[70240413 Statistical Machine Learning, Spring, 2018]

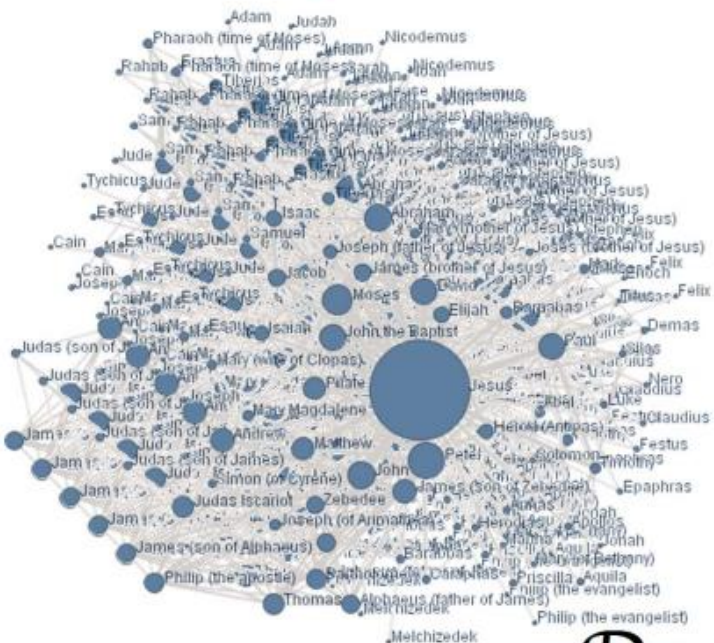
Probabilistic Graphical Models (I): Representation

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

May 22, 2018

What are Graphical Models?

Graph



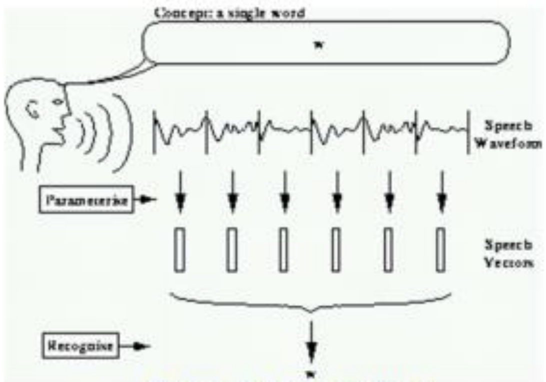
Model

$$\mathcal{M}_G$$

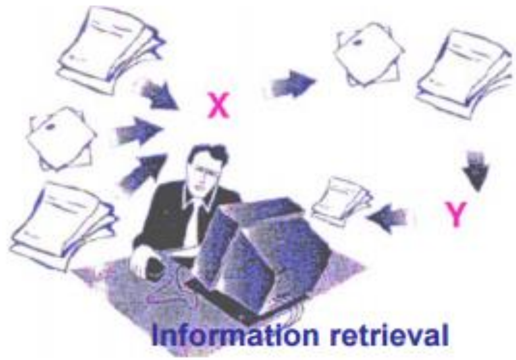
Data

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \ldots, X_m^{(i)}\}_{i=1}^{N}$$

Reasoning under uncertainty!

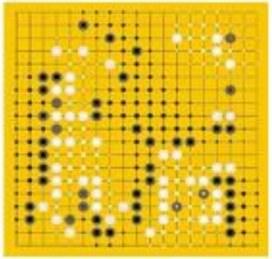


Speech recognition

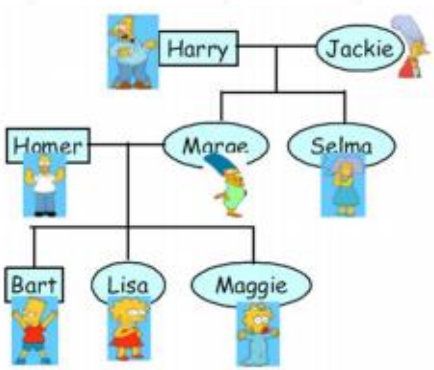


Information retrieval

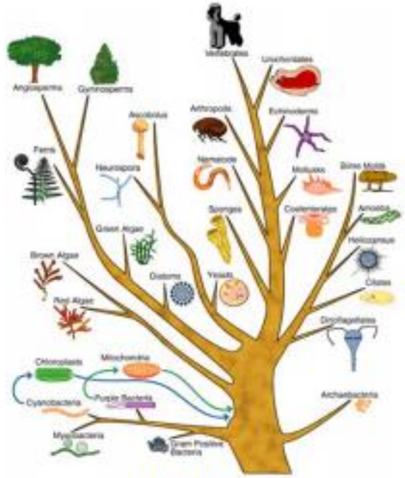

Computer vision


Games

Robotic control



Pedigree


Evolution


Planning

Three Fundamental Questions

- **Representation**
  - How to capture/model uncertainty in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?
- **Inference**
  - How do I answer questions/queries according to my model and/or based on given data?

$$\text{e.g.:}\quad P(X_i \mid \mathcal{D})$$

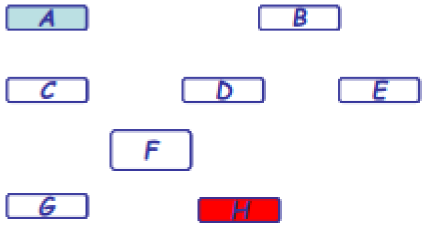
- **Learning**
  - What model is "right" for my data?

$$\text{e.g.:}\quad \mathcal{M} = \arg\max_{\mathcal{M} \in \mathrm{M}} F(\mathcal{D}; \mathcal{M})$$

Recap of Basic Prob. Concepts

- **Representation**: what is the joint prob. distribution on multiple variables

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



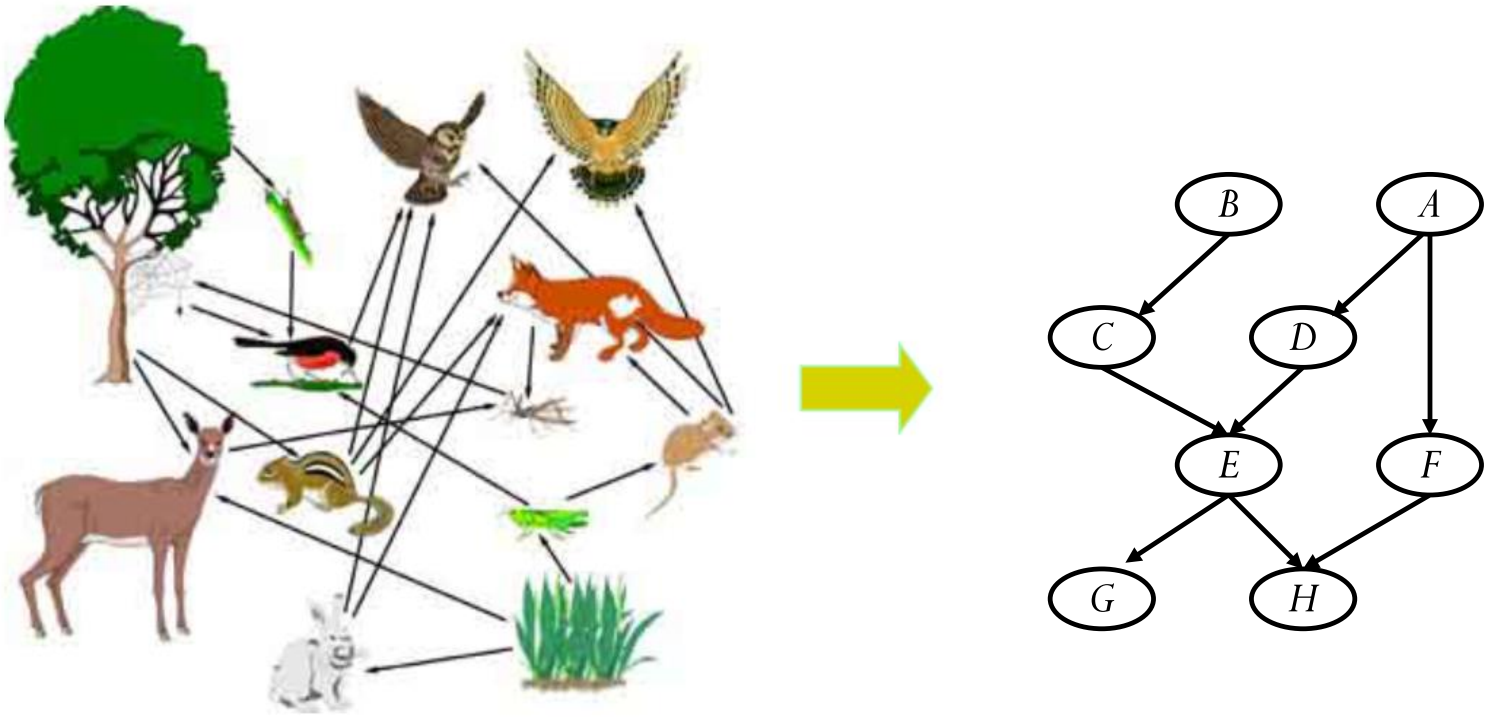
  - How many state configurations in total?
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?

- **Learning**: where do we get all this probabilities?
  - Maximum likelihood estimation? But how many data do we need?
  - Are there other estimation principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of probabilities?

- **Inference**: if not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing $p(\mathrm{H}|\mathrm{A})$ would require summing over all configurations of the unobserved variables

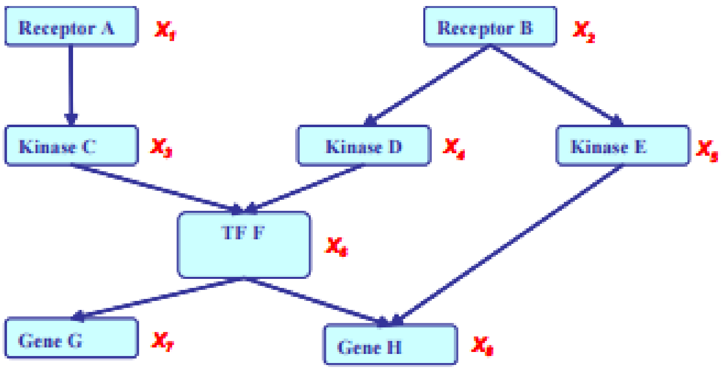
What is a Graphical Model?

- A multivariate distribution in high-dimensional space!
- An example with food web:



Probabilistic Graphical Models

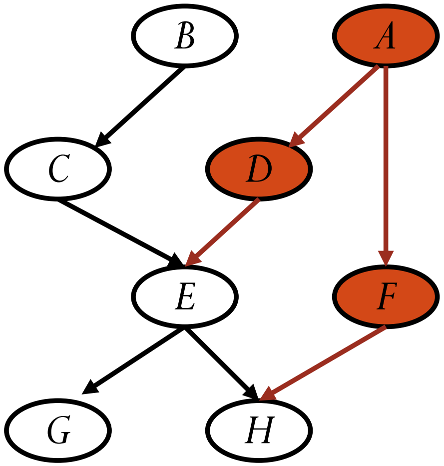
- If X_i's are conditionally independent (as described by a PGM), the joint can be factorized into a product of simpler terms, e.g.:



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

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures
    - How many parameters in the above factorized distribution?

