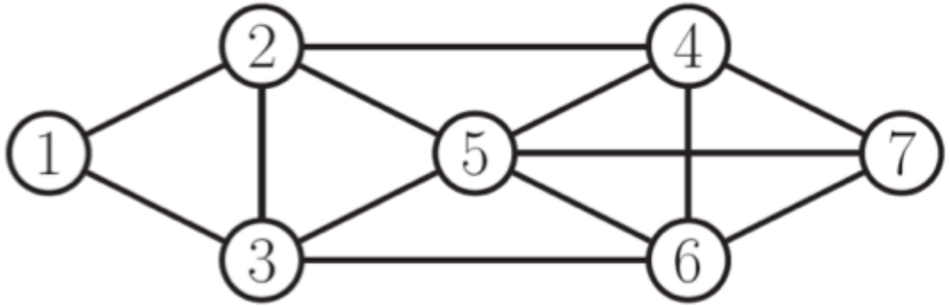
$1 \perp \text{rest} | ?$, $8 \perp \text{rest} | ?$

$1, 2 \perp 15, 20 | ?$





Example of Dependencies



Pairwise: $1 \perp 7 | \text{rest}$

Local: $1 \perp \text{rest} | 2, 3$

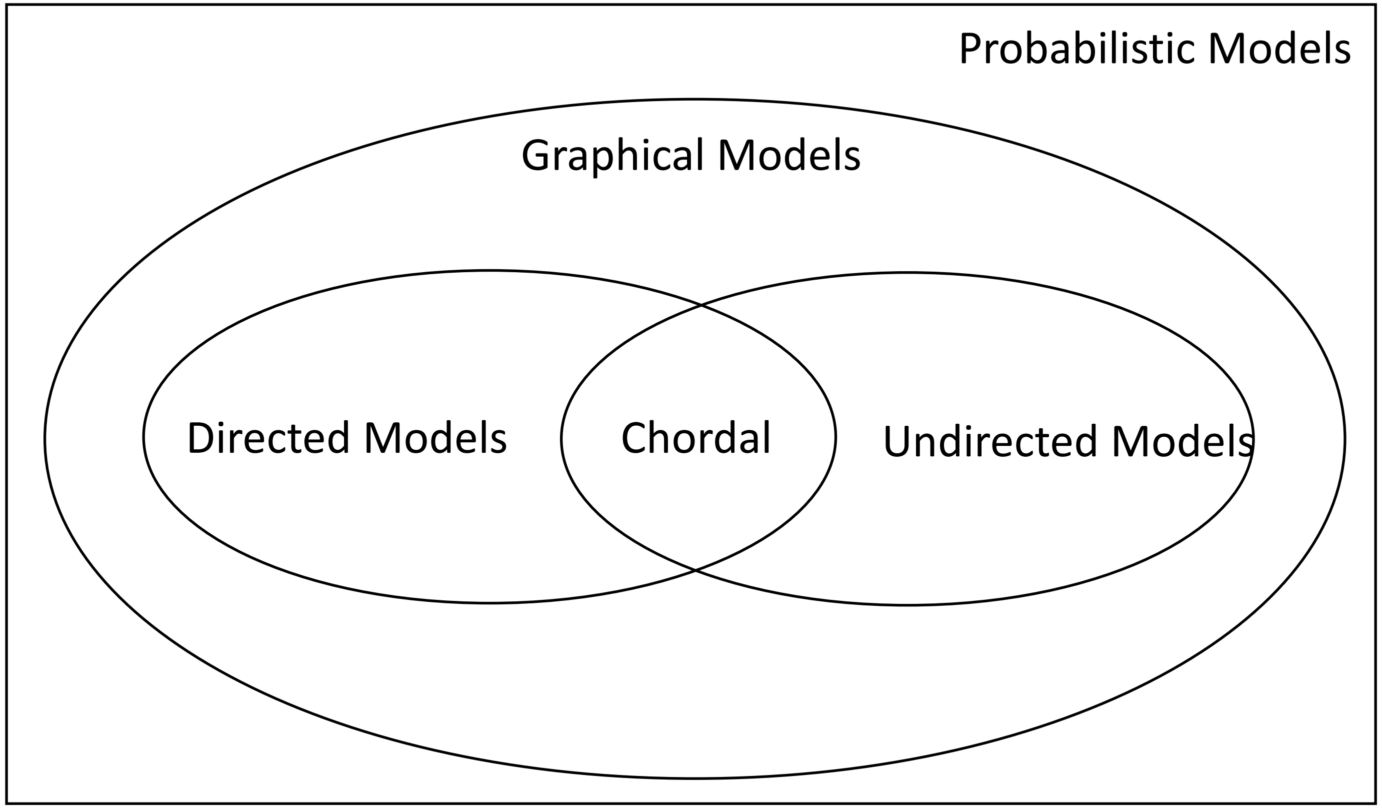
Global: $1, 2 \perp 6, 7 | 3, 4, 5$

Global $\Longrightarrow$ Local $\Longrightarrow$ Pairwise

$p(x) > 0$

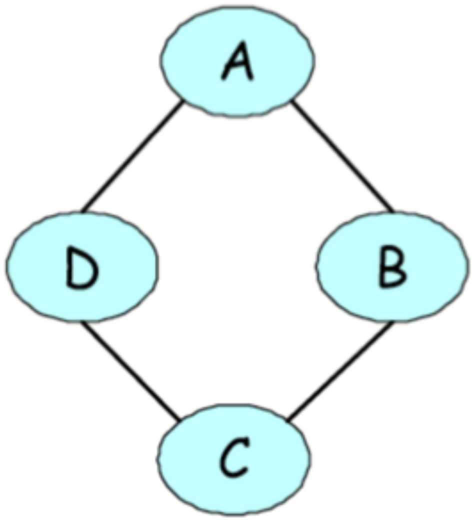
For proof: See page 119 of the book by Koller and Friedman

