PETUUM

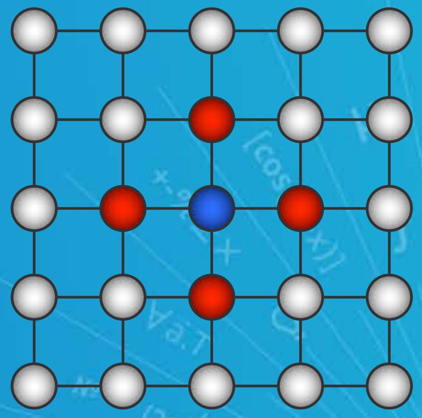
Carnegie Mellon University
School of Computer Science

Probabilistic Graphical Models

Representation of undirected GM

Eric Xing

Lecture 3, January 23, 2019

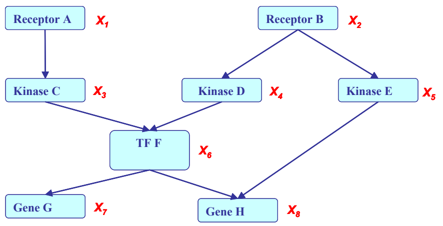


Reading: KF-chap4

Two types of GMs

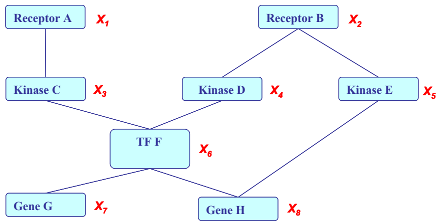
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3|\, X_1)\, P(X_4|\, X_2)\, P(X_5|\, X_2)\, P(X_6|\, X_3, X_4)\, P(X_7|\, X_6)\, P(X_8|\, X_5, X_6)$$



- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



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:

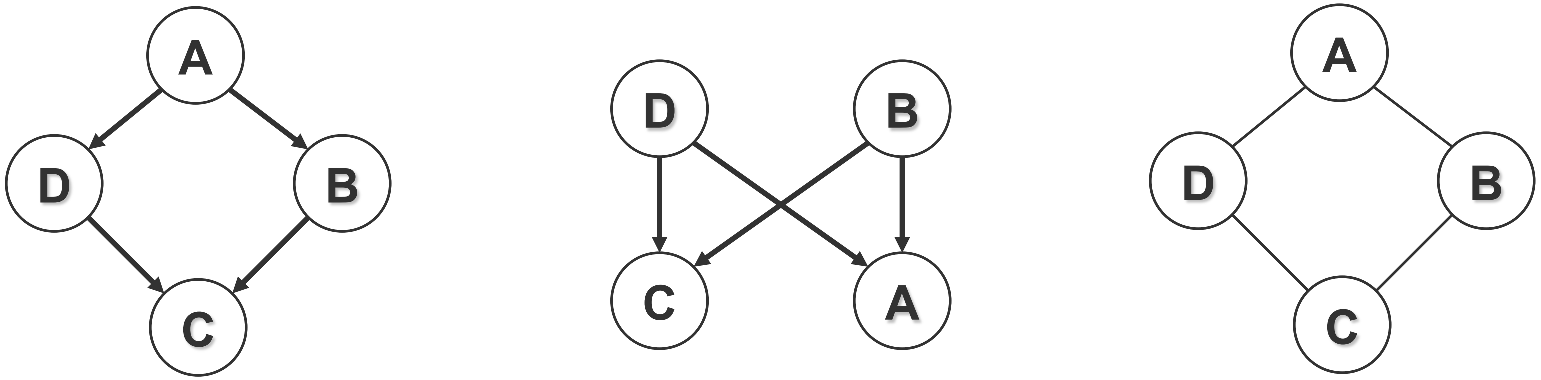
$$\mathrm{I}(G) = \left\{ X \perp Z \middle| Y : \mathrm{dsep}_G(X; Z \middle| Y) \right\}$$

- Defn: A DAG $\mathcal{G}$ is an I-map (independence-map) of $\mathcal{P}$ if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{P})$
- A fully connected DAG $\mathcal{G}$ is an I-map for any distribution, since $\mathcal{I}_l(\mathcal{G}) = \emptyset \subseteq \mathcal{I}(\mathcal{P})$ for any $\mathcal{P}$.
- Defn: A DAG $\mathcal{G}$ is a minimal I-map for $\mathcal{P}$ if it is an I-map for $\mathcal{P}$, and if the removal of even a single edge from $\mathcal{G}$ renders it not an I-map.
- A distribution may have several minimal I-maps
  - Each corresponding to a specific node-ordering

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 $B \perp D$.