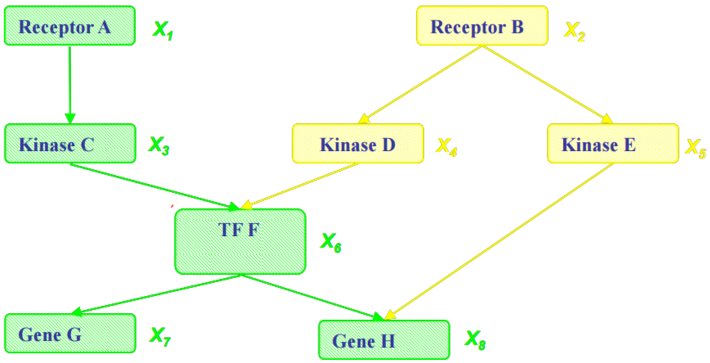
PGM: Data Integration



- More examples:
  - Text + Image + Network → Holistic Social Media

Probabilistic Graphical Models

- If X_i's are conditionally independent (as described by a PGM), the joint can be factorized into a product of simpler terms, e.g.:

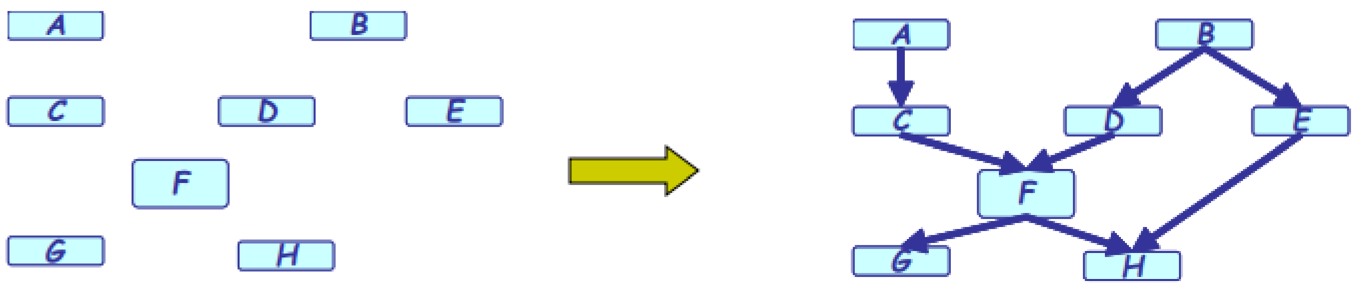


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_2)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_1)\, P(X_3 | X_1)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures
    - How many parameters in the above factorized distribution?
  - Modular combination of heterogeneous parts – data fusion!

So What is a PGM after all?

- **The informal blurb**:
  - It is a smart way to specify exponentially large prob. distributions without paying an exponential cost, and at the same time endow the distributions with structured semantics



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

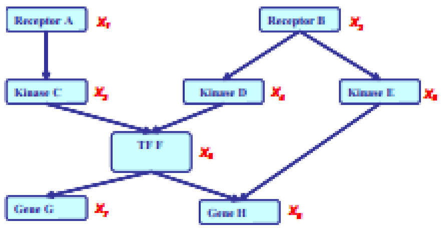
$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1 X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

- **A more formal description**:
  - It refers to a family of distributions on a set of RVs that are compatible with all the probabilistic independence propositions encoded by the graph that connects these variables

Two Types of PGMs

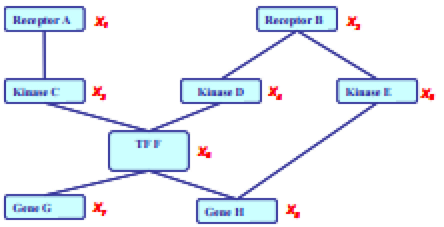
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Models)

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



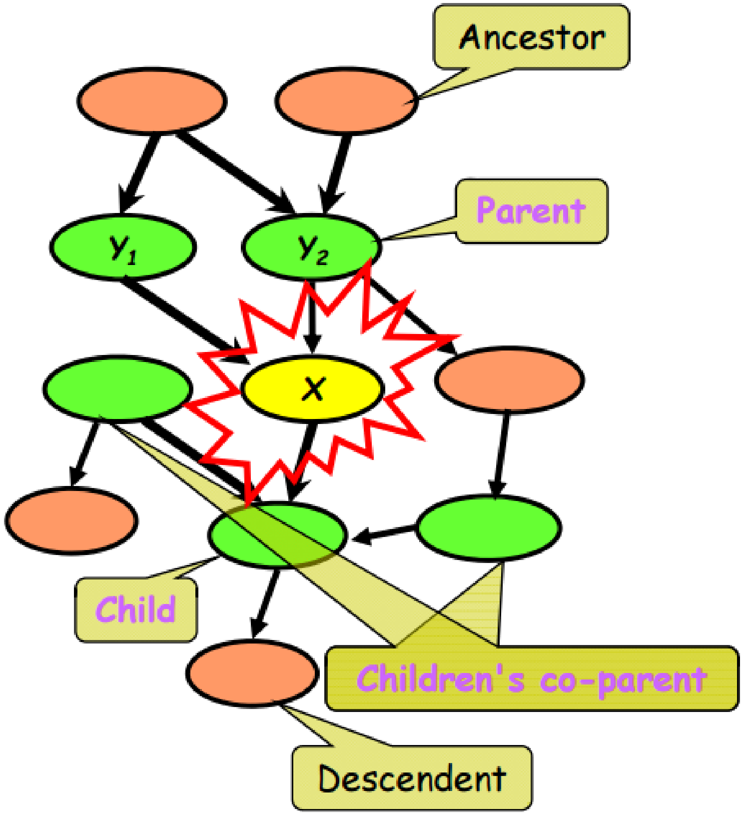
- Undirected edges give correlations between variables (Markov Random Field or Undirected Graphical Models)

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



Bayesian Networks

- Structure: ***DAG***
- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**
- Local conditional distributions (**CPD**) and the **DAG** completely determine the **joint** distribution

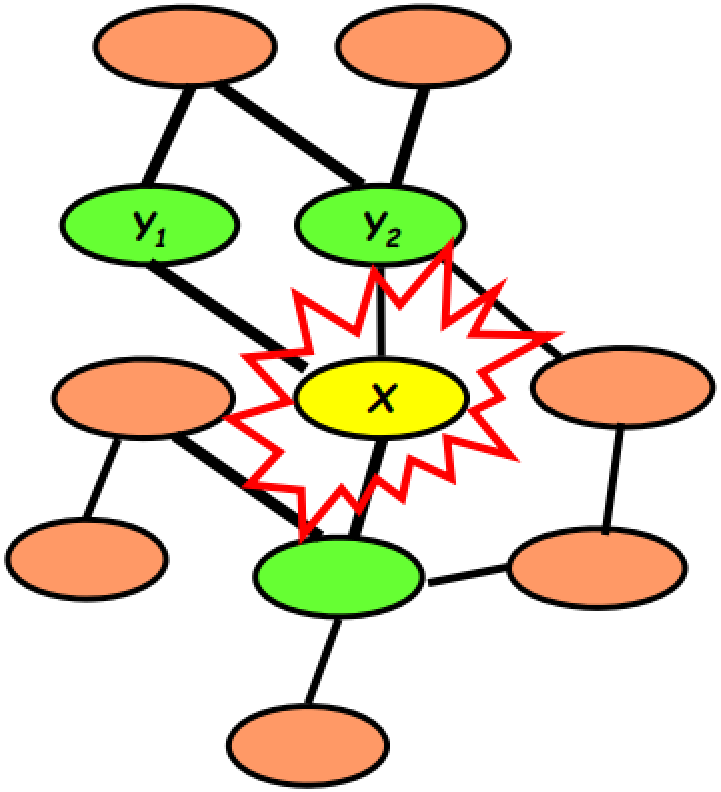


Markov Random Fields

- Structure: ***undirected graph***

- Meaning: a node is **conditionally independent** of every other node in the network given its **Direct Neighbors**

- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** distribution



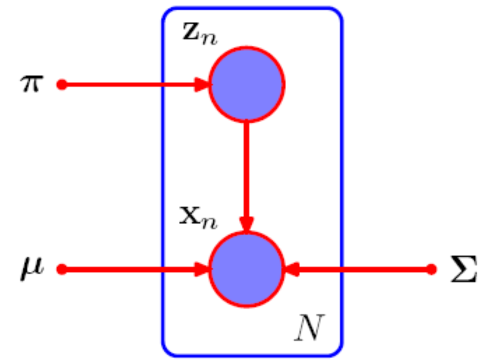
Towards Structural Specification of Probability Distribution

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

- The **Equivalence Theorem**:
  - For a graph G,
  - Let denote the family of distributions that satisfy I(G),
  - Let denote the family of distributions that factor according to G,
  - Then