UGM and DGM



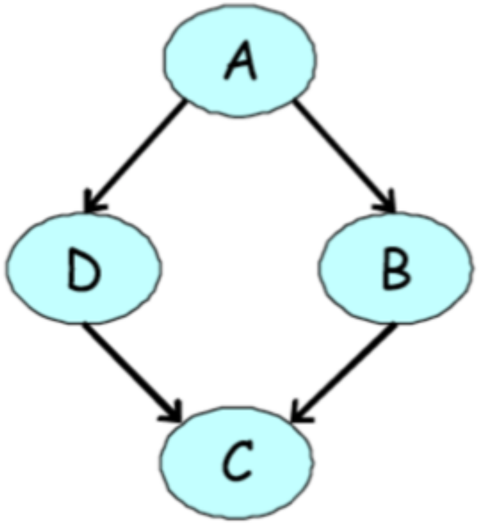
Triangulation: UGM $\Longrightarrow$ DGM
Moralization: DGM $\Longrightarrow$ UGM

Not all UGM can be represented as DGM



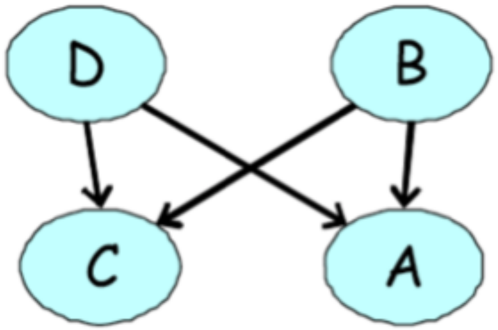
$A \perp C | D, B$

$B \perp D | A, C$



$A \perp C | D, B$

$B \perp D | A, C$ **X**

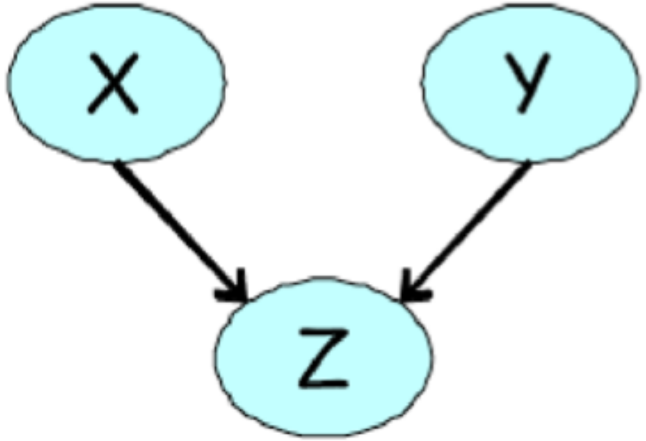


$A \perp C | D, B$

$B \perp D | A, C$ **X**

In this graph, B and D are marginally independent

Not all DGM can be represented as UGM



Undirected model fails to capture the marginal independence $(X \perp Y)$ that holds in the directed model at the same time as $\neg(X \perp Y|Z)$

What is this “Clique”?

Undirected graphical models (UGM)

Defn (also called Markov Network): For a set of variables $\mathcal{X} = \{x_1, \cdots, x_n\}$ a Markov network is defined as a product of potentials on subsets of the variables $\mathcal{X}_c \subseteq \mathcal{X}$

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

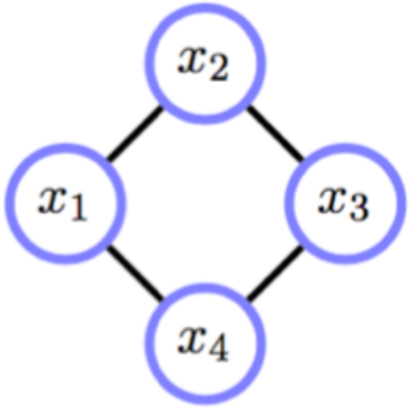
Normalizer to ensure it is a p is a probability

This is called potential ≥ 0 (this does not have to be probability)

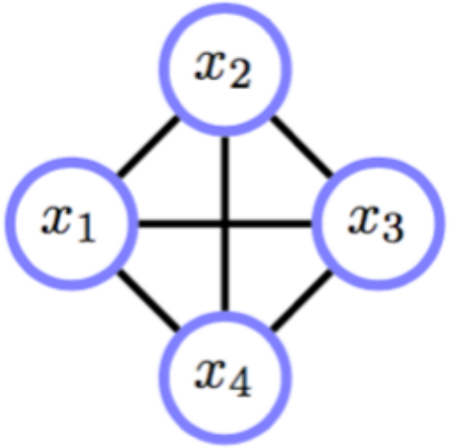
Maximal clique

Def: A maximal clique is a clique that cannot be extended by including one more adjacent vertex, meaning it is not a subset of a larger clique.

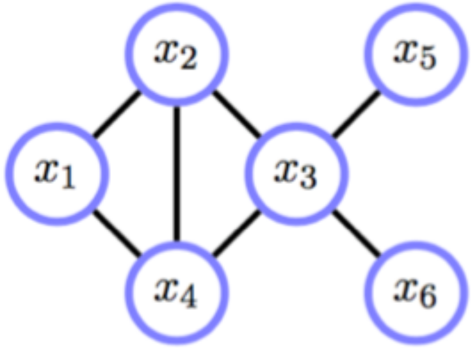
Examples



$$\phi(x_1, x_2)\phi(x_2, x_3)\phi(x_3, x_4)\phi(x_4, x_1)/Z_a$$

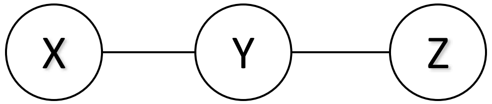


$$\phi(x_1, x_2, x_3, x_4)/Z_b$$



$$\phi(x_1, x_2, x_4)\phi(x_2, x_3, x_4)\phi(x_3, x_5)\phi(x_3, x_6)/Z_c$$

Interpretation of Clique Potentials



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

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

...but also we can write it

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

...but also ...