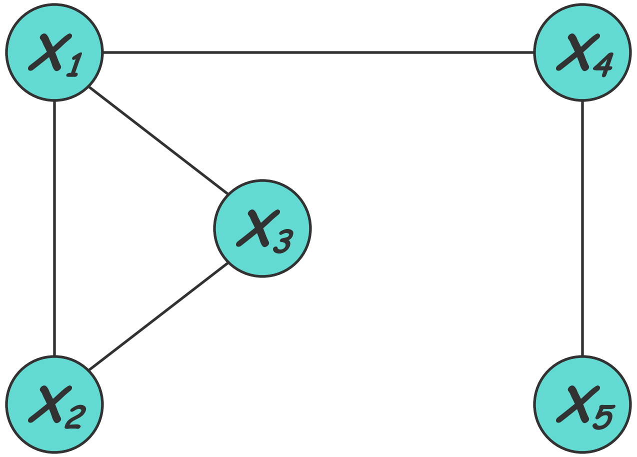


BN1 BN2 MRF

P-maps

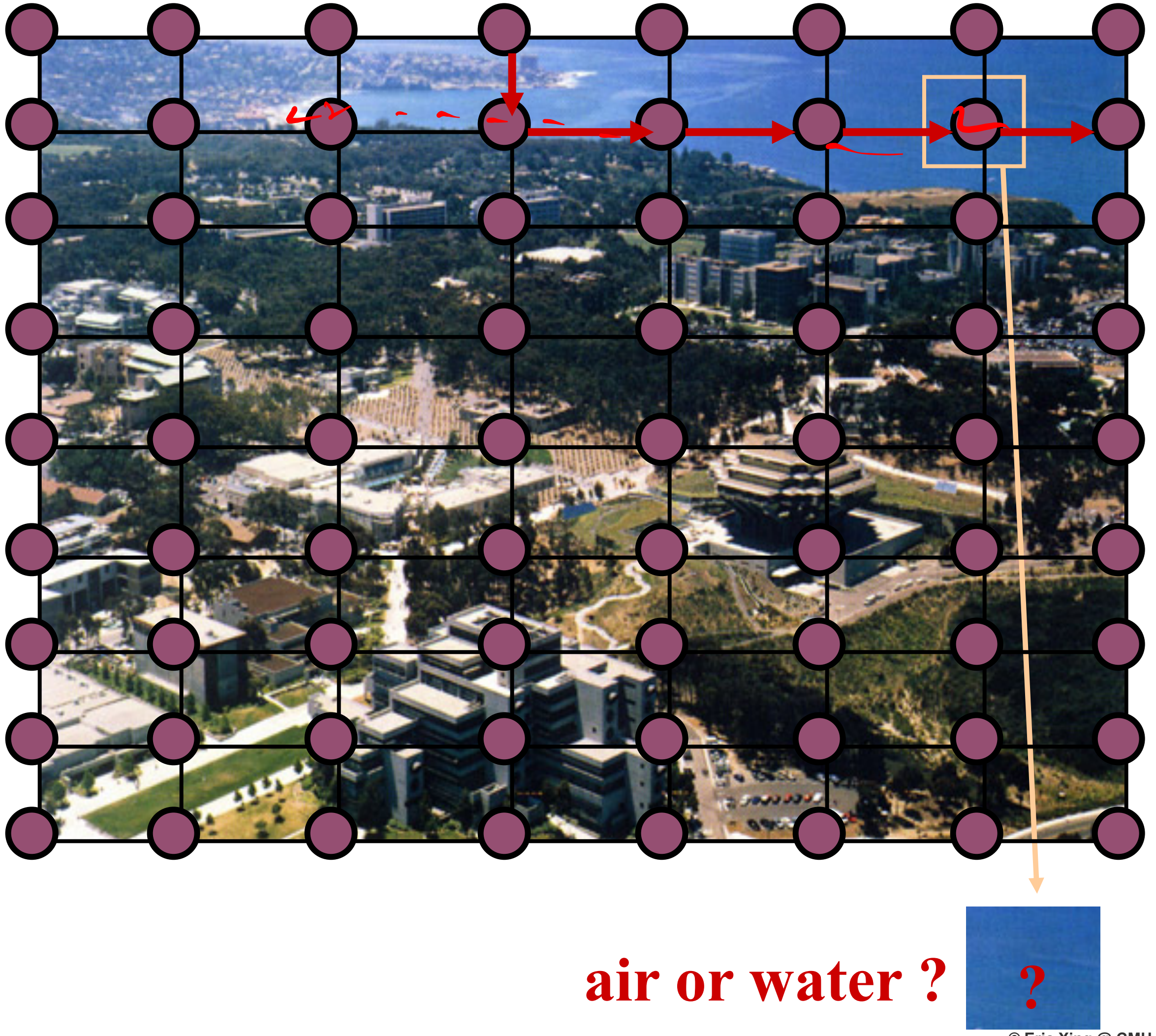
- Defn: A DAG G is a perfect map (P-map) for a distribution P if $I(P)=I(G)$.
- Thm: not every distribution has a perfect map as DAG.
  - Pf by counterexample. Suppose we have a model where $A \perp C \mid \{B,D\}$, and $B \perp D \mid \{A,C\}$. This cannot be represented by any Bayes net.
    - e.g.,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

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

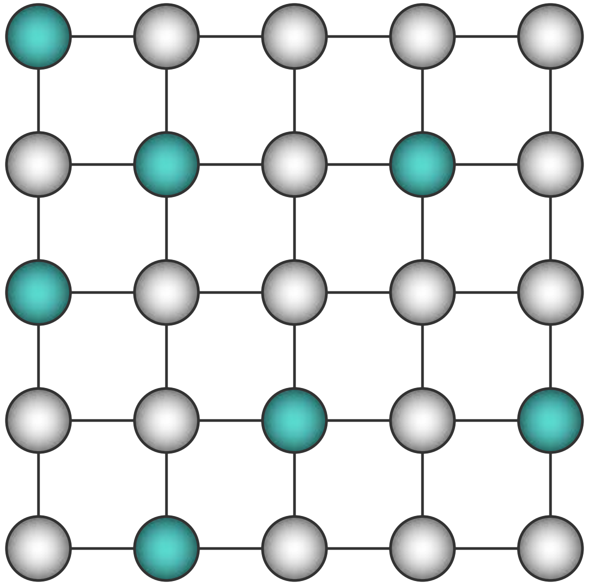
A Canonical Example: understanding complex scene …



air or water ?