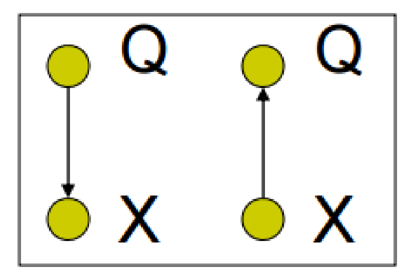
GMs are your old friends

- Clustering
  - GMMs

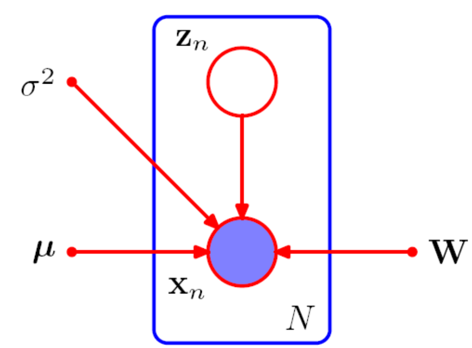


- Regression
  - Linear, conditional mixture

- Classification
  - Generative and discriminative approach



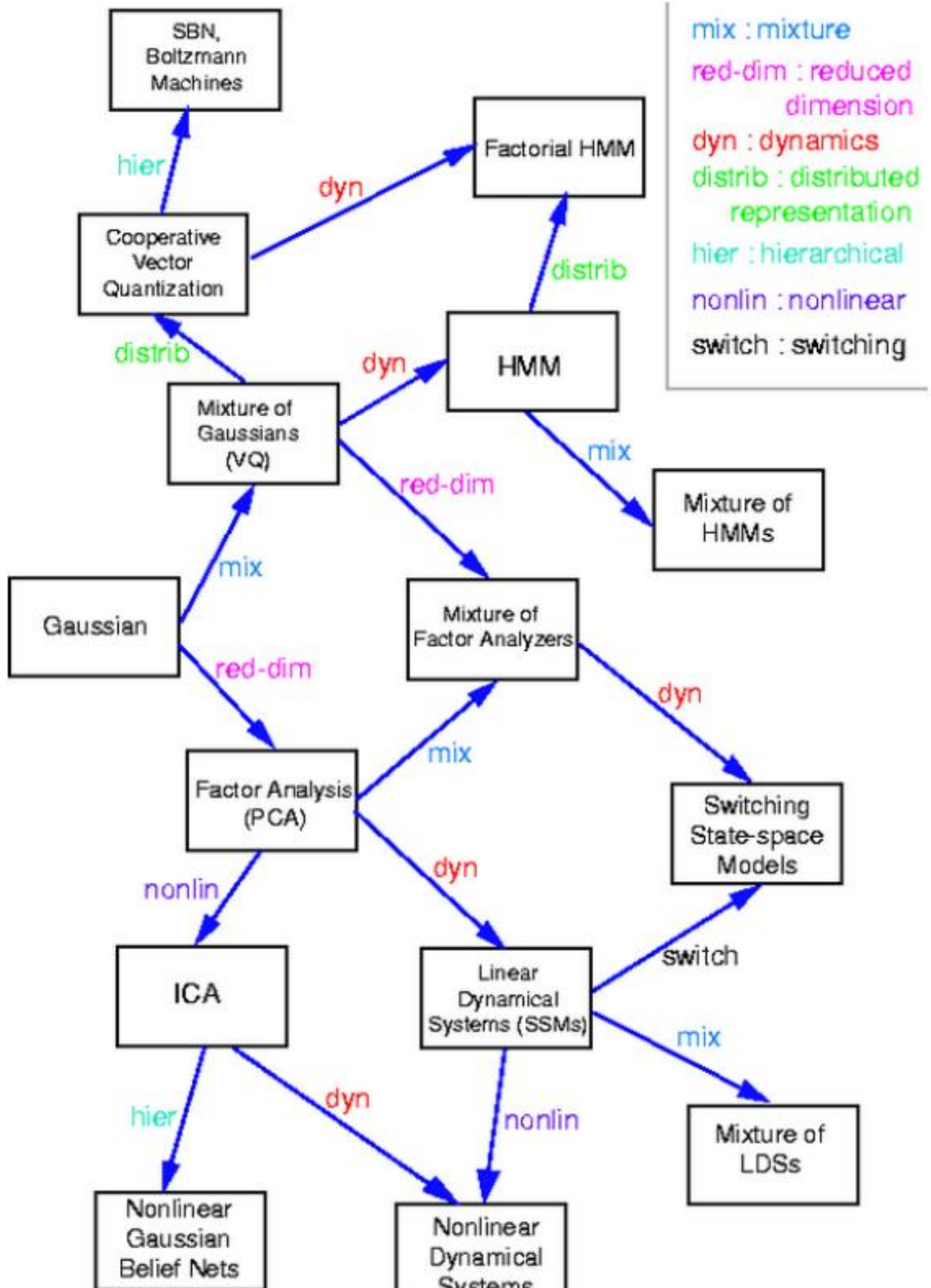
- Dimension reduction
  - PCA, FA, etc



- An (incomplete) genealogy of graphical models

1972-2010

- Picture by Zoubin Ghahramani & Sam Roweis

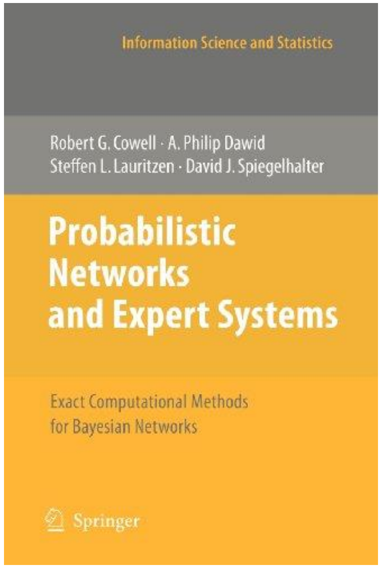


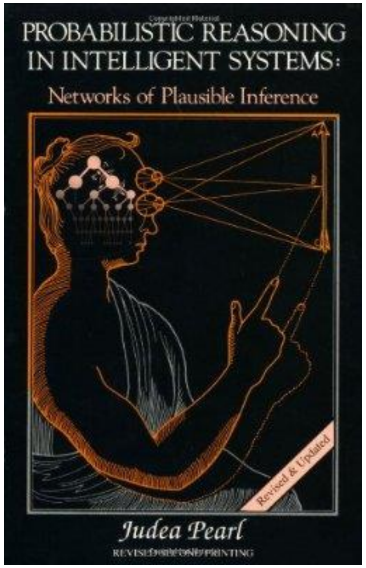
Application of PGMs

- Machine learning
- Computational statistics
- Computer vision and graphics
- Natural language processing
- Information retrieval
- Robot control
- Decision making under uncertainty
- Error-control codes
- Computational biology
- Genetics and medical diagnosis/prognosis
- Finance and economics
- Etc.

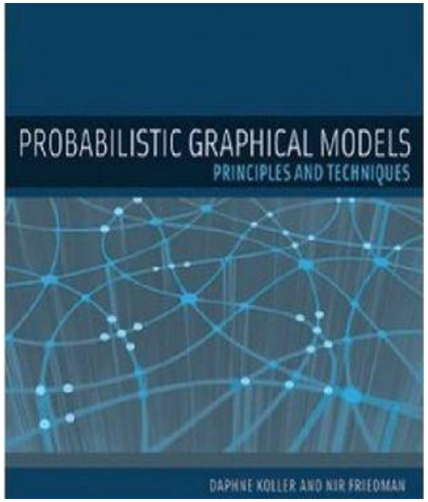
Why graphical models

- A language for communication
- A language for computation
- A language for development





- Origins:
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's



Why graphical models

- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data

- **Graph theory** provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms

- **Many of the classical multivariate probabilistic systems** studied in the fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**

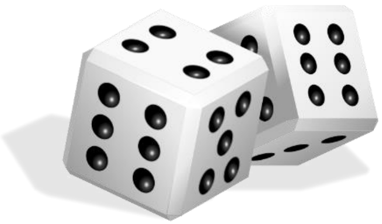
- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**

--- M. I. Jordan

Bayesian Networks

Example: The dishonest casino

- A casino has two dice:
  - Fair die: P(1)=P(2)=…=P(6)=1/6
  - Loaded die: P(1)=P(2)=…=P(5)=1/10; P(6)=1/2
- Casino player switches back & forth between fair and loaded die once every 20 turns



- Game:
  - You bet $1
  - You roll (always with a fair die)
  - Casino player rolls (maybe with fair die, maybe with loaded die)
  - Highest number wins $2

Puzzles regarding the dishonest casino

- **Given**: a sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

- **Questions**:
  - How likely is this sequence, given our model of how the casino works?
    - This is the **EVALUATION** problem
  - What portion of the sequence was generated with the fair die, and what portion with the loaded die?
    - This is the **DECODING** problem
  - How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?
    - This is the **LEARNING** problem

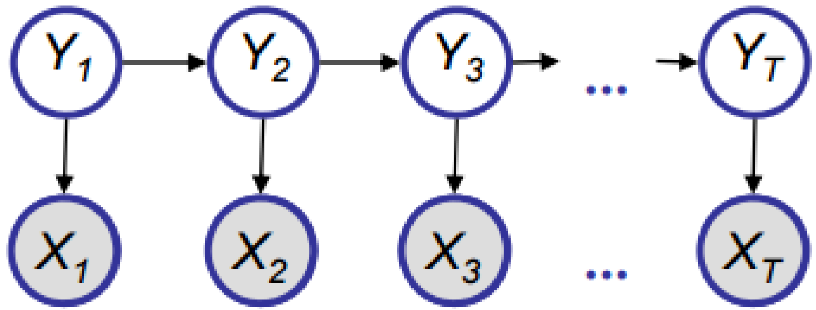
Hidden Markov Models (HMMs)

The underlying source:

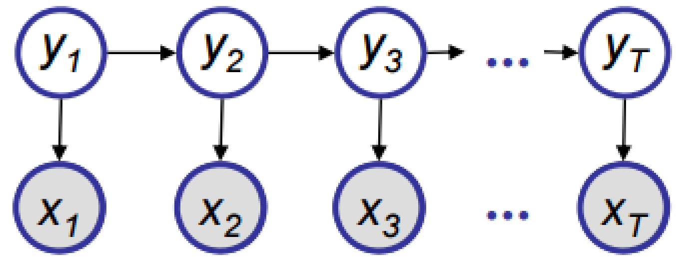
Speech signal
genome function
dice

The sequence:

Phonemes
DNA sequence
sequence of rolls



Probability of a parse



- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$
  and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T$
- To find how likely is the parse: (given our HMM and the sequence)

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}) &= p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \qquad \text{(Joint probability)} \\
&= p(y_1)\, p(x_1 \mid y_1)\, p(y_2 \mid y_1)\, p(x_2 \mid y_2) \ldots p(y_T \mid y_{T-1})\, p(x_T \mid y_T) \\
&= p(y_1)\, P(y_2 \mid y_1) \ldots p(y_T \mid y_{T-1}) \times p(x_1 \mid y_1)\, p(x_2 \mid y_2) \ldots p(x_T \mid y_T) \\
&= p(y_1, \ldots\ldots, y_T)\, p(x_1 \ldots\ldots x_T \mid y_1, \ldots\ldots, y_T)
\end{aligned}
$$

  - Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \pi_{y_1} \prod_{t=2}^{t} a_{y_{t-1}, y_t} \prod_{t=1}^{t} p(x_t \mid y_t)$
  - Posterior probability: $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$
- We will learn how to do this explicitly (**polynomial time**)

Bayesian Networks in a Nutshell

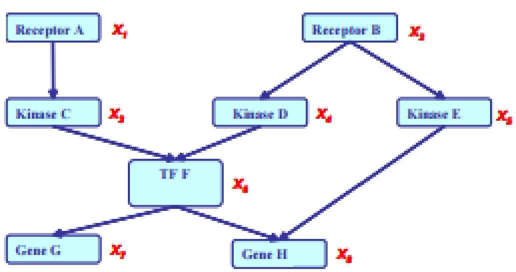
- A BN is a directed graph whose nodes represent the RVs and whose edges represent direct influence of one variable on another
- It is a data structure that provides the skeleton for representing a **joint distribution** compactly in a **factorized** way
- It offers a compact representation for **a set of conditional independence assumptions** about a distribution
- We can view the graph as encoding a **generative sampling process** executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents.