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

$\phi_1(x, y)$

...but also ...

$$\frac{f_1(x, y)}{Z_1}\frac{f_2(y, z)}{Z_2}$$

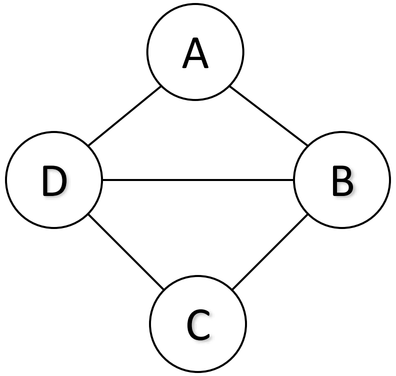
$\phi_2(y, x)$

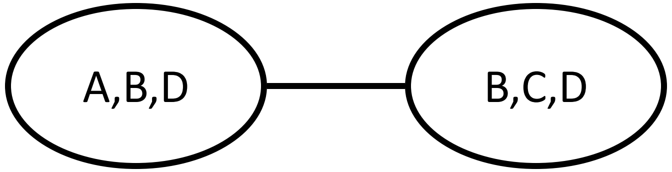
Interpretation of Clique Potentials

Take-home message about potentials:

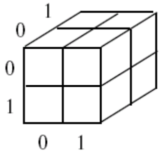
- Those are not necessarily marginals or conditionals.
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

Example UGM – using max cliques





$\psi_c(\mathbf{x}_{124})$ $\psi_c(\mathbf{x}_{234})$

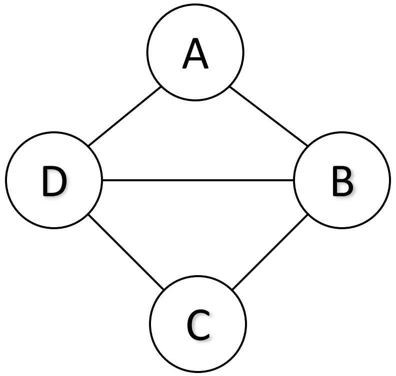


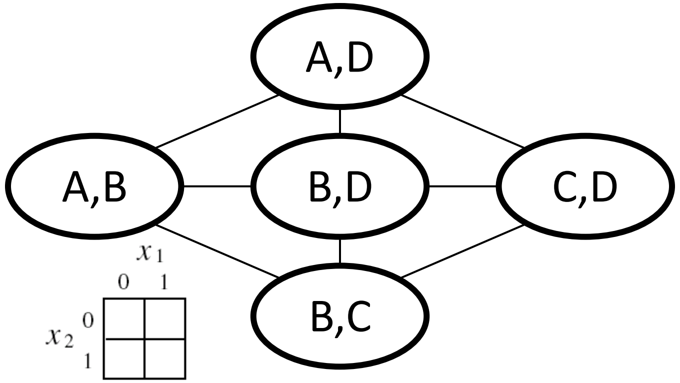
$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

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

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

Example UGM – using subcliques



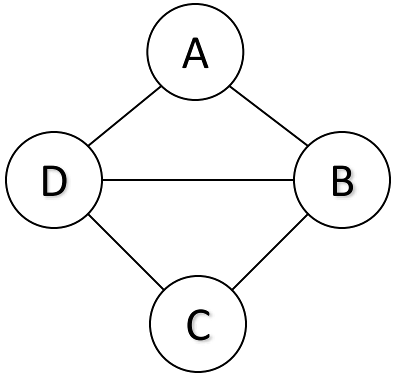


$$P''(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

$$Z = \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij}) \qquad = \frac{1}{Z} \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34})$$

- We can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table
- Pair MRFs, a popular and simple special case
- Are two graphs equivalent ($\mathcal{I}(P')$ and $\mathcal{I}(P'')$)?

Example UGM – canonical representation



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

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

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

Hammersley-Clifford Theorem

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

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

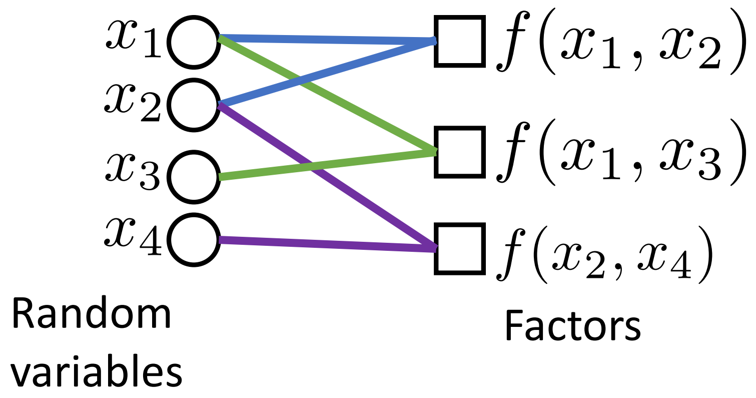
$$Z = \sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

  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*.

Factor Graphs

Factor Graph

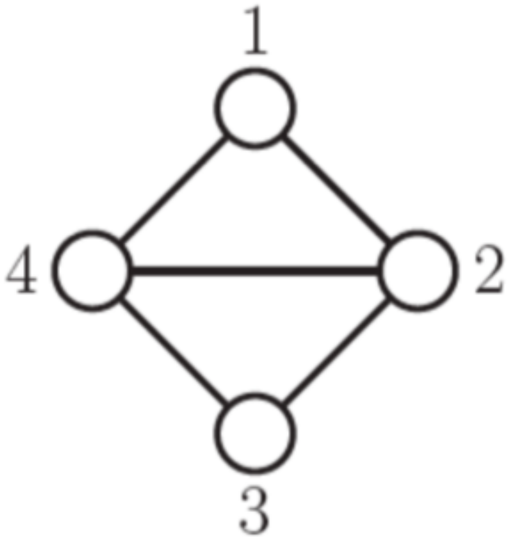


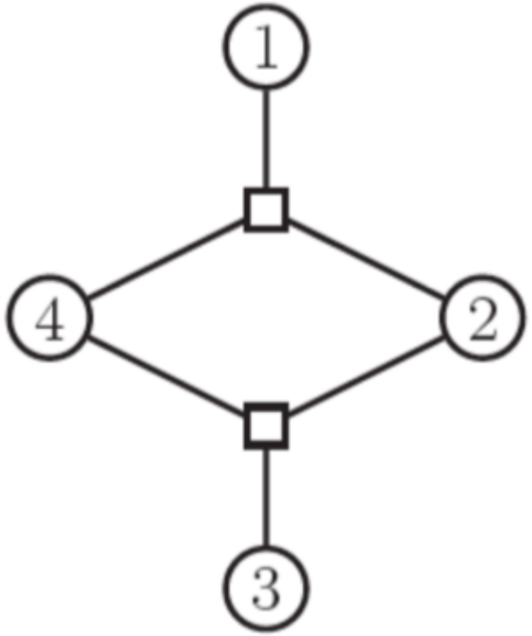
- A factor graph is a graphical model representation that unifies directed and undirected models
- It is an undirected bipartite graph with two kinds of nodes.
  - Round nodes represent variables,
  - Square nodes represent factors

