电子科技大学
University of Electronic Science and Technology of China

An Introduction of Probabilistic Graph Model

Chen Huang

Data Mining Lab,
Big Data Research Center, UESTC
Email： huangchen.uestc@gmail.com

Further Ado

数据挖掘实验室
Data Mining Lab

- **Open Course**
  → "*Probabilistic Graphical Models*"
  by Eric Xing

- **Book**
  → "*Probabilistic Graphical Models*", Ch. 1-4
  by Koller and Friedman

Table of Contents

数据挖掘实验室
Data Mining Lab

Part One

What Is Graph Model

How to do data mining

- We get data, then we do mining
- Data Representation → Feature Vector

To describe things from different angles

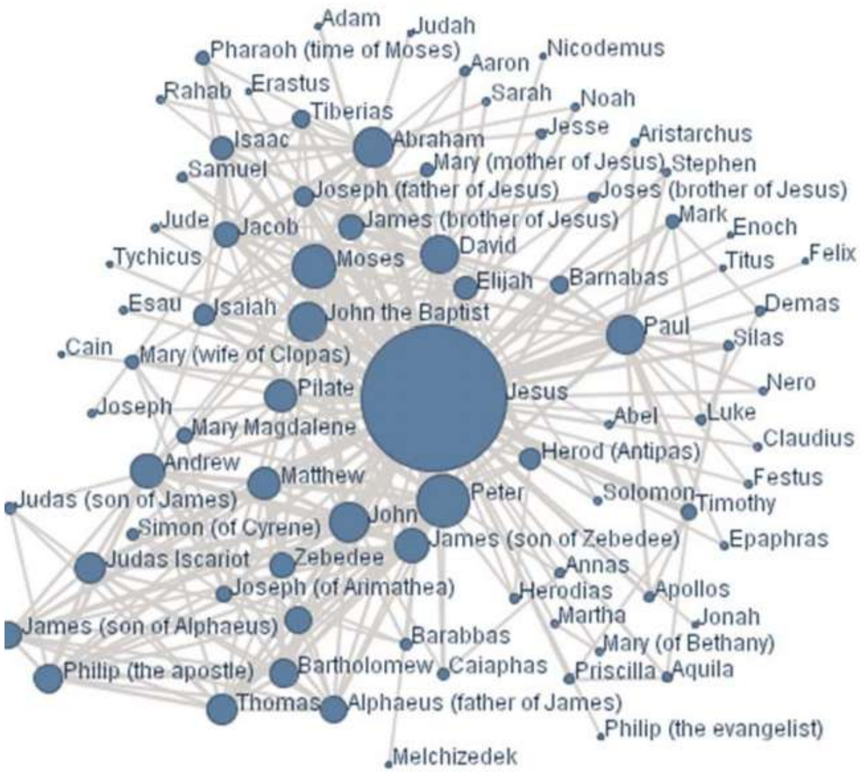
海绵宝宝的形状的傅立叶变换就是派大星的形状

- Data Relationship (Similarity)
  → Data Matrix

How you do data mining

DM LESS IS MORE 数据挖掘实验室 Data Mining Lab