Bayesian Network: Factorization Theorem

- **Theorem**:
  - Given a DAG, the most general form of the probability distribution that is consistent with the graph factors according to "node given its parents":

$$P(\mathbf{X}) = \prod_{i=1:d} P(X_i \mid \mathbf{X}_{\pi_i})$$

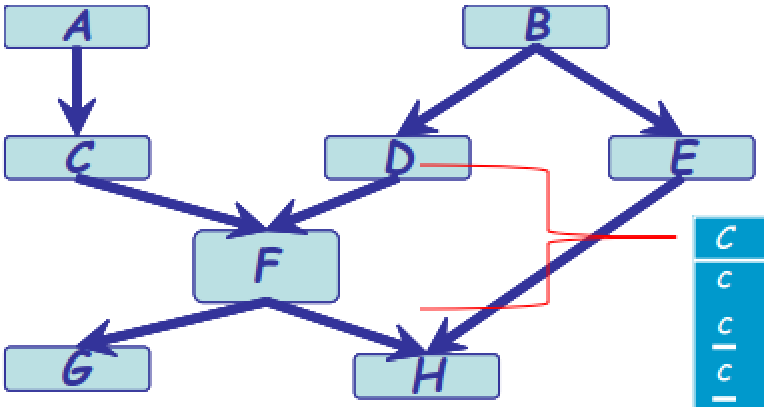
  - where $\mathbf{X}_{\pi_i}$ is the set of parents of X_i, d is the number of nodes (variables) in the graph



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

Specification of a Directed GM

- There are two components to any GM:
  - The qualitative specification
  - The quantitative specification



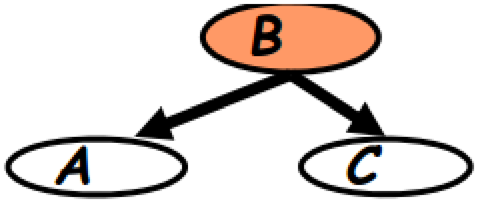
C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply like a certain architecture (e.g., a layered graph)
  - …

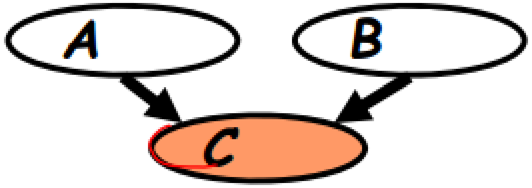
Local Structure & Independence

- Common parent
  - Fixing B decouples A and C



- Cascade
  - Knowing B decouples A and C

- V-structure
  - Knowing C couples A and B because A can “explain away” B w.r.t C



- The language is compact, the concepts are rich!

I-Maps

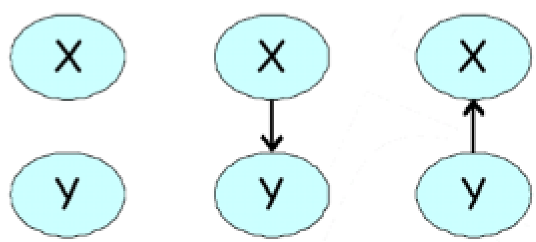
- **Defn :** Let P be a distribution over $\mathcal{X}$. We define I(P) to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P (however how we set the parameter-values).

- **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, $I(K) \subseteq I$.

- 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* 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*
- Example: (who is P1 / P2's I-map?)



$\mathcal{G}_\emptyset$ $\quad$ $\mathcal{G}_{X \to Y}$ $\quad$ $\mathcal{G}_{Y \to X}$

P_1

X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

P_2

X	Y	$P(X,Y)$
x^0	y^0	0.4
x^0	y^1	0.3
x^1	y^0	0.2
x^1	y^1	0.1

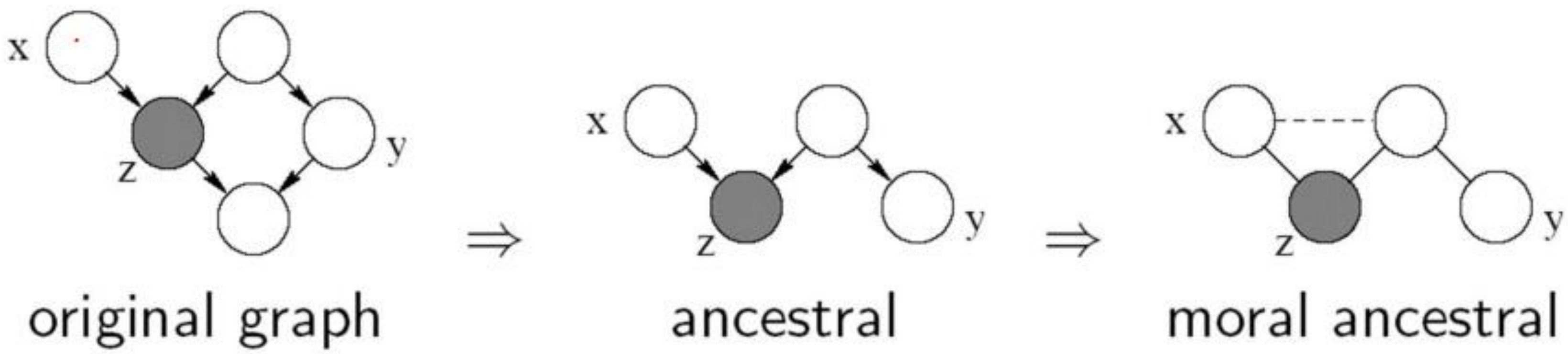
- Complete graph is an I-map for any distribution, right?
  - Yet it does not reveal any independence structure in the distribution

Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

  Defn: variables x and y are *D-separated* (conditionally independent) given z if they are separated in the *moralized* ancestral graph

- Example:



Active trail

- **Causal trail** X → Z → Y **:** active if and only if Z is not observed.

- **Evidential trail** X ← Z ← Y **:** active if and only if Z is not observed.

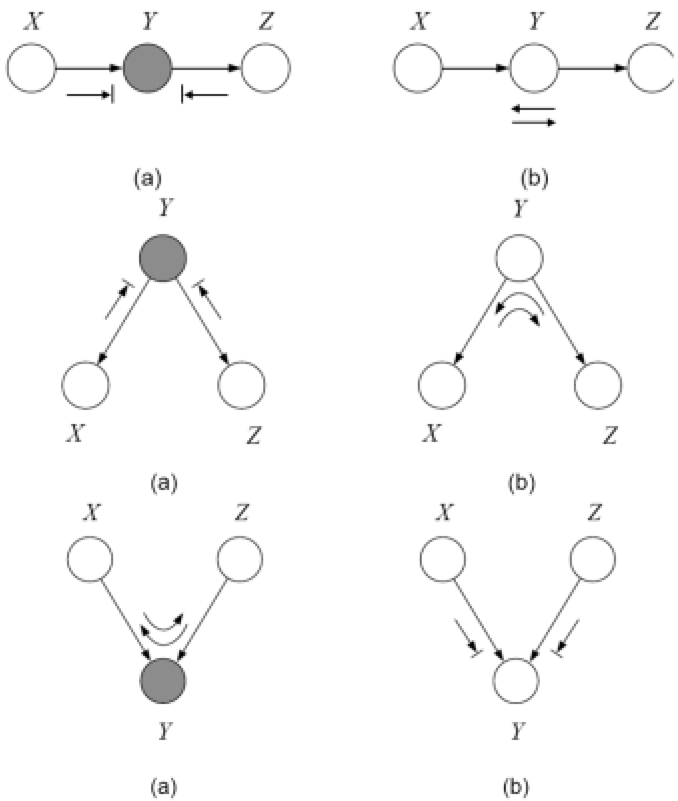
- **Common cause** X ← Z → Y **:** active if and only if Z is not observed.

- **Common effect** X → Z ← Y **:** active if and only if either Z or one of Z's descendants is observed

Definition : Let X, Y , Z be three sets of nodes in G. We say that X and Y are *d-separated given* Z, denoted *d-sep*$_{\mathcal{G}}(X;Y \mid Z)$, if there is no active trail between any node $X \in X$ and $Y \in Y$ given Z.

What is in I(G): Global Markov Property

- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

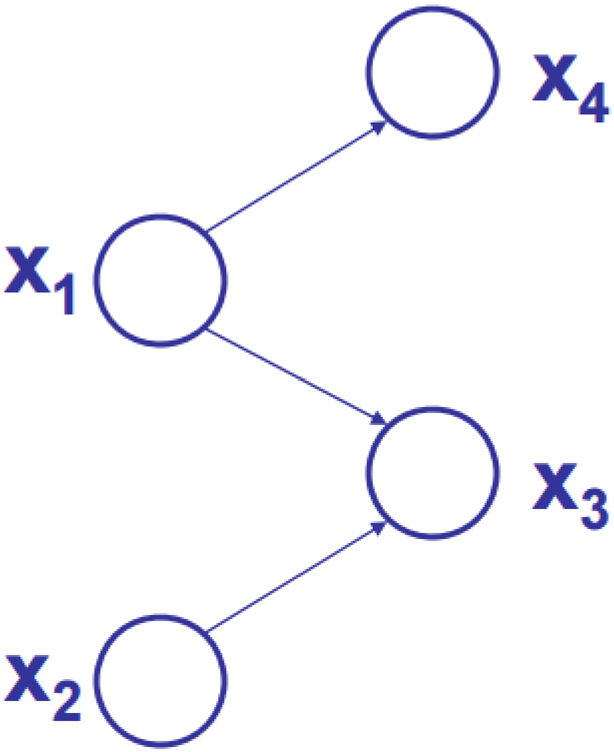


- **Defn: $\mathcal{I}(G)$=all independence properties that correspond to d-separation:**

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

Example

- Complete the I(G) of this graph:



Quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables

- The Equivalence Theorem:

  For a graph G,

  Let $\mathcal{D}_1$ denote the family of **all distributions** that satisfy I(G),

  Let $\mathcal{D}_2$ denote the family of **all distributions** that factor according to G,

  $$P(\mathbf{X}) = \prod_{i=1:d} P(X_i \mid \mathbf{X}_{\pi_i})$$

  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

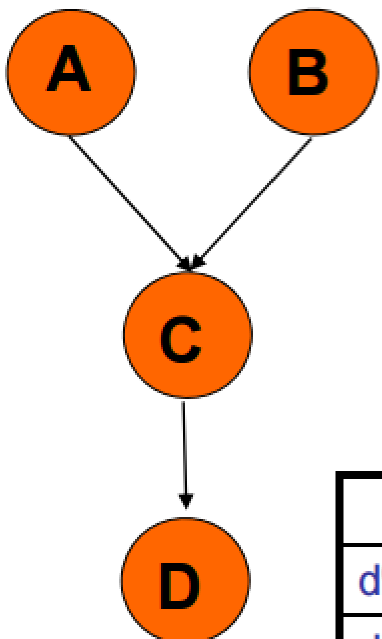
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