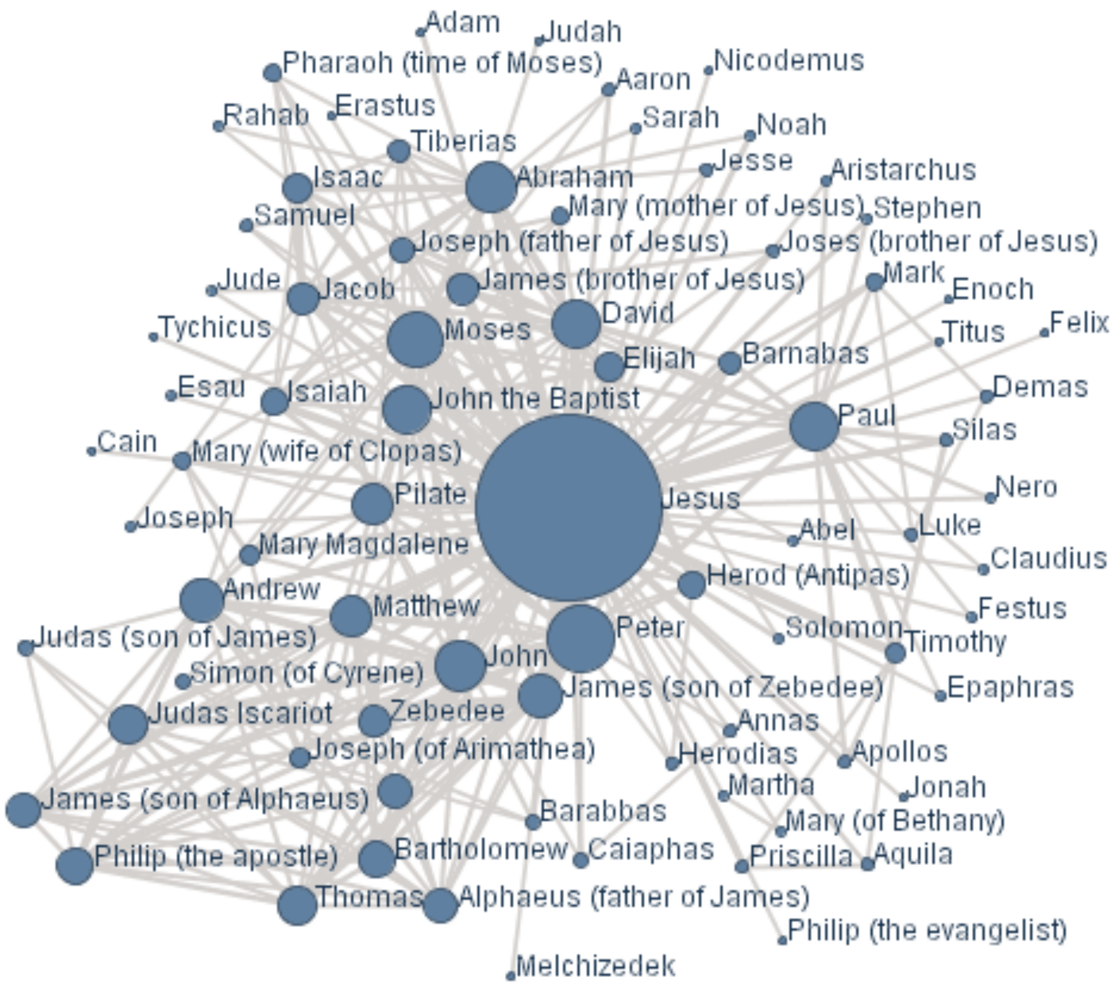
A Canonical Example

- The grid model



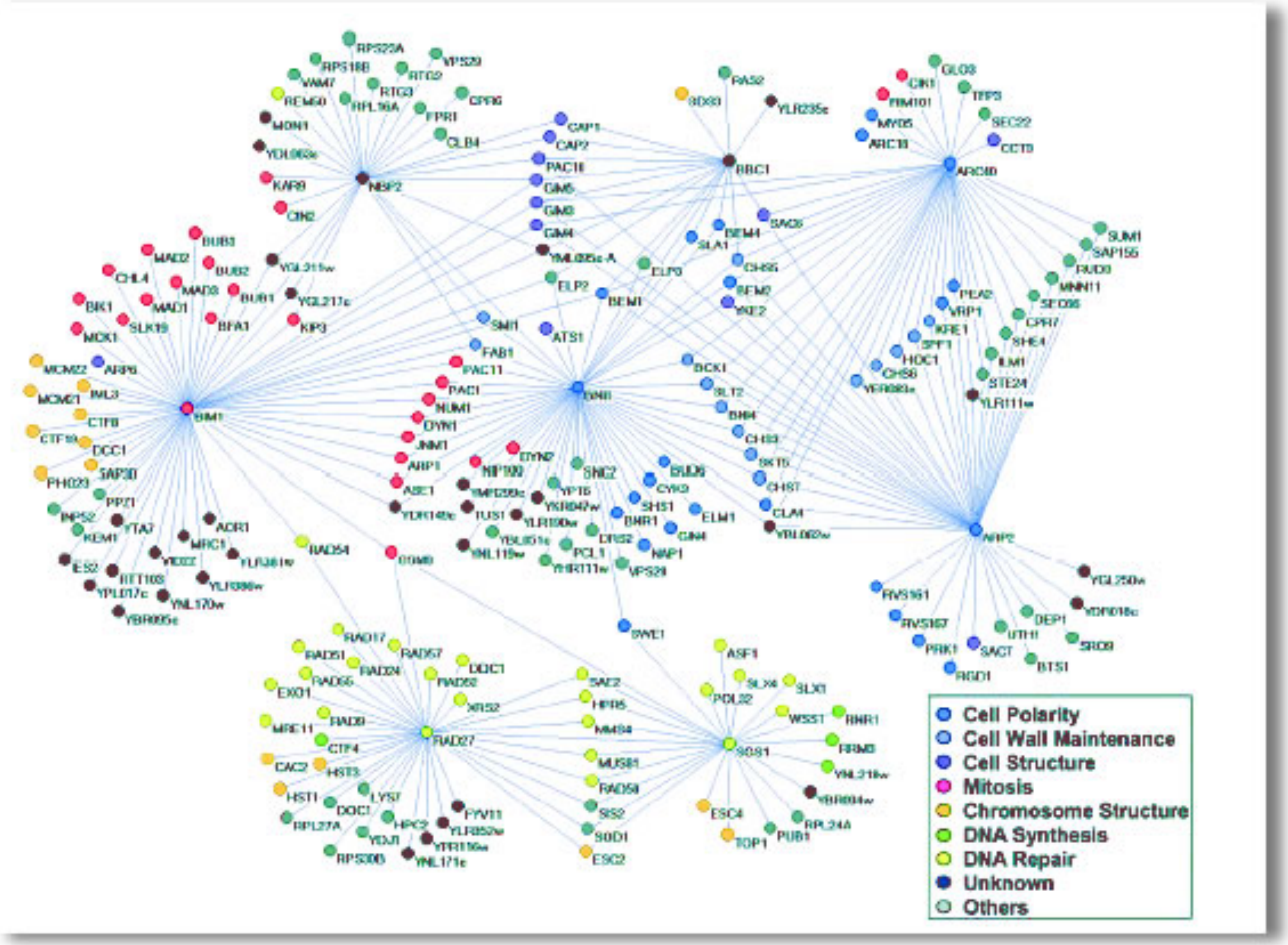
- Naturally arises in image processing, lattice physics, etc.
- Each node may represent a single "pixel", or an atom
  - The states of adjacent or nearby nodes are "coupled" due to pattern continuity or electro-magnetic force, etc.
  - Most likely joint-configurations usually correspond to a "low-energy"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