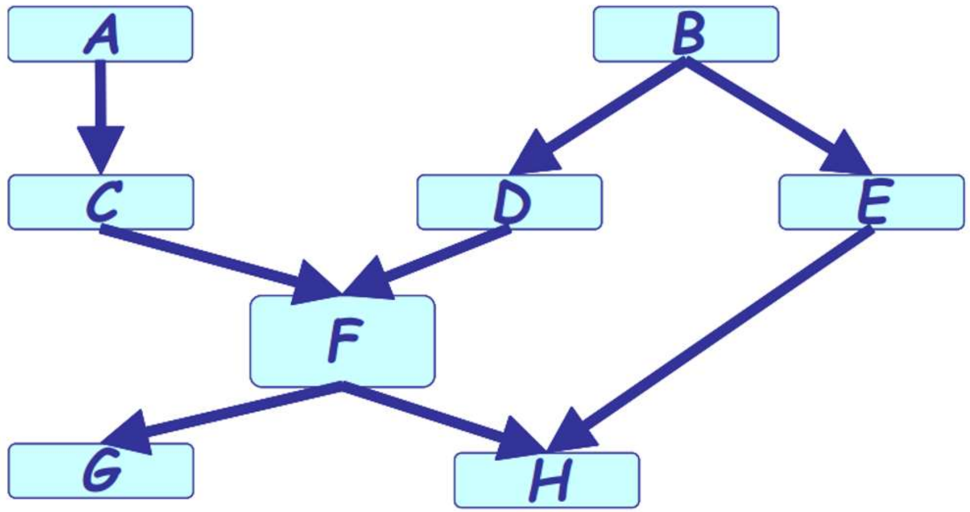
- We apply some kind of algorithm(model) on data matrix, which in some case, can be regarded as a **data graph**.



- The question is that **"Graph-based mining algorithms are the graph model ?".**

Graph Model

- GM refers to a family of **distributions on** a set of **random variables** that are compatible with all the **probabilistic independence** propositions **encoded by a graph** that connects these variables.



- GM = Multivariate statistics + Structure
- GM is a language that used for writing down a fancy model.

--- Eric Xing

Graph Model

Graph

- Reveal the relationships among random variables

Probability

- Reveal the multivariate joint distribution among random variables

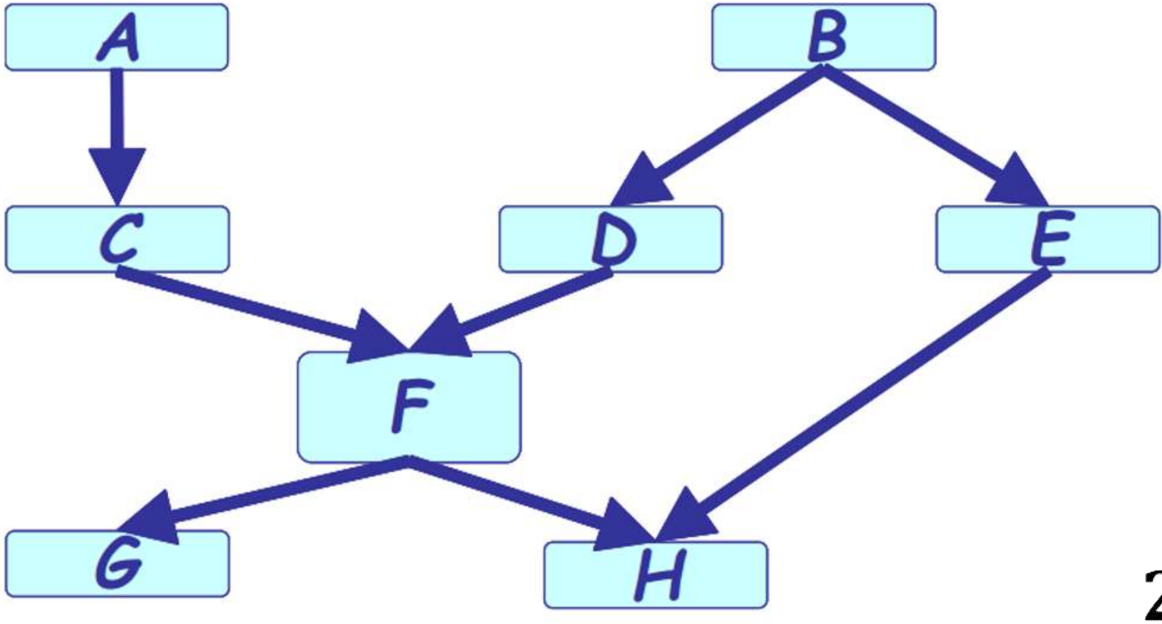
<u>Compact Skeleton</u>

Elimination Ordering

.....