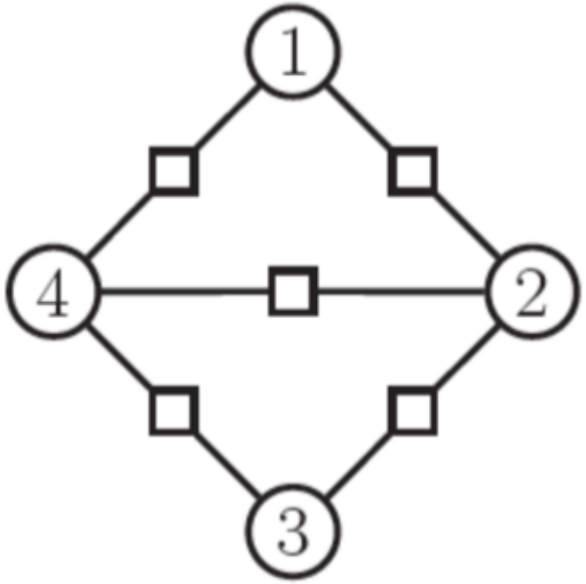
and there is an edge from each variable to every factor that mentions it.

- Represents the distribution more uniquely than a graphical model

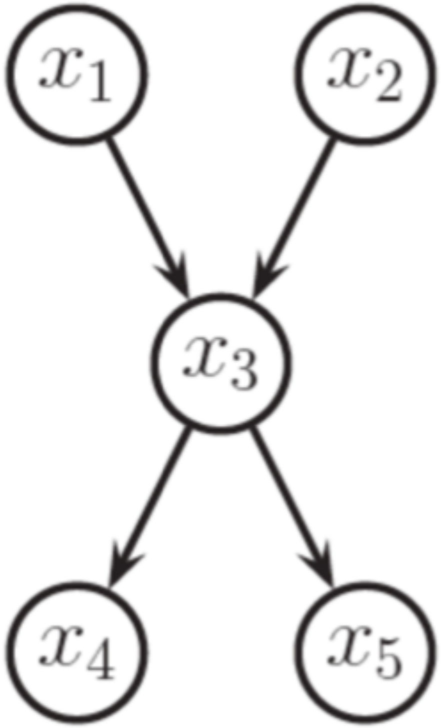
Factor Graph for UGM

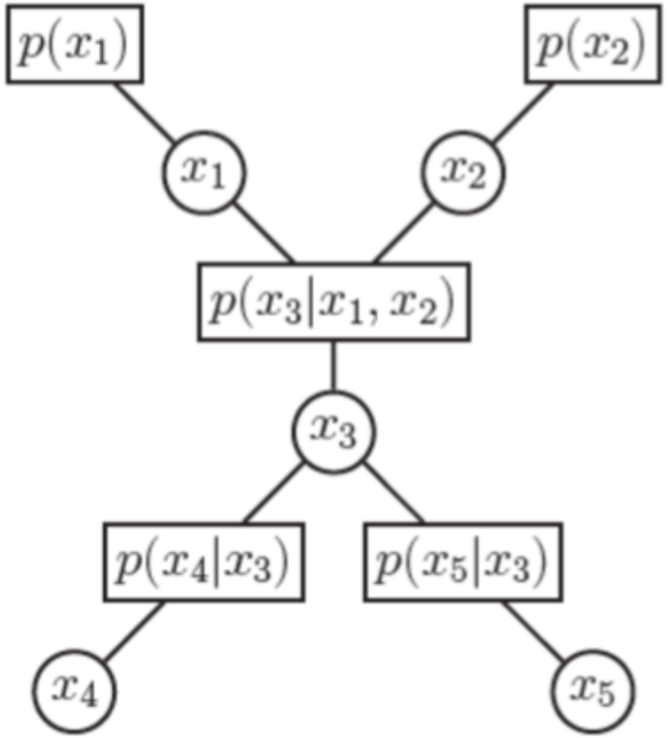






Factor Graph for DGM





One factor per CPD (conditional distribution) and connect the factor to all the variables that use the CPD

Practical Examples

Exponential Form

Remember the Gibbs distribution:

$$p(x_1, \cdots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \boxed{\psi_c(\mathcal{X}_c)}$$

So-called Potentials > 0

$$p(x_1, \cdots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \exp\left(-\boxed{\phi_c(\mathcal{X}_c)}\right)$$

Energy of the clique, can be positive/negative

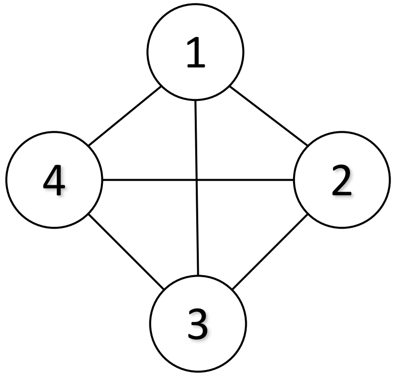
Free Energy of the system (log of prob):

$$H(x_1, \cdots, x_n) = \sum_c \phi_c(\mathcal{X}_c)$$

A powerful parametrization (log-linear model):

$$H(x_1, \cdots, x_n; \underbrace{\boxed{\theta}}_{\text{Param}}) = \sum_c \underbrace{\boxed{f_c(\mathcal{X}_c)}}_{\text{Feature function}}{}^T \theta_c$$

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

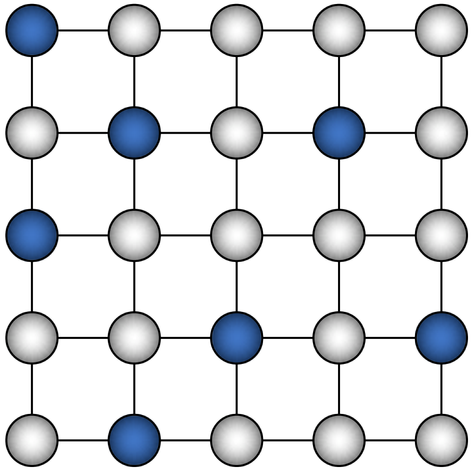
$$p(x_1, x_2, x_3, x_4; \theta; \alpha) = \frac{1}{Z(\theta, \alpha)} \exp\left[\sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i\right]$$