Conditional Probability Tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

P(a,b,c.d) =
P(a)P(b)P(c|a,b)P(d|c)



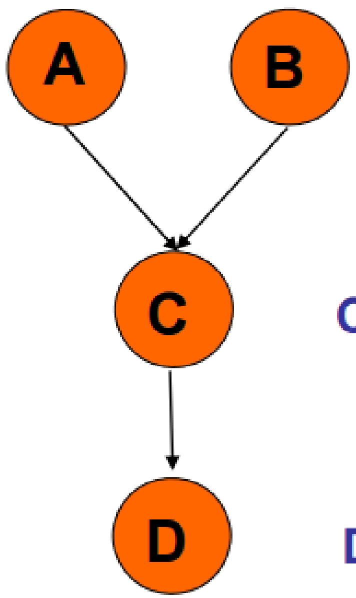
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional Probability Density Functions (CPDs)

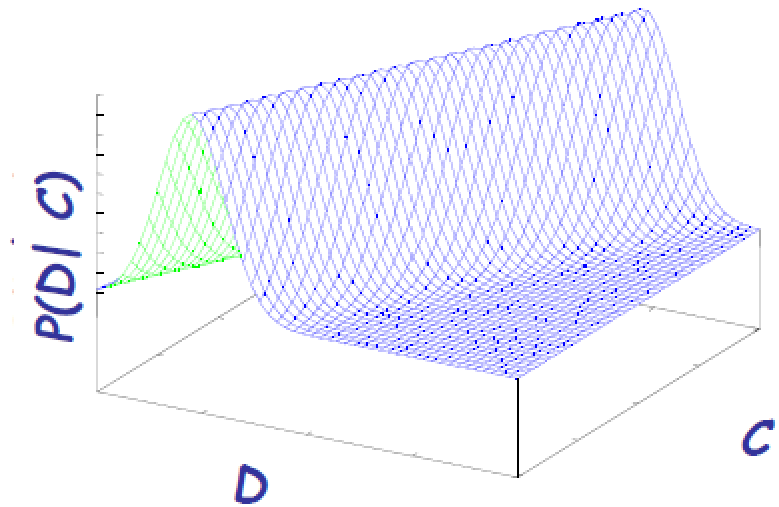
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

P(a,b,c.d) =
P(a)P(b)P(c|a,b)P(d|c)



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d + C, \Sigma_d)$



Summary of BN Semantics

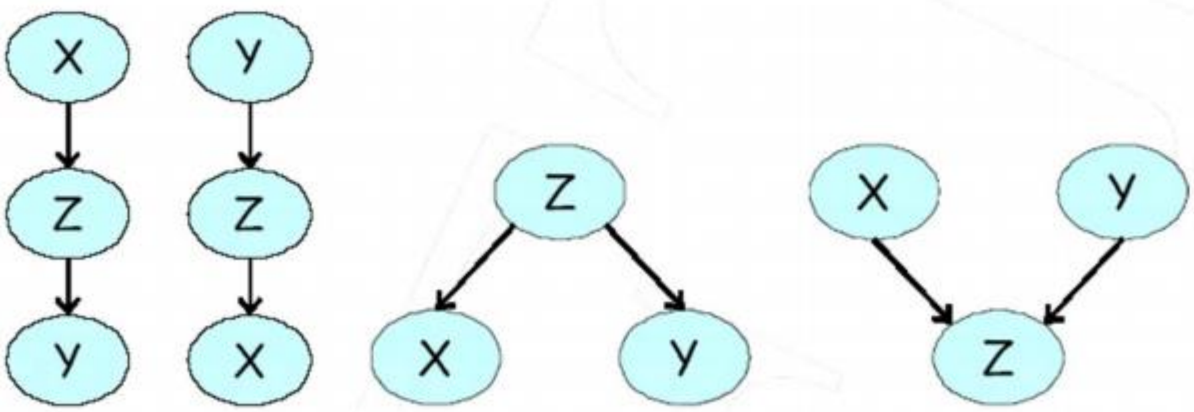
- **Defn :** A *Bayesian network* is a pair (G, P) where P factorizes over G, and where P is specified as set of CPDs associated with G's nodes.

  - Conditional independencies imply factorization
  - Factorization according to G implies the associated conditional independencies.

- D-separation is sound and complete w.r.t BN factorization law

Uniqueness of BN

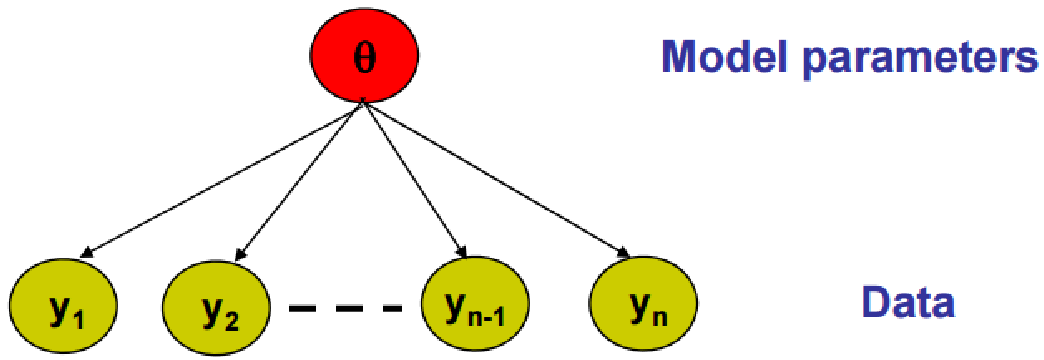
- Very different BN graphs can actually be equivalent, in that they encode precisely the same set of conditional independence assertions.



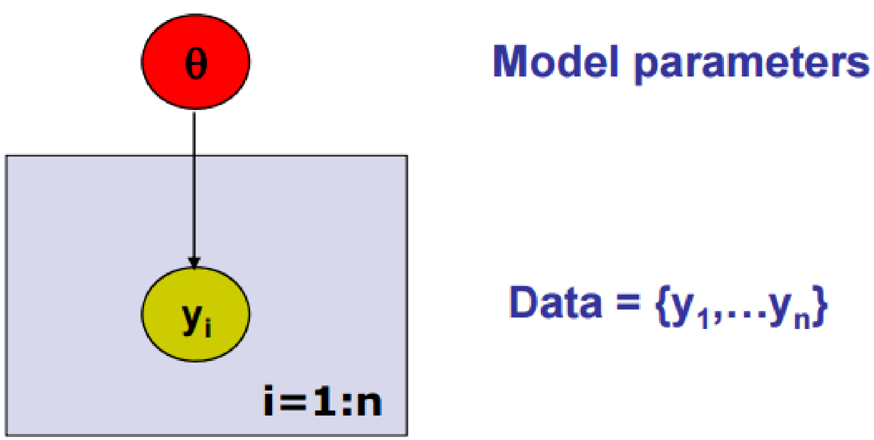
(a) (b) (c) (d)

$(X \perp Y \mid Z)$.

Simple BNs: Conditionally Indep. Observations

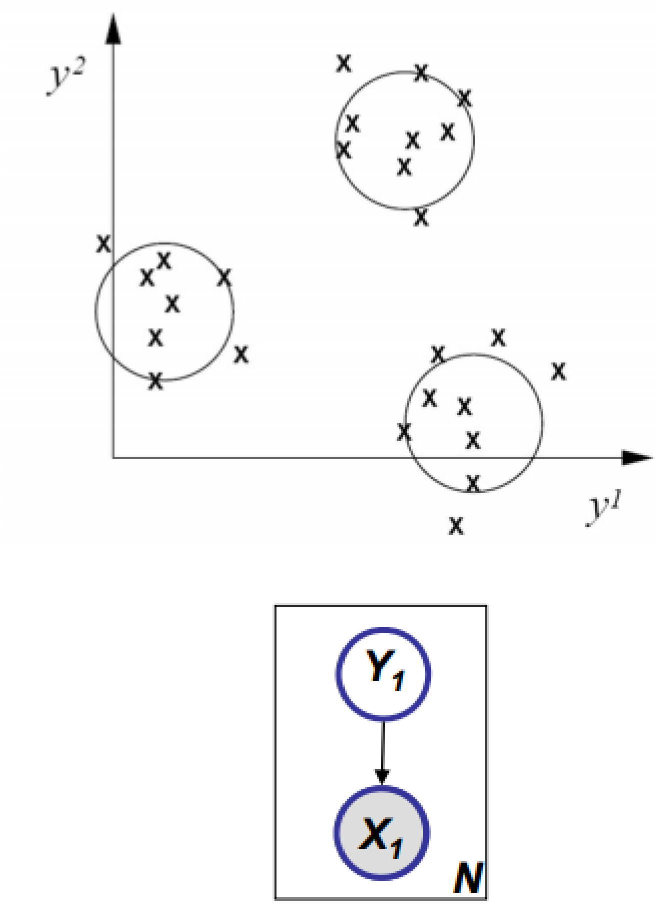


- The "Plate" Micro:

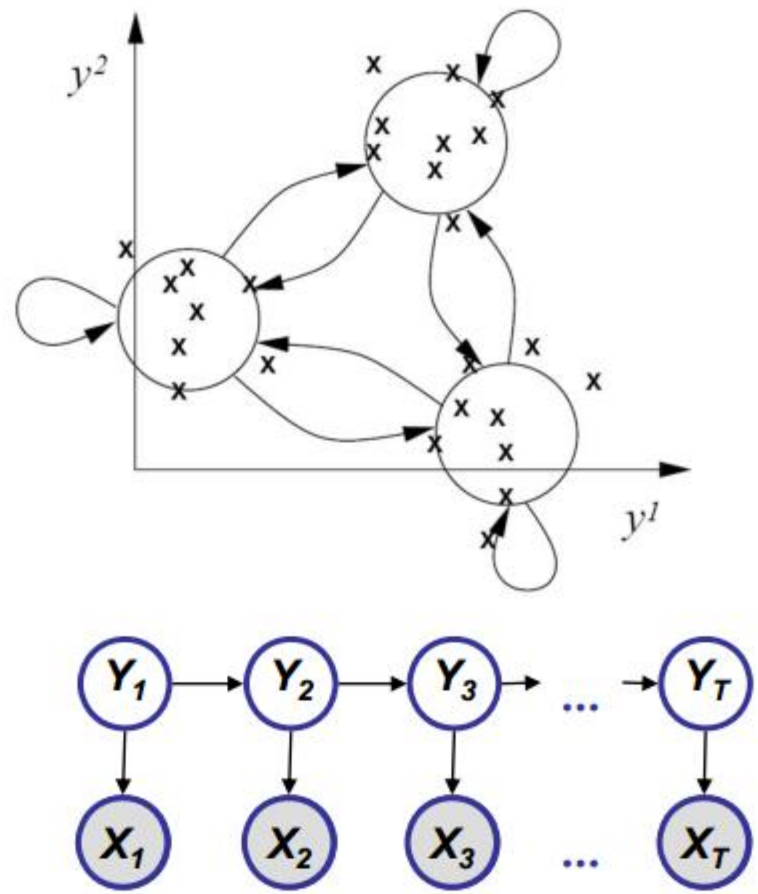


Hidden Markov Model: from static to dynamic mixture

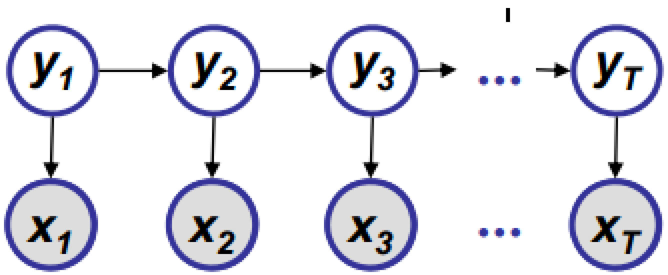
Static mixture



Dynamic mixture



Definition of HMM



- Observation space
  - **Alphabetic set:** $\mathbb{C} = \{c_1, c_2, \cdots, c_K\}$
  - **Euclidean space:** $\mathbb{R}^d$
- Index set of hidden states