Social networks

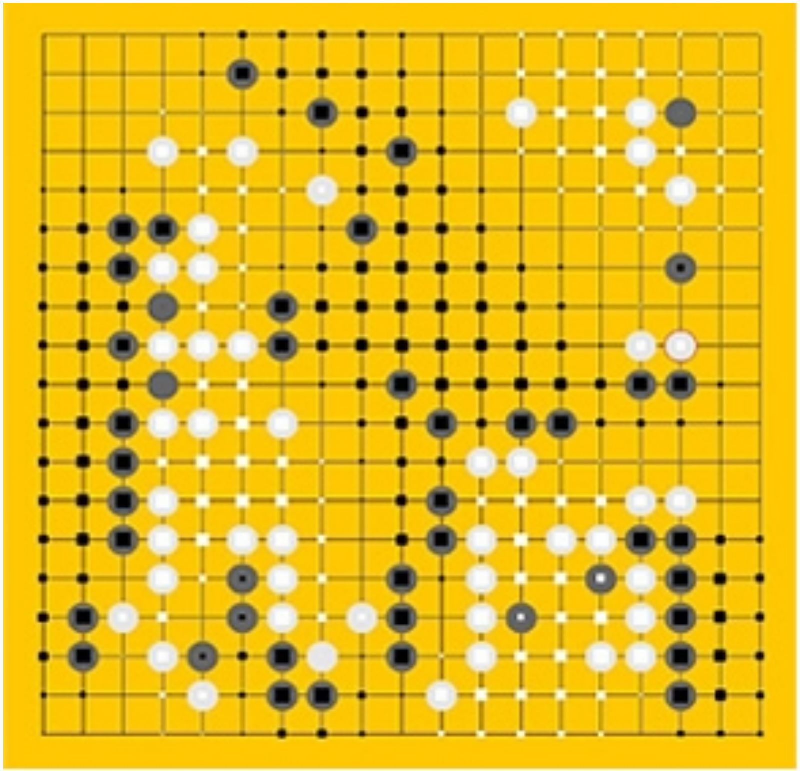


The New Testament Social Networks

Protein interaction networks

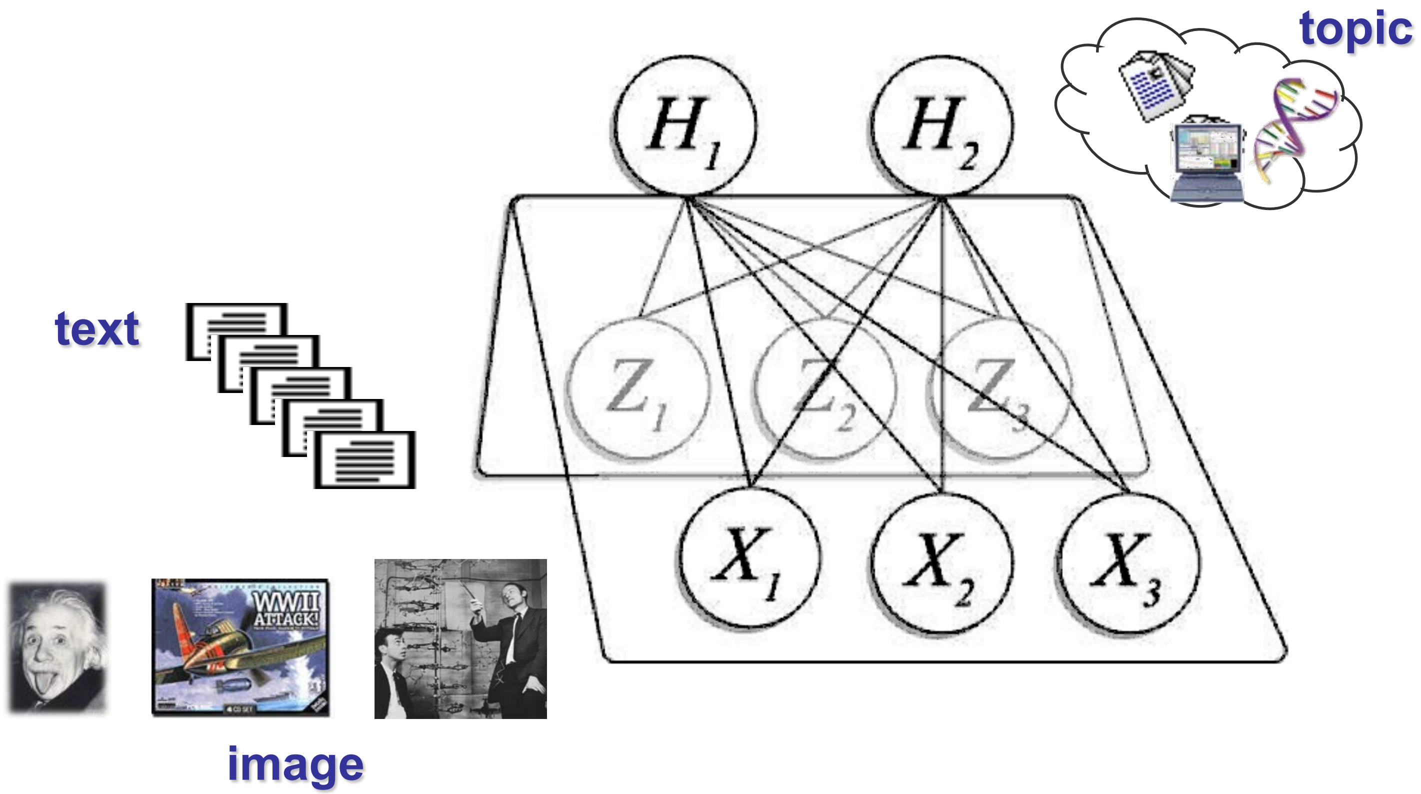


Modeling Go



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

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

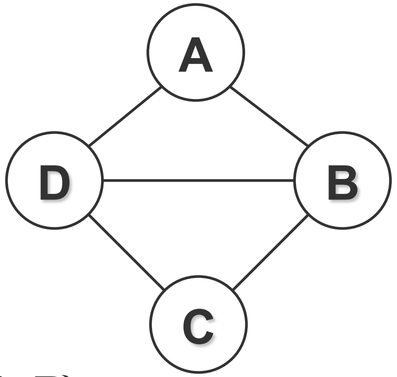
where Z is known as the partition function:

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

- 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.

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

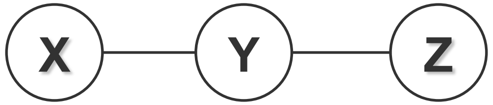
(A Gibbs distribution)

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.

Interpretation of Clique Potentials



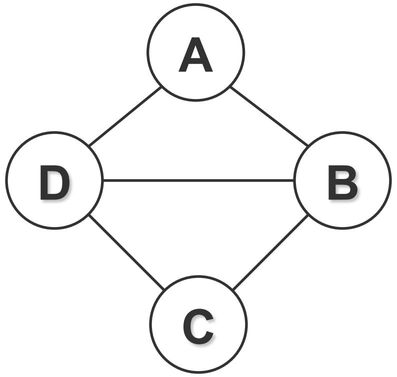
- The model implies $X \perp Z \mid Y$. This independence statement implies (by definition) that the joint must factorize as:

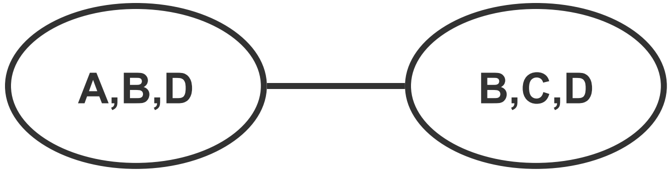
$$p(x,y,z) = p(y)p(x \mid y)p(z \mid y)$$

- We can write this as: $p(x,y,z) = p(x,y)p(z \mid y)$, $p(x,y,z) = p(x \mid y)p(z,y)$, but
  - cannot have all potentials be marginals
  - cannot have all potentials be conditionals

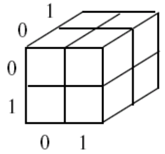
- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

Example UGM – using max cliques





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

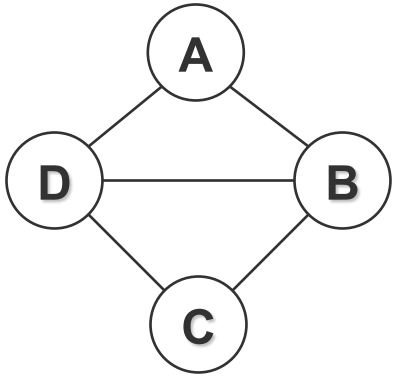


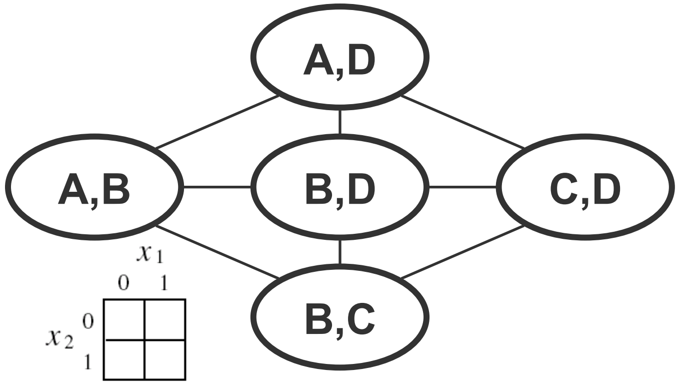
$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