Hence the overall energy function has a quadratic form.

$$H(\mathbf{x}; \Theta, \mu) = (\mathbf{x} - \mu)^T \Theta (\mathbf{x} - \mu)$$

Ising models

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

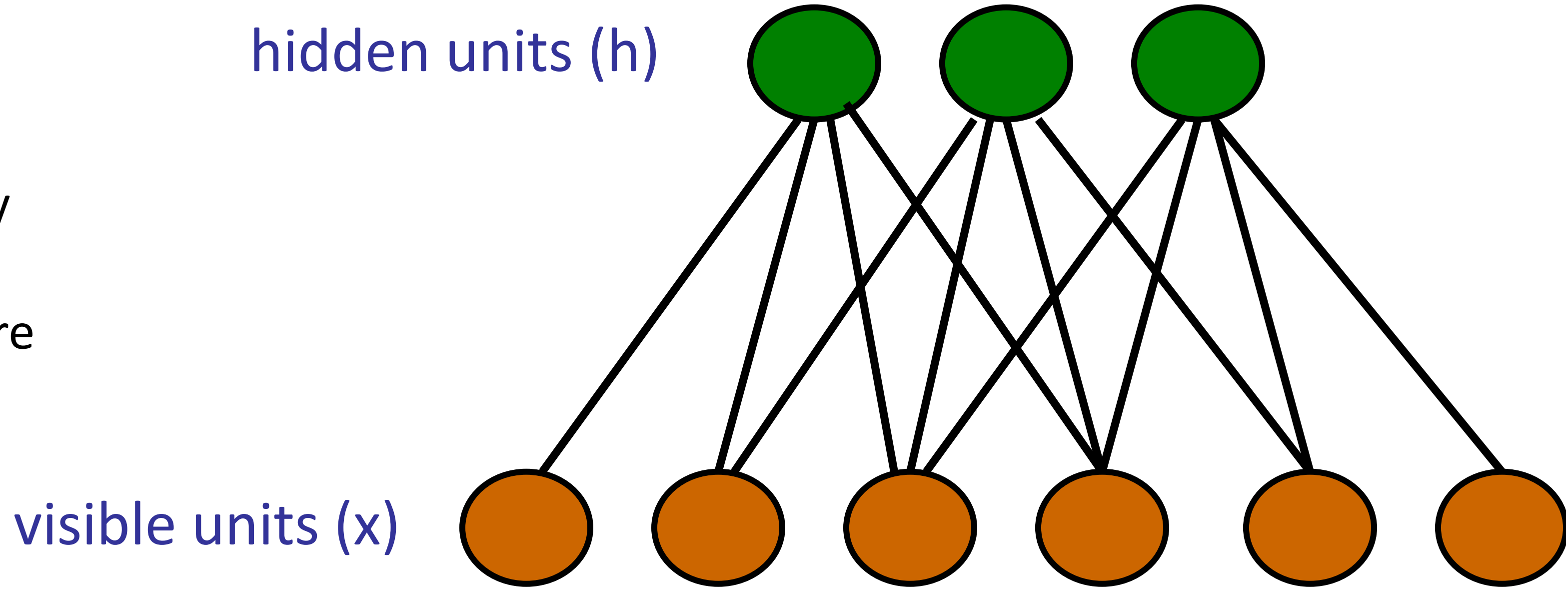


$$p(X) = \frac{1}{Z} \exp\left\{ \sum_{i,j \in N_i} \theta_{ij} X_i X_j + \sum_i \theta_{i0} X_i \right\}$$

- 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 (RBM)

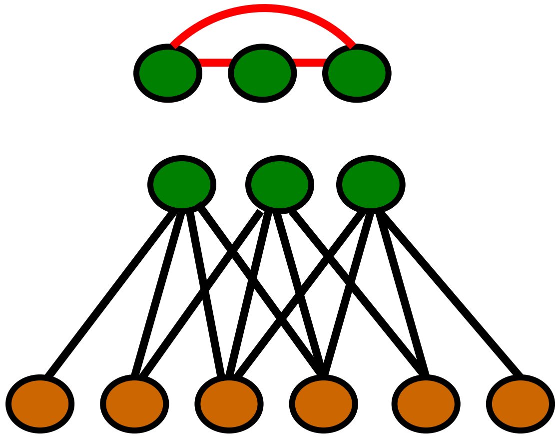
- Observed can pixels, signal in speech, word in a document
- Unobserved has “a notion” of summary of data
- One can use it as building block for more complicated models



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

Properties of RBM

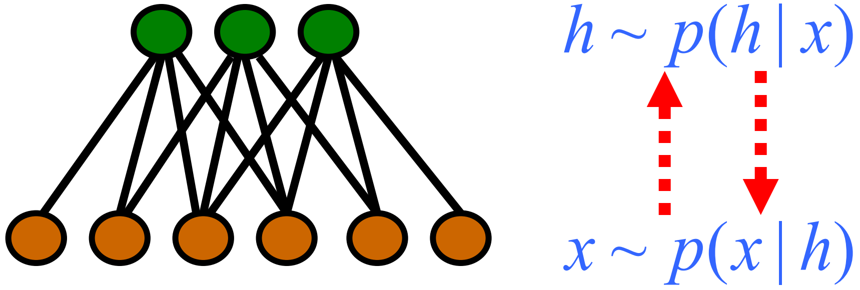
- Factors are marginally *dependent*.