Compact Skeleton

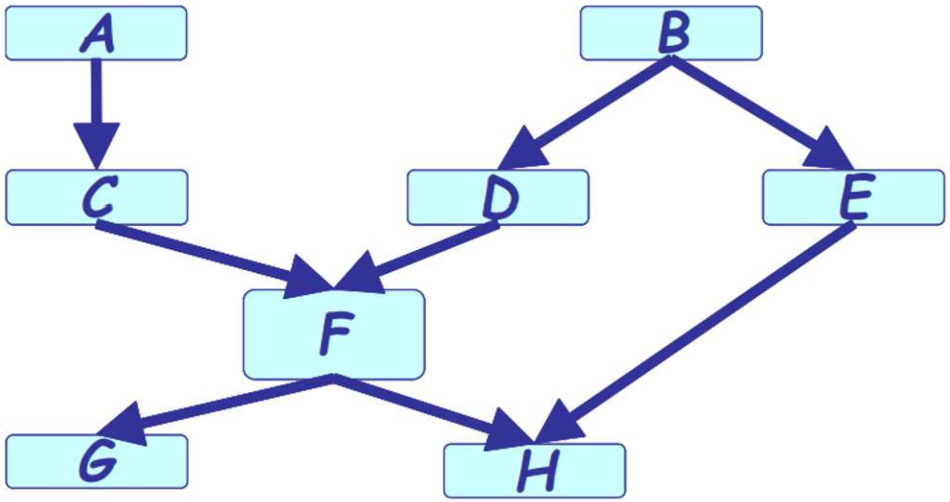
- If I have n = 8 discrete random variables(0/1), how can I write down the full probability distribution?



$$2^8 - 1$$

Compact Skeleton

- If I have n = 8 discrete random variables(0/1), how can I write down the full probability distribution?

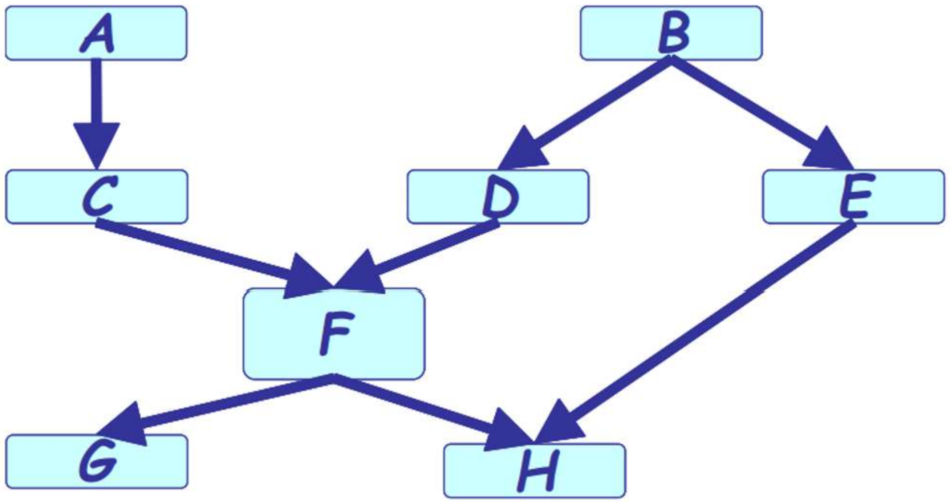


$$P(A\ to\ H) = P(A)P(B|A)P(C|AB)P(D|ABC)P(E|ABCD)$$
$$P(F|ABCDE)P(G|ABCDEF)P(H|ABCDEFG)$$

More natural ! $\mathbf{2^8 - 1}$

Compact Skeleton

- If I have **$n = \infty$** discrete random variables(0/1), how can I write down the full probability distribution now?