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

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

Example UGM – using subcliques





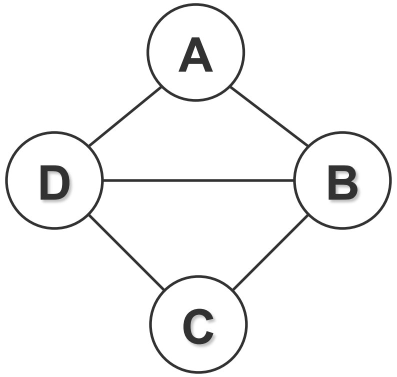
$$P''(x_1,x_2,x_3,x_4) = \frac{1}{Z}\prod_{ij}\psi_{ij}(\mathbf{x}_{ij})$$

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

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

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

Example UGM – canonical representation



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

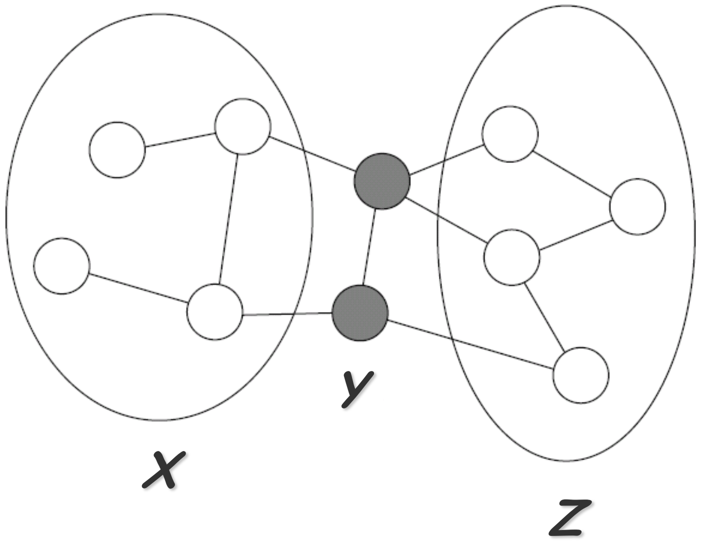
$$
Z = \sum_{x_1, x_2, x_3, x_4} \begin{array}{l} \psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234}) \\ \times \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34}) \\ \times \psi_1(x_1) \psi_2(x_2) \psi_3(x_3) \psi_4(x_4)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

- Most general, subsume P' and P" as special cases
- I(P) vs. I(P') vs. I(P")
- D(P) vs. D(P') vs. D(P")

II: Independence properties:

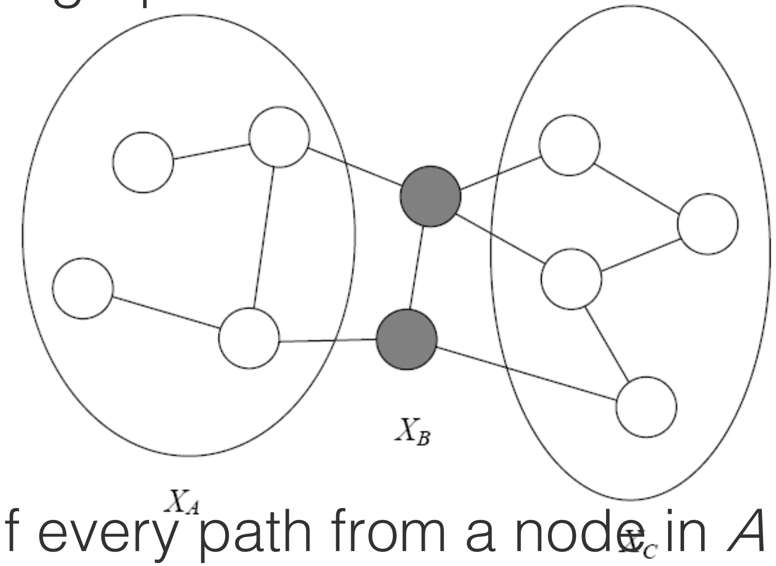
- Now let us ask what kinds of distributions can be represented by undirected graphs (ignoring the details of the particular parameterization).
- Defn: the global ~~Markov~~ properties of a UG *H* are

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



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*:

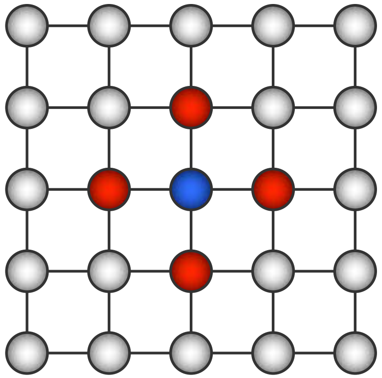
$$\mathrm{sep}_H(A;C|B)$$

- 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 $\neg\text{sep}_H(X; Z|Y)$, then $X \perp_P Z | Y$ 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

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

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

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 $P \models I(H)$ 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

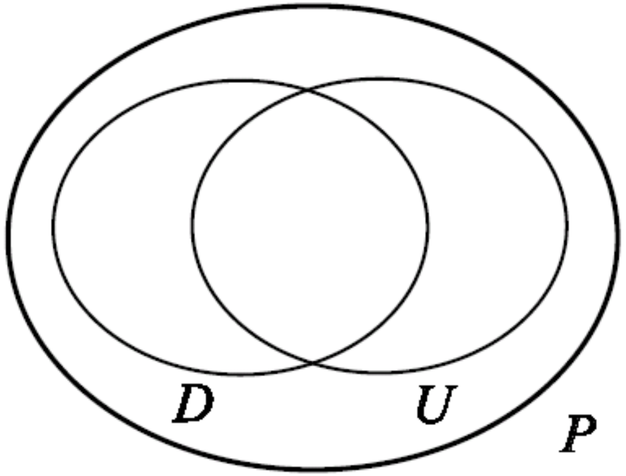
- Thm : Let P be a positive distribution over $\mathbf{V}$, and *H* a Markov network graph over $\mathbf{V}$. If <u>*H* is an I-map for P</u>, then P is a Gibbs distribution over *H*.

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 $X \rightarrow Z \leftarrow Y$.



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 strcuture

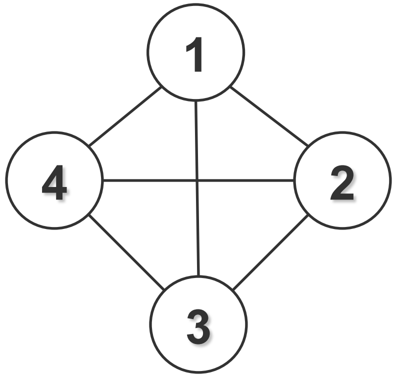
$$p(\mathbf{x}) = \frac{1}{Z}\exp\left\{-\sum_{c\in C}\phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z}\exp\{-H(\mathbf{x})\}$$

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

$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z}\exp\left\{\sum_{ij}\phi_{ij}(x_i, x_j)\right\}$$

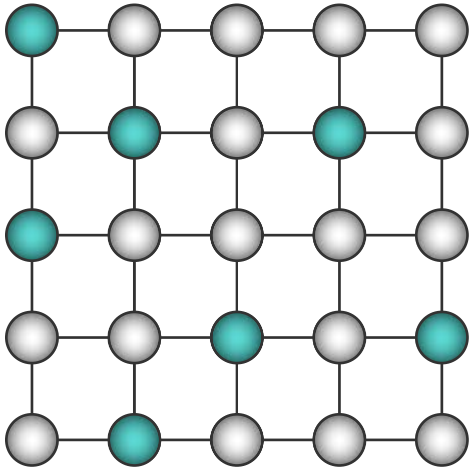
$$= \frac{1}{Z}\exp\left\{\sum_{ij}\theta_{ij}x_i x_j + \sum_i \alpha_i x_i + C\right\}$$

- 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.