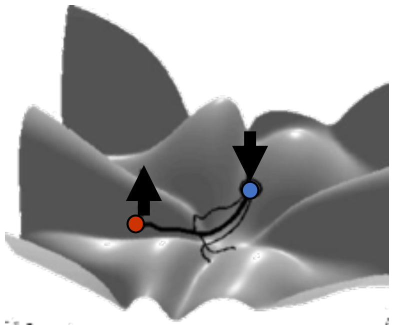
- Factors are conditionally *independent* given observations on the visible nodes.

$$p(h_1, \cdots, h_M|\mathbf{x}) = \prod_m p(h_m|\mathbf{x})$$

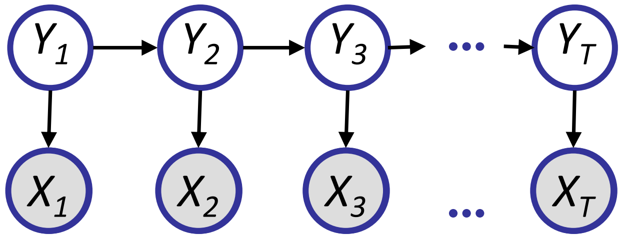
- Iterative Gibbs sampling to generate pairs of (x,h).

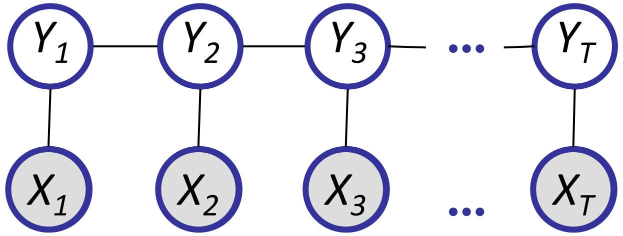


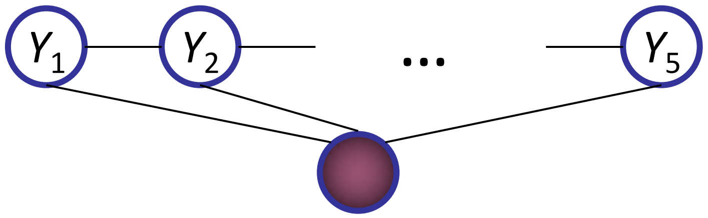
- Learning with contrastive divergence



Conditional Random Fields







- For example: part of speech labeling
- We are interested in Discriminative (not joint):

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

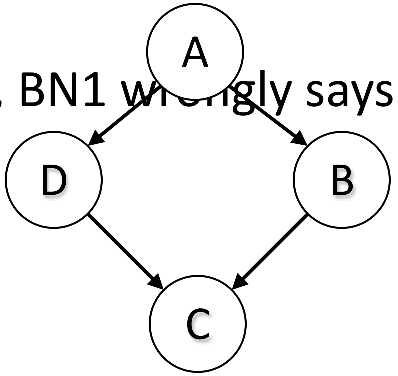
Summary

- 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

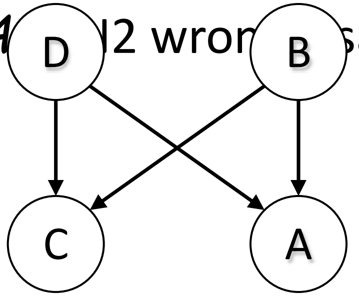
Extra slides

P-maps

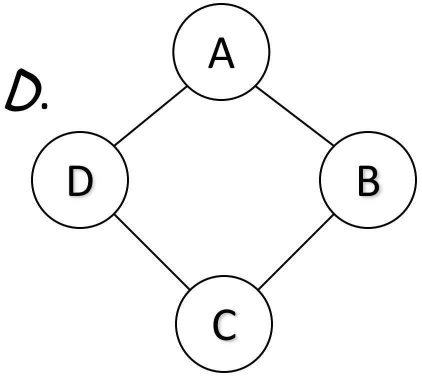
- Defn: A DAG G is a **perfect map** (P-map) for a distribution P if $I(P)=I(G)$.
- Thm: not every distribution has a perfect map as DAG.
  - Pf by counterexample. Suppose we have a model where $A \perp C \mid \{B,D\}$, and $B \perp D \mid \{A,C\}$. This cannot be represented by any Bayes net.
    - e.g., BN1 wrongly says $B \perp D \mid A$, BN2 wrongly says $B \perp D$.



BN1



BN2



MRF

P-maps

- Defn: A DAG G is a **perfect map** (P-map) for a distribution P if $I(P)=I(G)$.
- Thm: not every distribution has a perfect map as DAG.
  - Pf by counterexample. Suppose we have a model where $A \perp C \mid \{B,D\}$, and $B \perp D \mid \{A,C\}$.
    This cannot be represented by any Bayes net.
    - e.g., BN1 wrongly says $B \perp D \mid A$, BN2 wrongly says $B \perp D$.
  - The fact that G is a minimal I-map for P is far from a guarantee that G captures the independence structure in P
  - The P-map of a distribution *is* unique up to I-equivalence between networks. That is, a distribution P can have many P-maps, but all of them are I-equivalent.

Representation

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

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

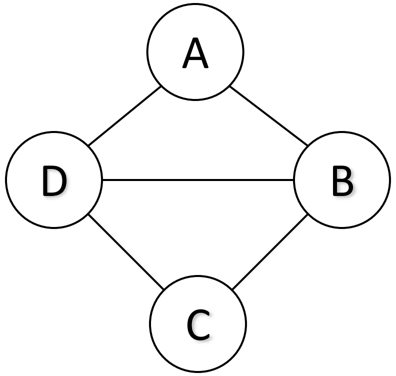
where Z is known as the partition function:

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

- 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 = $\{A,B\}$, $\{C,D\}$, …→ all edges and singletons

Gibbs Distribution and Clique Potential

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

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

where Z is known as the partition function:

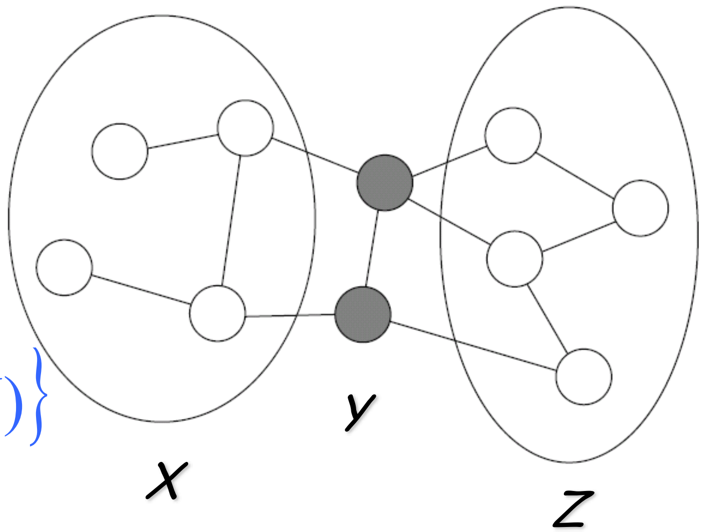
$$Z = \sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- 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.