  $\mathbb{I} = \{1, 2, \cdots, M\}$
- Transition probabilities between any two states

  $p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j}$,

  or $p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,1}, \ldots, a_{i,M}), \forall i \in \mathbb{I}.$
- Start probabilities

  $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$
- Emission probabilities associated with each state

  $p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,1}, \ldots, b_{i,K}), \forall i \in \mathbb{I}.$

  or in general:

  $p(x_t \mid y_t^i = 1) \sim f(\cdot \mid \theta_i), \forall i \in \mathbb{I}.$

Markov Random Fields

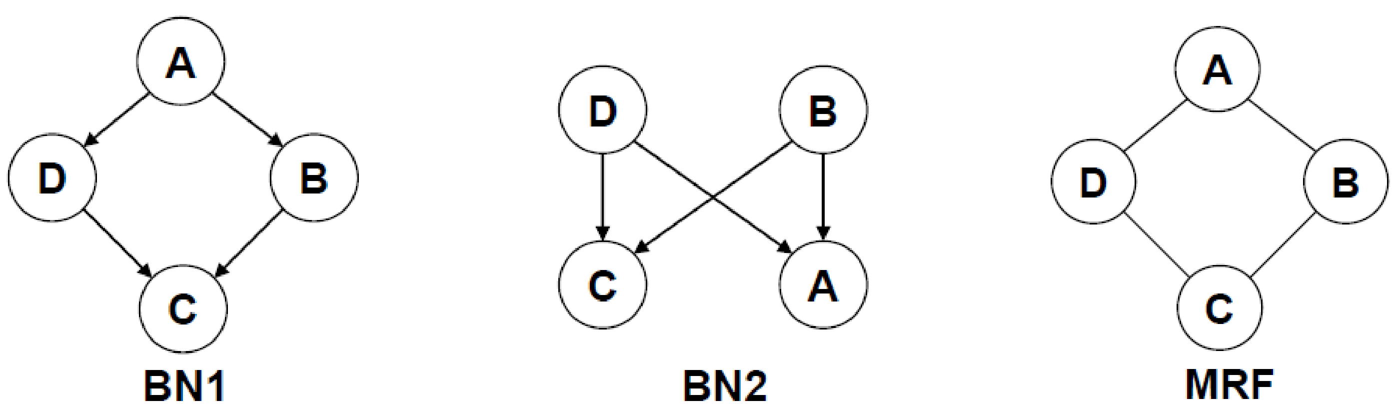
P-maps

- **Definition**: A DAG G is a perfect map (P-map) for a distribution P is $\mathrm{I}(P) = \mathrm{I}(G)$

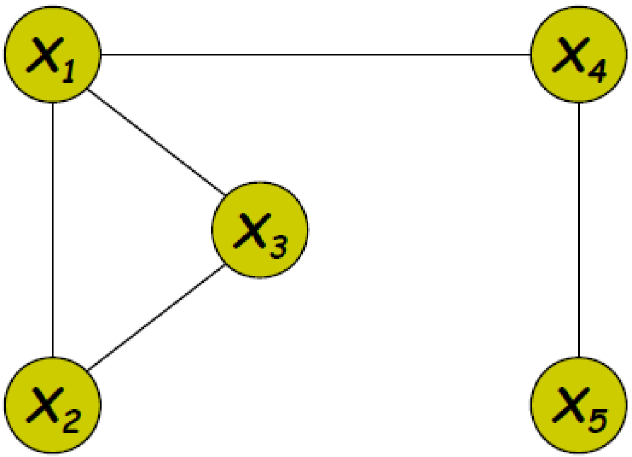
- **Theorem**: not every distribution has a perfect map as DAG
  - Proof by counterexample: suppose we have a model where

$$A \perp C \mid \{B, D\}, \text{ and } B \perp D \mid \{A, C\}.$$

  - This cannot be represented by any Bayes net

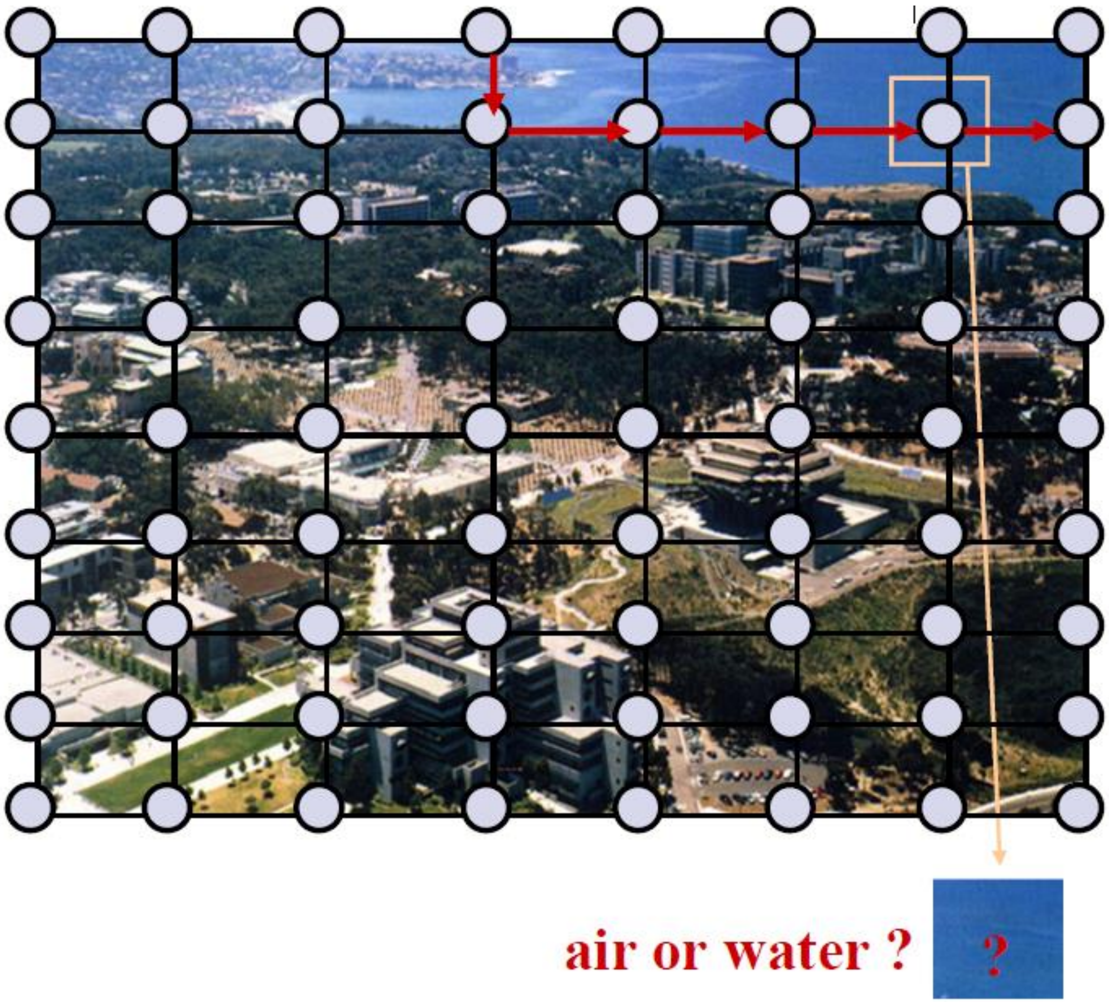


Undirected Graphical Models (UGM)



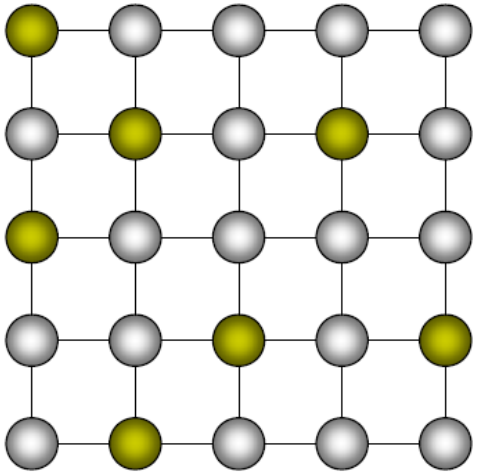
- Pairwise (non-causal) relationships
- Can write down model, and score specific configurations of the graph, but no explicit way to generate samples
- Contingency constrains on node configuration

A Canonical Example: understanding complex scene



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

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.

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

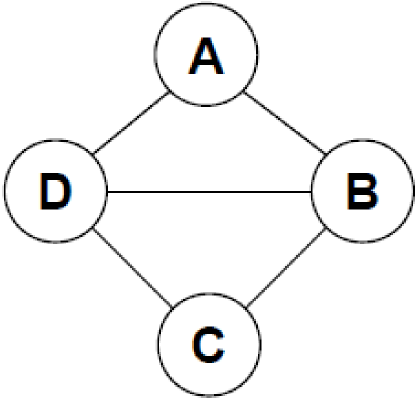
where Z is known as the partition function:

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

- 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.