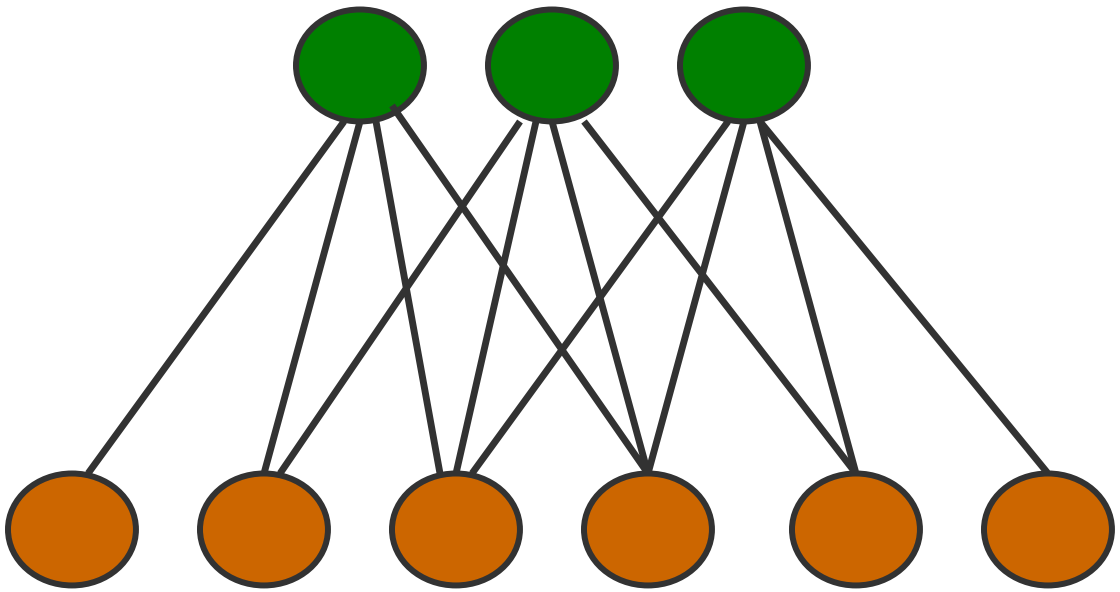
$$p(X) = \frac{1}{Z}\exp\left\{\sum_{i,j\in N_i}\theta_{ij}X_iX_j + \sum_i \theta_{i0}X_i\right\}$$

- 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

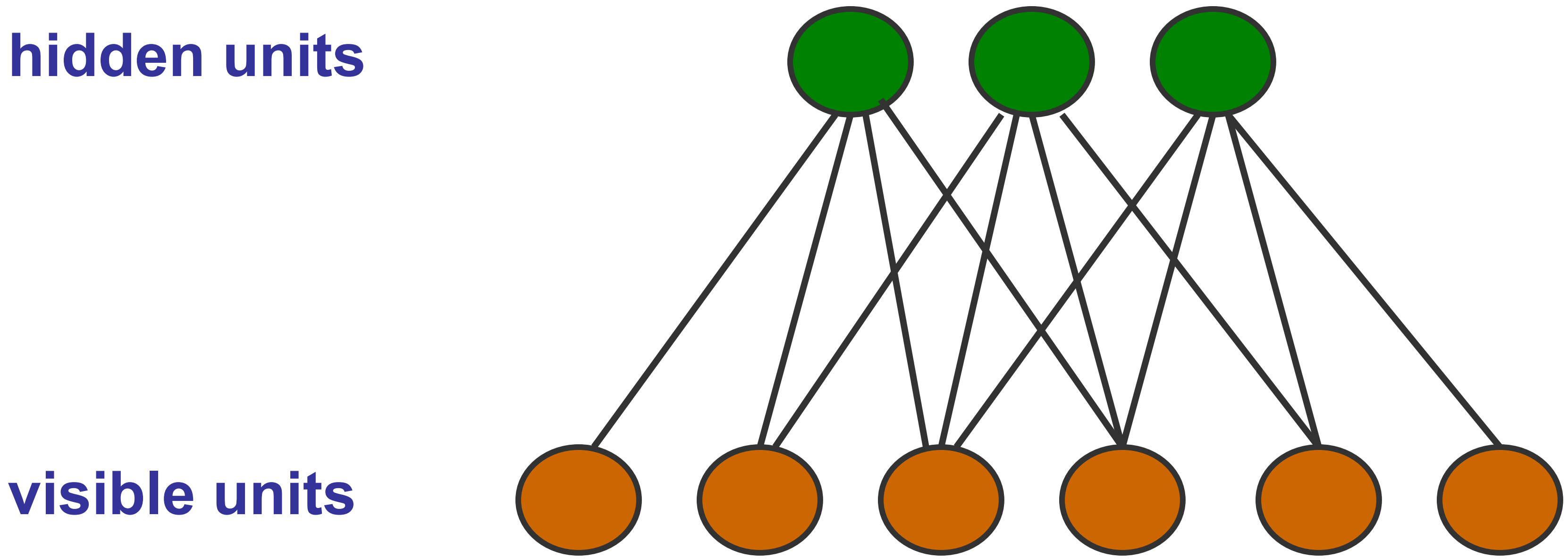
visible units



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

Restricted Boltzmann Machines

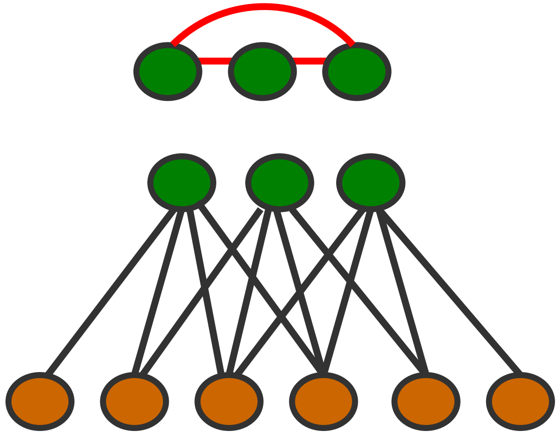
The Harmonium (Smolensky –'86)



<u>History:</u>
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)

Properties of RBM

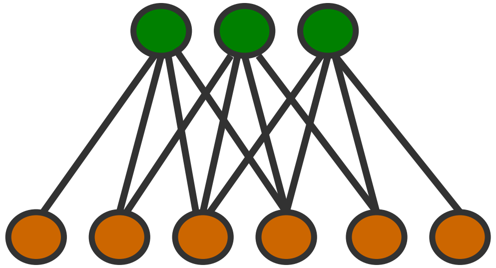
- Factors are marginally *dependent*.