II: Independence properties:

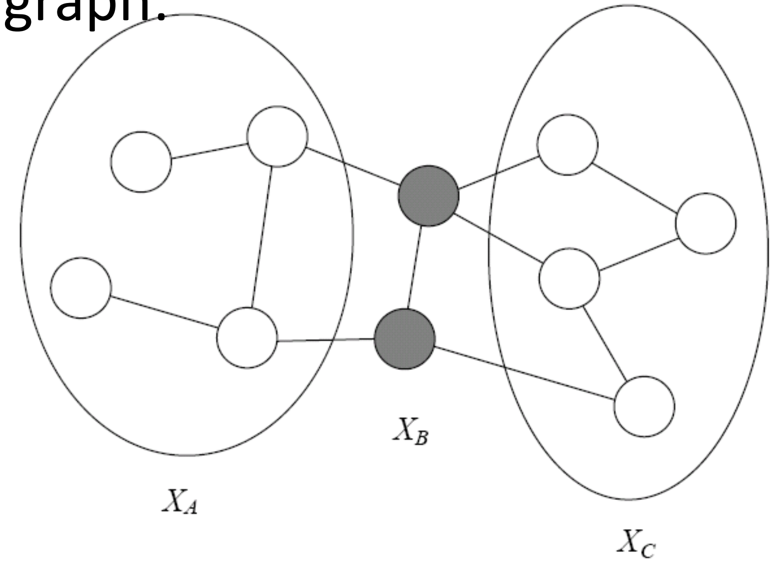
- Now let us ask what kinds of distributions can be represented by undirected graphs (ignoring the details of the particular parameterization).
- Defn: the global Markov properties of a UG *H* are

$$I(H) = \left\{ X \perp Z \middle| Y) : \mathrm{sep}_H(X; Z \middle| Y) \right\}$$



Global Markov Independencies

- Let H be an undirected graph:

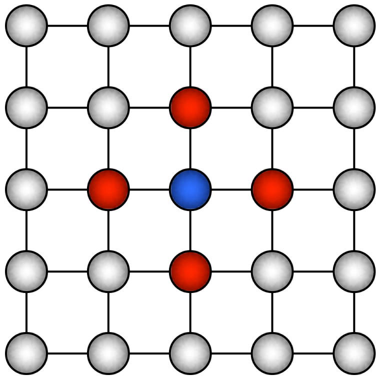


- B ***separates*** A and C if every path from a node in A to a node in C passes through a node in B: $\mathrm{sep}_H(A;C|B)$
- A probability distribution satisfies the ***global Markov property*** if for any disjoint A, B, C, such that B separates A and C, A is independent of C given B:

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

Local Markov independencies

- For each node $X_i \in \mathbf{V}$, there is *unique Markov blanket* of X_i, denoted MB_{Xi}, 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:

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

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 $I(H) \subseteq I(P)$, i.e., P entails $I(H)$.
- Defn: P is a Gibbs distribution over H if it can be represented as

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

- Thm (soundness): If P is a Gibbs distribution over H, then H is an I-map of P.
- Thm (completeness): If $\neg \text{sep}_H(X; Z \mid Y)$, then $X \not\perp_P Z \mid Y$ in ***some*** P that factorizes over H.

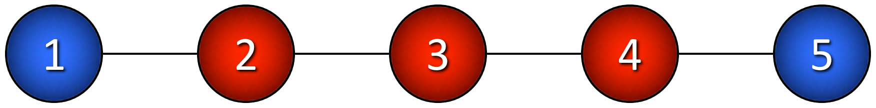
Other Markov properties

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

$$\mathrm{I}_p(H) = \{X \perp Y \mid V \setminus \{X,Y\} : \{X,Y\} \notin E\}$$

$$X_1 \perp X_5 \mid \{X_2, X_3, X_4\}$$

- e.g.,



Relationship between local and global Markov properties

- Thm 5.5.5. If $P \models I_l(H)$ then $P \models I_p(H)$.
- Thm 5.5.6. If $P = I(H)$ then $P \models I_l(H)$.
- Thm 5.5.7. If $P > 0$ and $P \models I_p(H)$, then $P \models I(H)$.

- **Corollary (5.5.8):** The following three statements are equivalent for a *positive distribution* P:

$$P \models I_l(H)$$
$$P \models I_p(H)$$
$$P \models I(H)$$

  - This equivalence relies on the positivity assumption.
  - We can design a distribution locally

Hammersley-Clifford Theorem

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

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

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

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

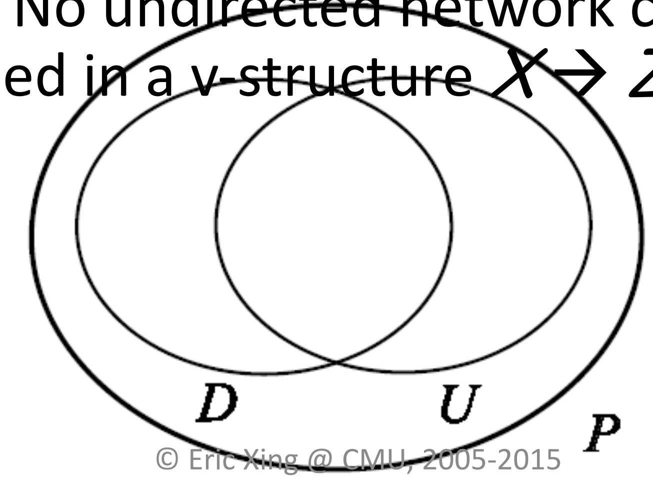
- **Thm :** Let P be a positive distribution over $\mathbf{V}$, and *H* a Markov network graph over $\mathbf{V}$. If *H* is an I-map for P, then P is a Gibbs distribution over *H*.

Perfect maps

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

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

- Thm: not every distribution has a perfect map as UGM.
  - Pf by counterexample. No undirected network can capture all and only the independencies encoded in a v-structure $X \rightarrow Z \leftarrow Y$.