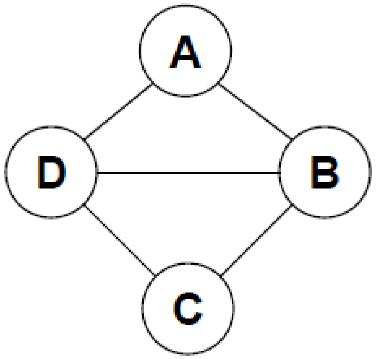
I. Quantitative Specification: Cliques

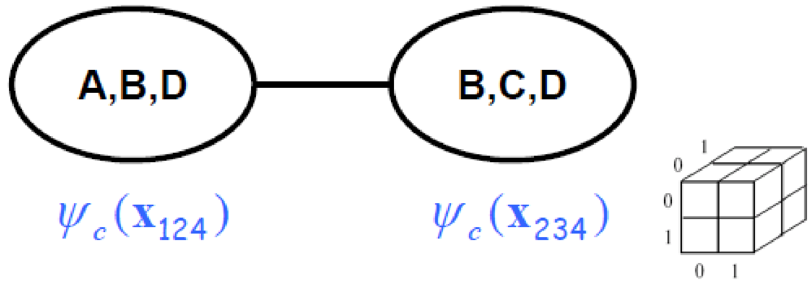
- For $G=\{V,E\}$, a complete subgraph (clique) is a subgraph $G'=\{V'\subseteq V, E'\subseteq E\}$ such that nodes in V' are fully interconnected
- A (maximal) clique is a complete subgraph s.t. any **superset** $V''\supset V'$ is not complete.
- A sub-clique is a not-necessarily-maximal clique.



- Example:
  - max-cliques = $\{A,B,D\}$, $\{B,C,D\}$,
  - sub-cliques = $\{A,B\}$, $\{C,D\}$, …→ all edges and singletons

Example UGM – using max cliques



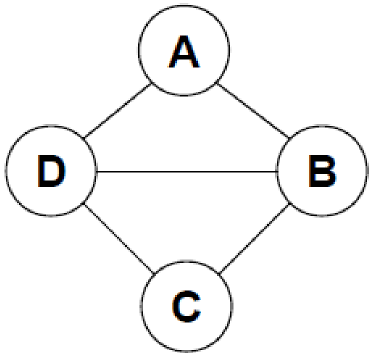


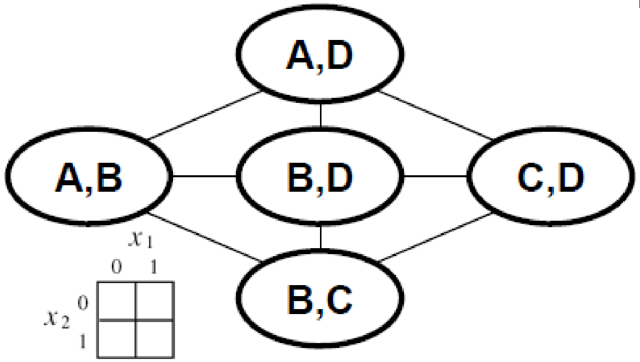
$$P'(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_c(\mathbf{x}_{124}) \times \psi_c(\mathbf{x}_{234})$$

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

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

Example UGM – using subcliques





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

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

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

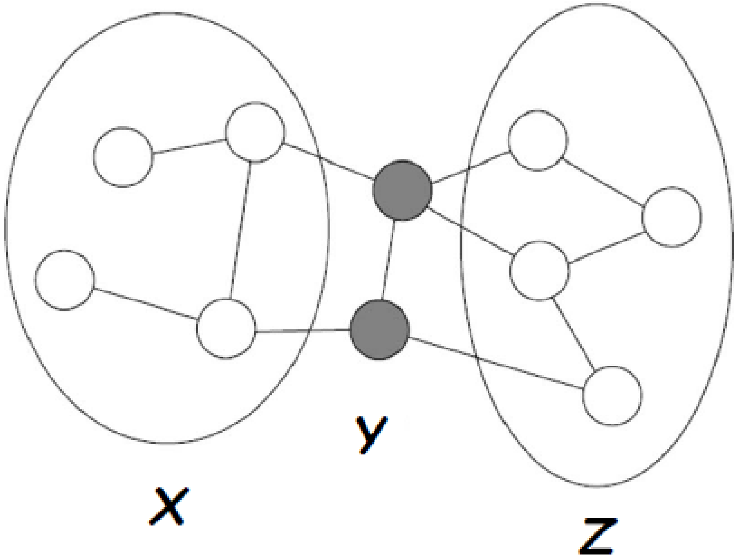
- We can represent $P(X_{1:4})$ as 5 2D tables instead of one 4D table
- Pair MRFs, a popular and simple special case

II: Independence Properties

- Now let us ask what kinds of distributions can be represented by undirected graphs (ignoring the details of the particular parameterization).

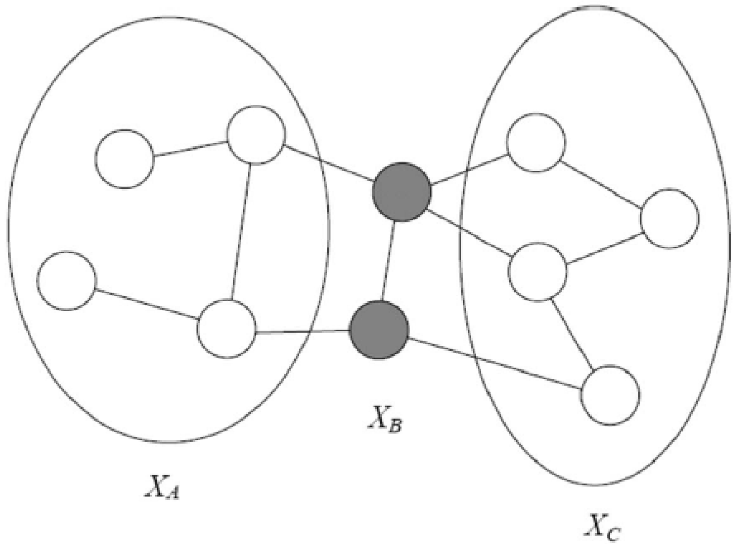
- Defn: the global Markov properties of a UG *H* are

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



Global Markov Properties

- Let *H* be an undirected graph:

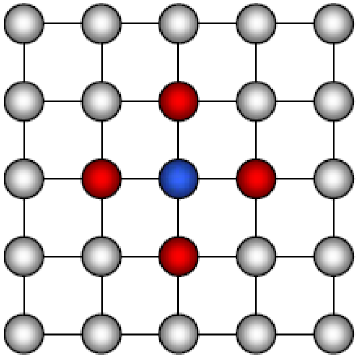


- *B* ***separates*** *A* and *C* if every path from a node in *A* to a node in *C* passes through a node in *B*: $\mathrm{sep}_H(A; C \mid B)$

- 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*: $\mathrm{I}(H) = \left\{ A \perp C \mid B : \mathrm{sep}_H(A; C \mid B) \right\}$

Local Markov Properties

- 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:

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

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 $I(H) \subseteq I(P)$, 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 $\neg \mathrm{sep}_H(X; Z \mid Y)$, then $X \not\perp_P Z \mid Y$ in ***some*** P that factorizes over H.

Hammersley-Clifford Theorem

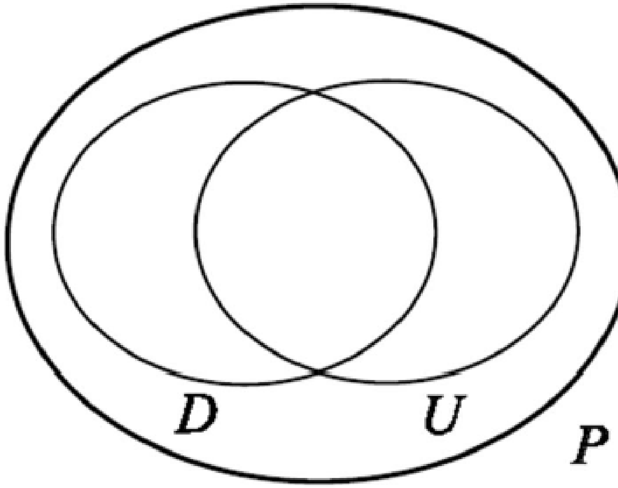
- **Thm :** Let P be a positive distribution over $\mathbf{V}$, and *H* a Markov network graph over $\mathbf{V}$. If <u>*H* is an I-map for P</u>, then P is a Gibbs distribution over *H*.

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

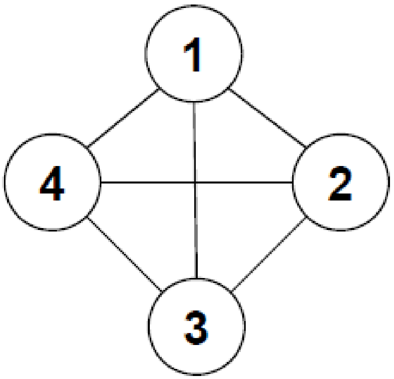
$$p(\mathbf{x}) = \frac{1}{Z}\exp\left\{-\sum_{c\in C}\phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z}\exp\{-H(\mathbf{x})\}$$

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

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

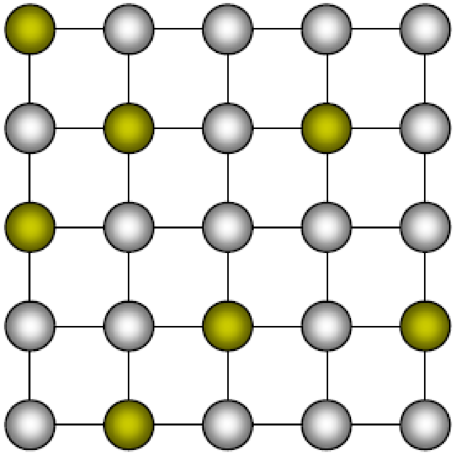
$$= \frac{1}{Z}\exp\left\{\sum_{ij}\theta_{ij}x_i x_j + \sum_i \alpha_i x_i + C\right\}$$

- 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 Model

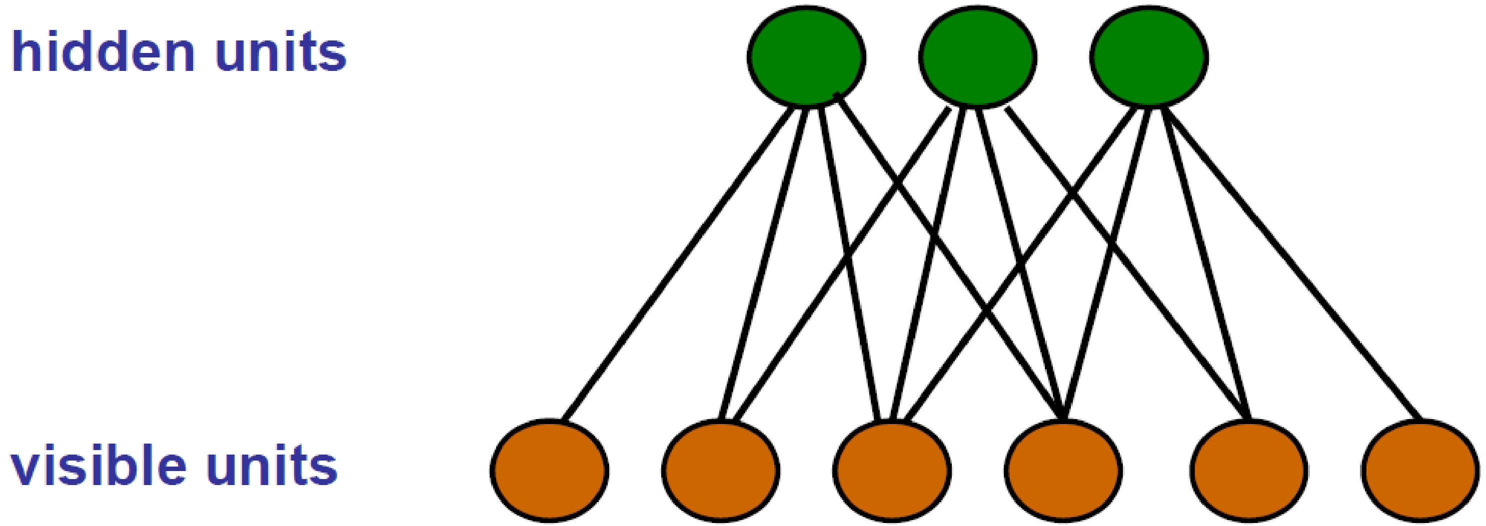
- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.