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

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

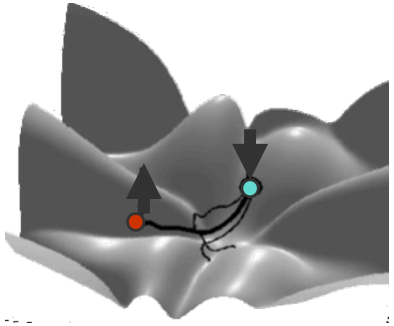
- Iterative Gibbs sampling.



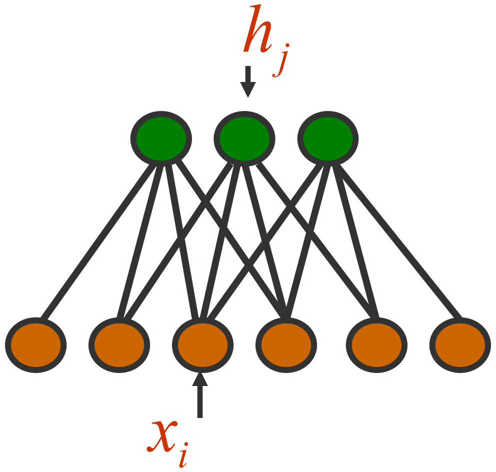
$$h \sim p(h \mid x)$$

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

- Learning with contrastive divergence



A Constructive Definition



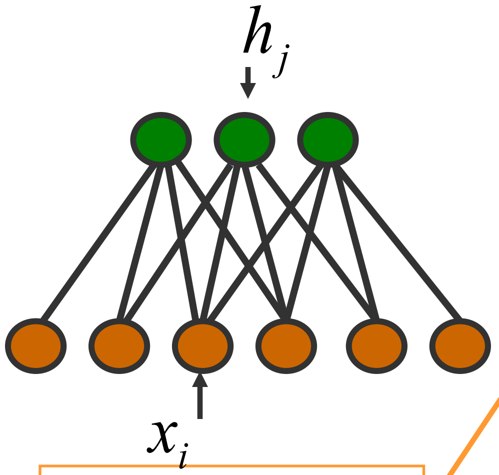
$$p_{\text{ind}}(\mathbf{h}) \propto \prod_j \exp\{\ \theta_j g_j(h_j)\ \}$$

how do we couple them?

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

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

A Constructive Definition



coupling in the log-domain with shifted parameters

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

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

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

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

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

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

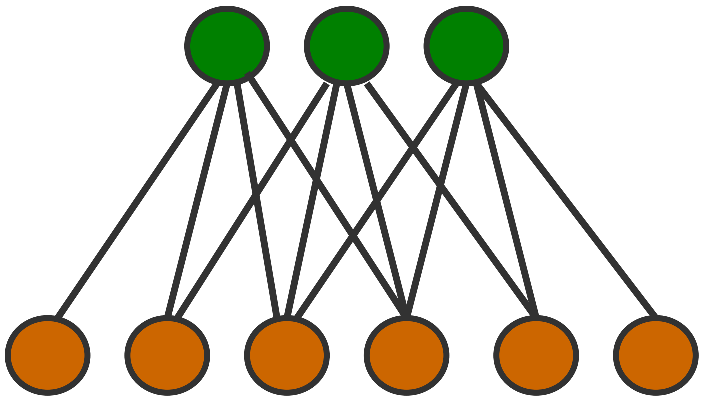
vector of local sufficient statistics (features)

They map to the RBM random field:

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

An RBM for Text Modeling

topics



words counts

$h_j = 3$: topic j has strength 3

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

$x_i = \mathbf{n}$: word i has count n

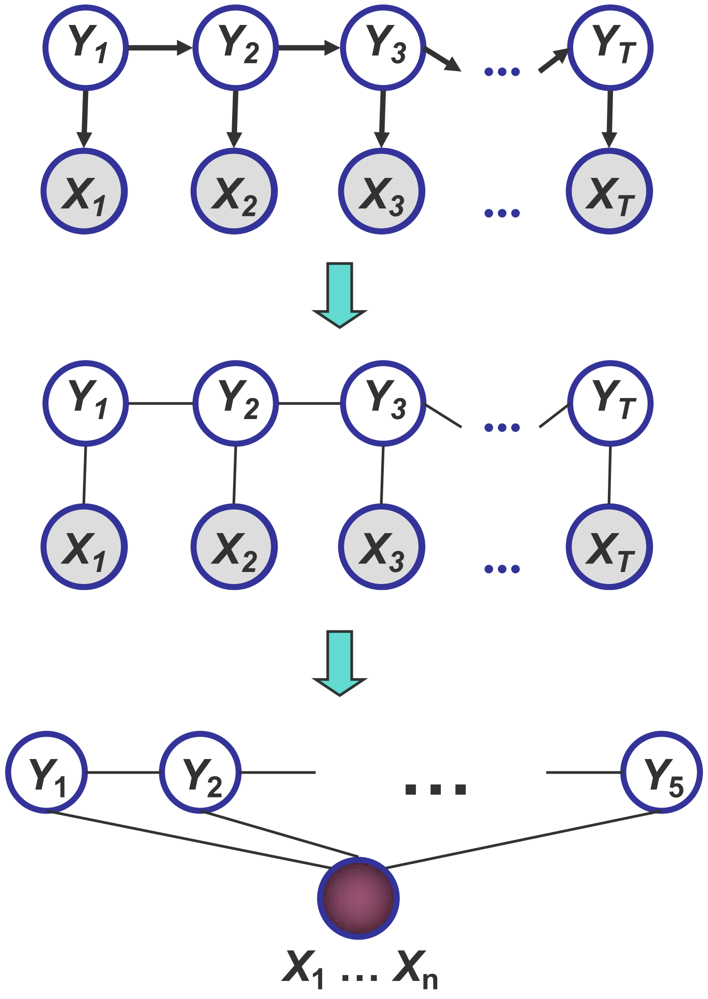
$x_i \in \mathbf{I}$

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

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

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

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(label sequence $\mathbf{y}$ | 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

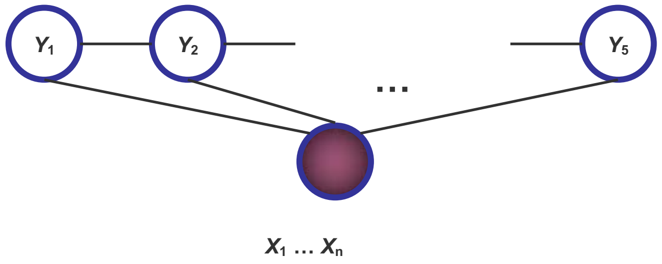
- Relax strong independence assumptions in generative models

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:

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

  - x is a data sequence
  - 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
  - $\theta = (\lambda_1, \lambda_2, \cdots, \lambda_n; \mu_1, \mu_2, \cdots, \mu_n); \lambda_k$ and μ_k are parameters to be estimated
  - $\mathrm{y}|_e$ is the set of components of y defined by edge e
  - $\mathrm{y}|_v$ is the set of components of y defined by vertex v