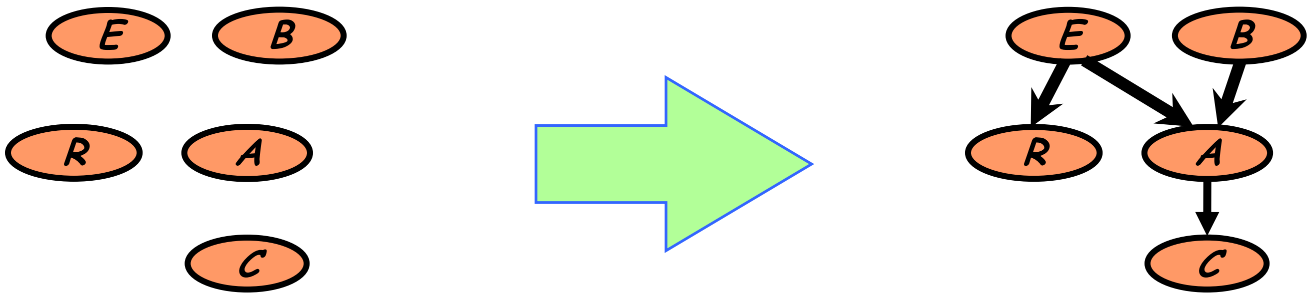


Where is the graph structure come from?

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



(B,E,A,C,R)=(T,F,F,T,F)
(B,E,A,C,R)=(T,F,T,T,F)
.... (B,E,A,C,R)=(F,T,T,T,F)

Information Theoretic Interpretation of ML

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

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

Information Theoretic Interpretation of ML (con'd)

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

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:

$$\ell(\theta_G, G; D) = \log \hat{p}(D \mid \theta_G, G)$$
$$= M\sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)}) - M\sum_i \hat{H}(x_i)$$

$$\Rightarrow \boxed{C(G) = M\sum_i \hat{I}(x_i, \mathbf{x}_{\pi_i(G)})}$$

- Chow-Liu:
  - For each pair of variable x_i and x_j
    - Compute empirical distribution: $$\hat{p}(X_i, X_j) = \frac{count(x_i, x_j)}{M}$$
    - Compute mutual information: $$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{p}(x_i, x_j) \log \frac{\hat{p}(x_i, x_j)}{\hat{p}(x_i)\hat{p}(x_j)}$$
  - Define a graph with node $x_1, \ldots, x_n$
    - Edge (I,j) gets weight $\hat{I}(X_i, X_j)$

Chow-Liu algorithm (con'd)

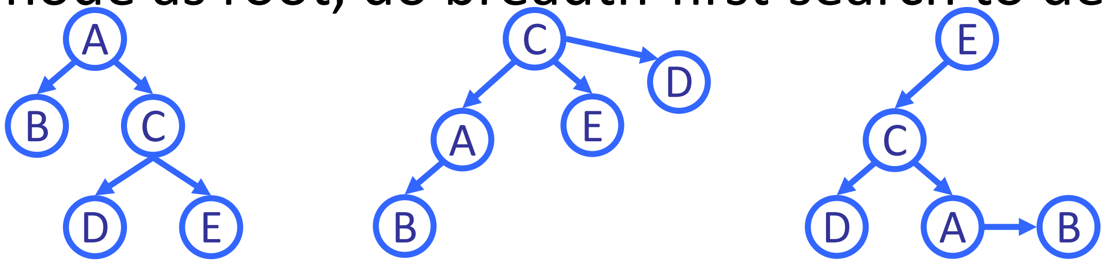
- Objection function:

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

- 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) $d \geq 2$

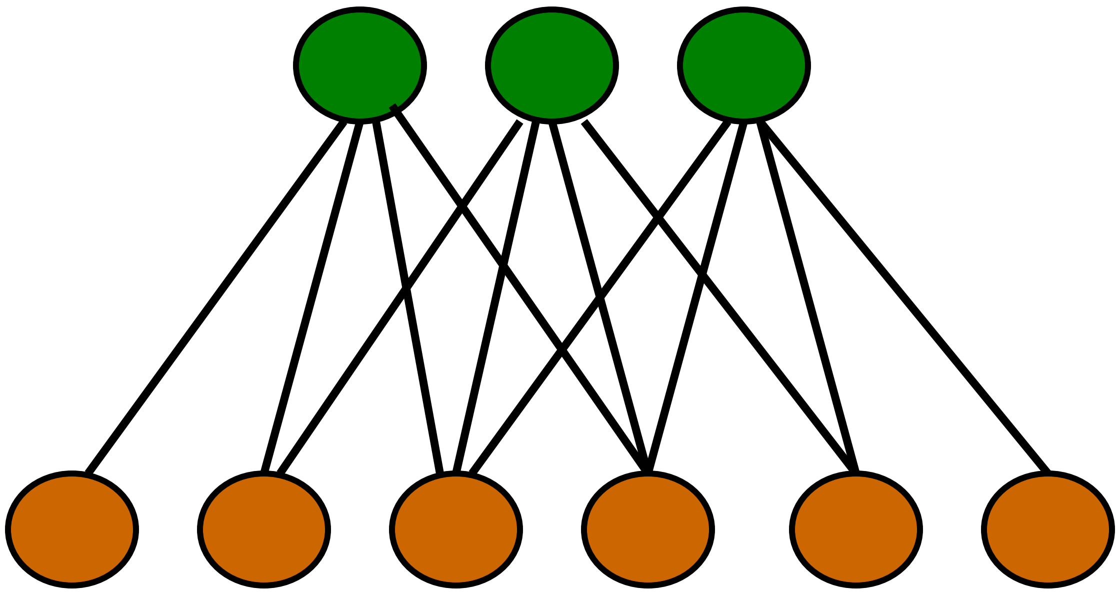
- Most structure learning approaches use heuristics
  - Exploit score decomposition
  - Two heuristics that exploit decomposition in different ways
    - Greedy search through space of node-orders
    - Local search of graph structures

Restricted Boltzmann Machines

The Harmonium (Smolensky –'86)

hidden units

visible units



History:
Smolensky ('86), Proposed the architechture.
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)