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

- 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.

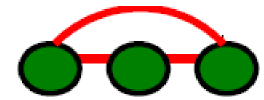
Restricted Boltzmann Machines



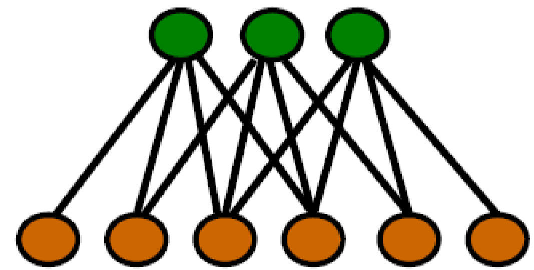
$$p(x,h \mid \theta) = \exp\left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\boldsymbol{\theta}) \right\}$$

Properties of RBM

- Factors are marginally *dependent*.

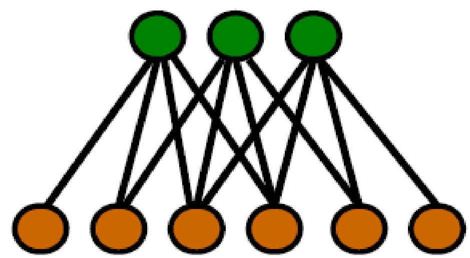


- Factors are conditionally *independent* given observations on the visible nodes.



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

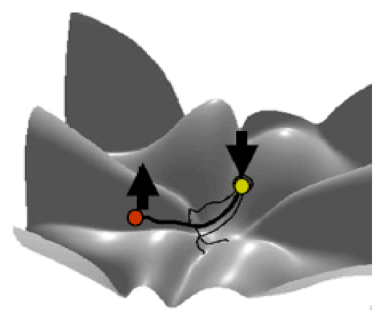
- Iterative Gibbs sampling.



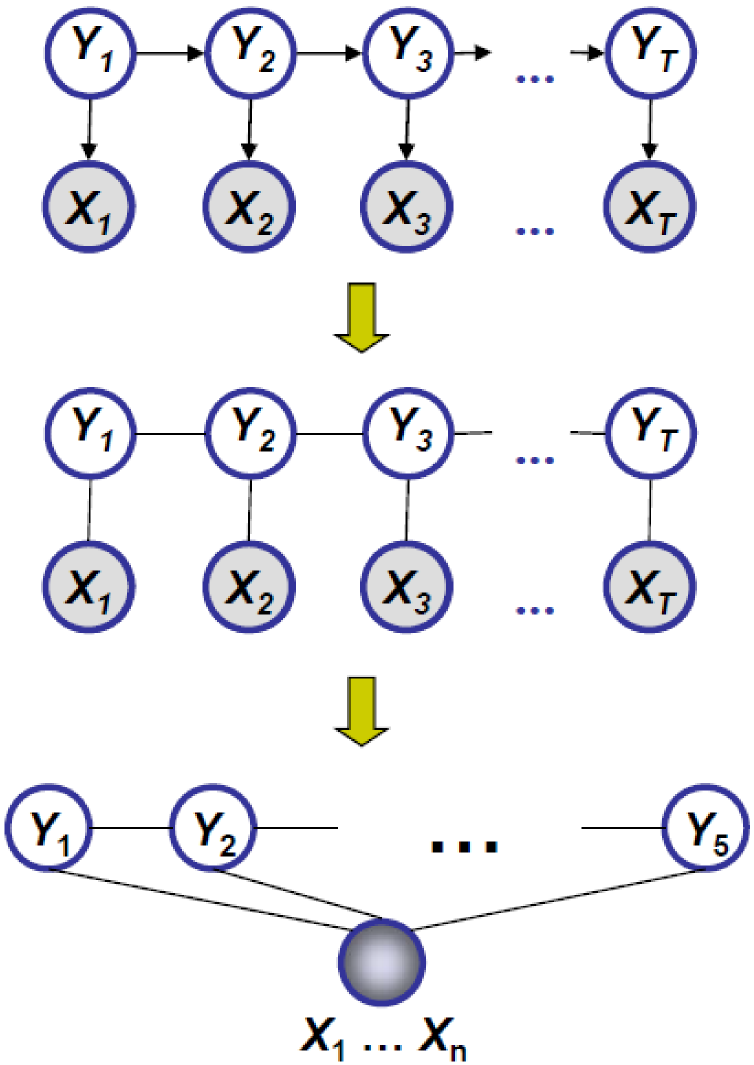
$$h \sim p(h \mid x)$$

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

- Learning with contrastive divergence



Conditional Random Fields



- Discriminative

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

- Doesn't assume that features are independent

- When labeling X_i future observations are taken into account

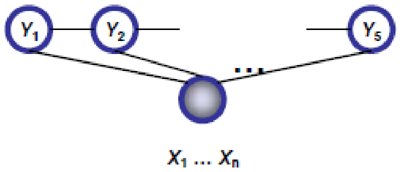
Conditional Models

- 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

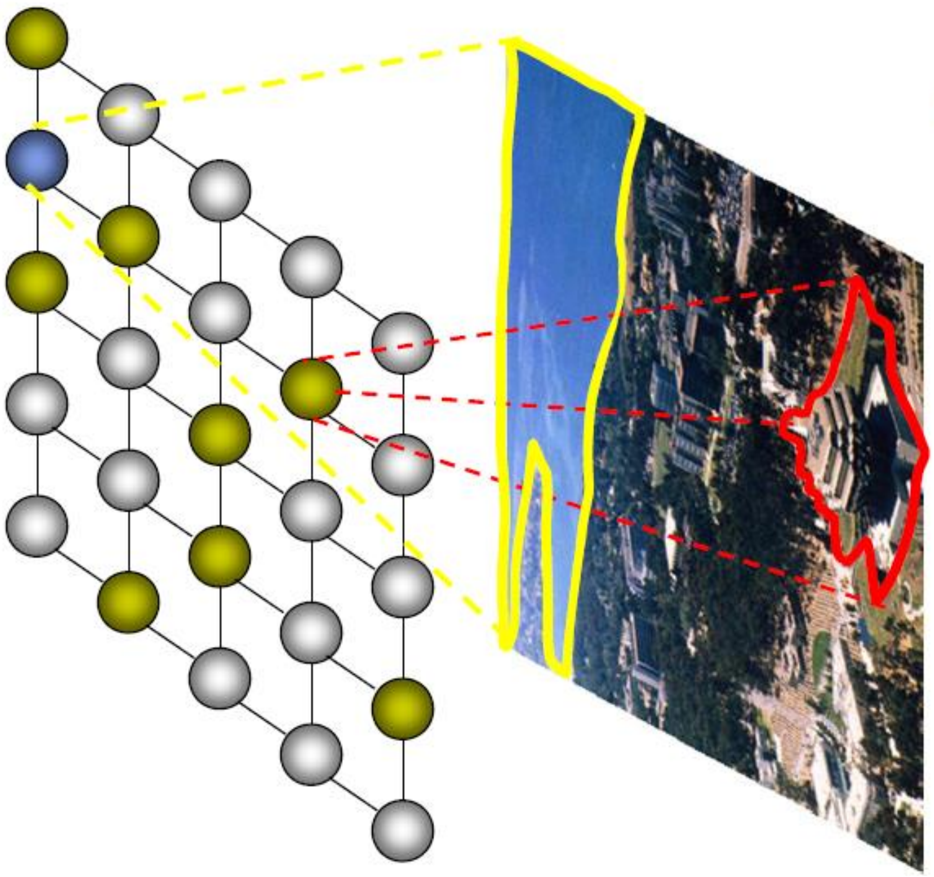
- If the graph $G = (V, E)$ of $\mathbf{Y}$ is a tree, the conditional distribution over the label sequence $\mathbf{Y} = \mathbf{y}$, given $\mathbf{X} = \mathbf{x}$, by the Hammersley Clifford theorem of random fields is:

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



  - 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*

CRFs

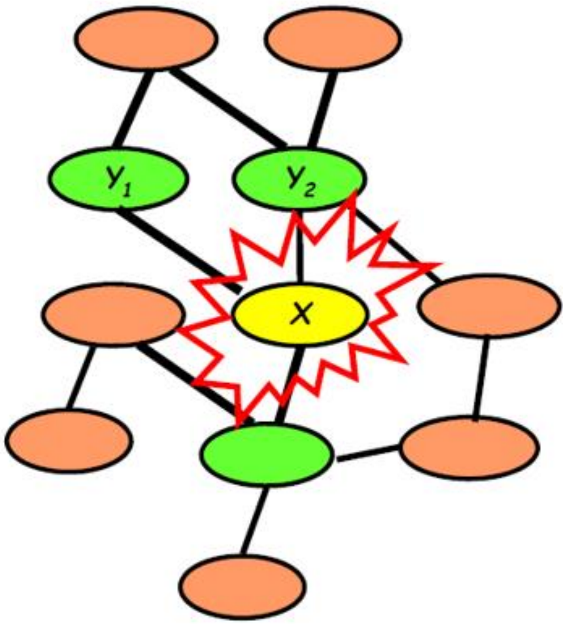


$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

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

Summary: Cond. Indep. Semantics in MRF

- **Structure: an undirected graph**
  - Meaning: a node is conditionally independent of every other node given its directed neighbors
  - Local potential functions 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 variables
  - Local and global independence properties 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"
  - Not only inference but also likelihood-based learning is difficult in general

- Important special cases
  - Ising models; RBMs; CRFs

References

- Lecture notes from "Probabilistic Graphical Models", 10-708, Spring 2015. Eric Xing, CMU

- Daphne Koller and Nir Friedman, Probabilistic Graphical Models: Principles and Techniques