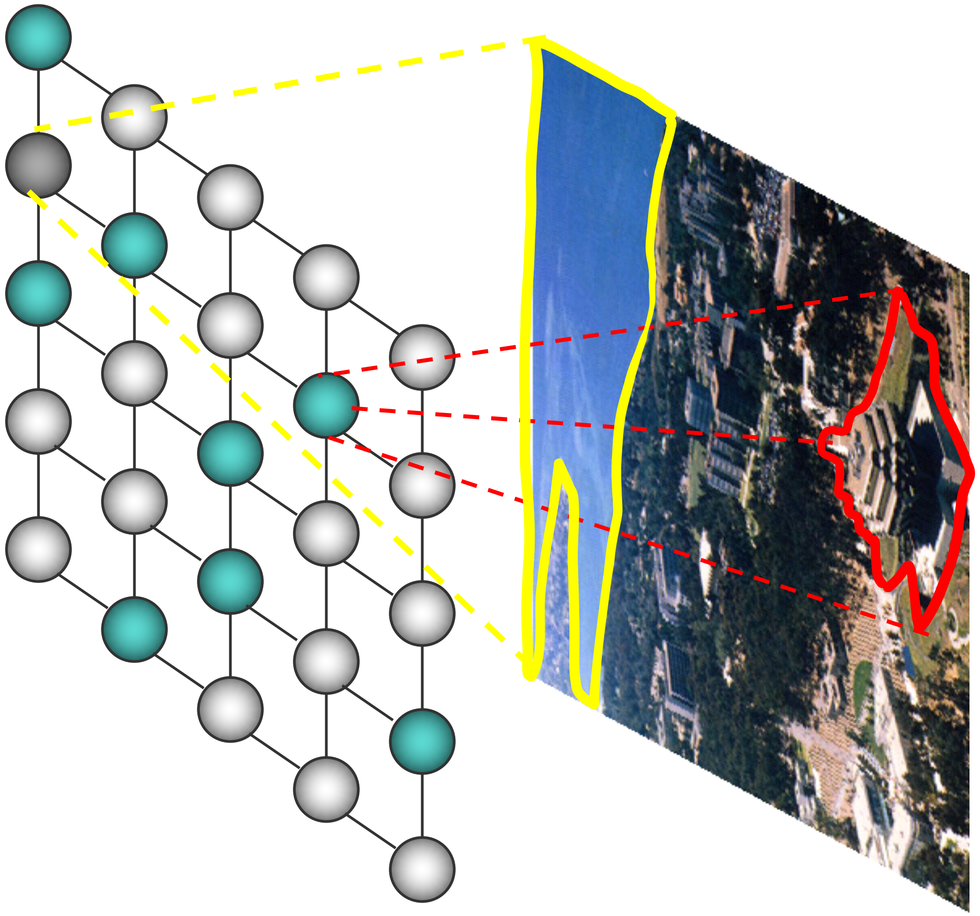
Conditional Distribution (cont'd)

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

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

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

Conditional Random Fields

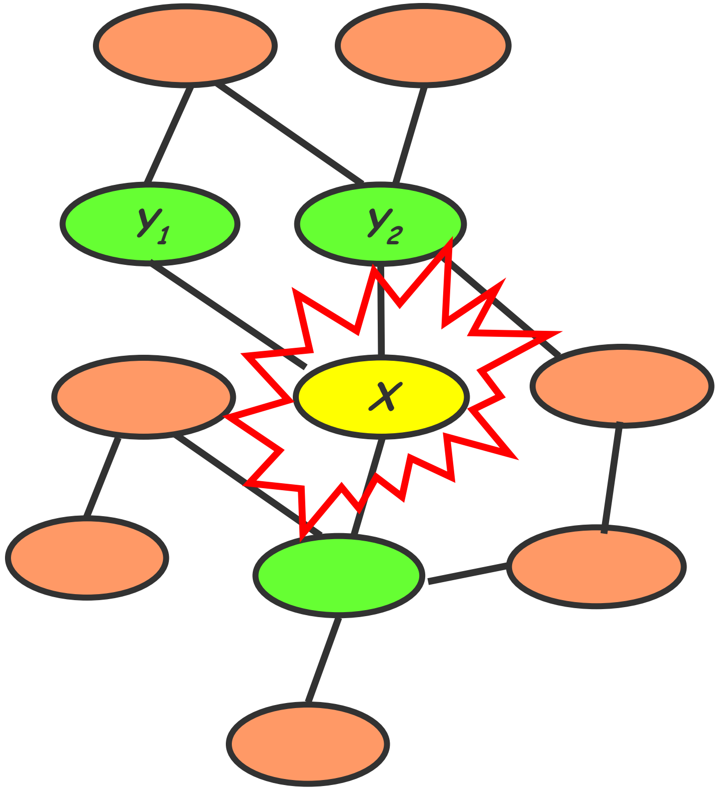


$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{\sum_c \theta_c f_c(x, y_c)\right\}$$

- 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

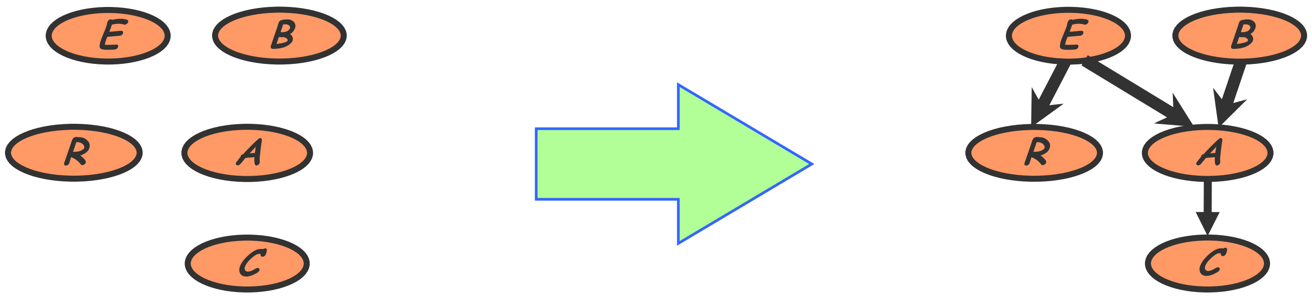
Supplementary:

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



(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)

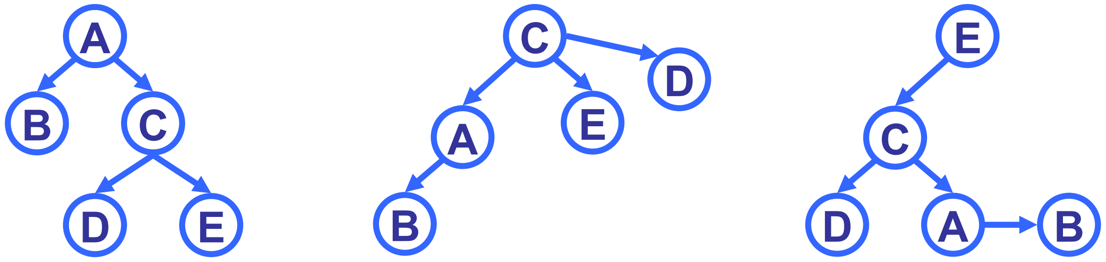
- Objection function:

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

- 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