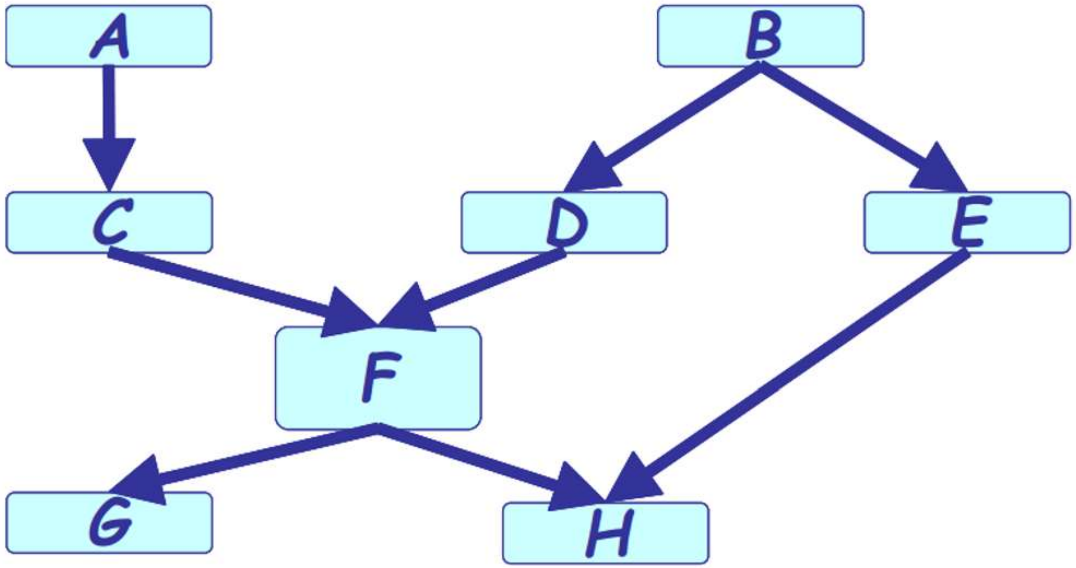


- Count all configurations? → Big Table
  - Lose the insight of the graph
  - Calculation

More natural but not compact !

Why GM

➤ More intuitively in **a factorized** way, we have



20

$$P(A\ to\ H) = P(A)P(B)P(C|A)P(D|B)P(E|B)$$
$$P(F|CD)P(G|F)P(H|EF)$$

Compromise..

- What we gain: Calculation or Cost saving
- What we loss: Variables relation may not be independent

More Information on GM

- **Two types of GM**
  - Bayesian Network (Directed)
  - Markov Random Field (Undirected)

- **What can GM do**
  - **Representation**

    Capture uncertainties
  - **Inference**

    Probability of A under the observation of B
  - **Learning**

    Find "right" model for my data

More Information on GM

➢ What can GM do

- **Representation**

  Capture uncertainties

- **Inference**

  Message-passing (sum-product, belief propagation)

  The junction tree algorithms

  MCMC

  Variational algorithms

  ….

- **Learning**

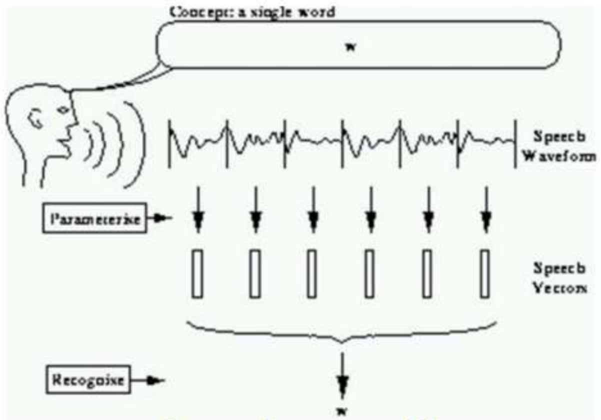
  Chow-Liu Algorithm

  …

More Applications on GM

➢ If you still remember…

- *LDA* (Topic Model) --- NLP
- *Hidden Markov Model* by $Ming^2$
- *Bayes Network*
- *Dirichlet Process* by $Yu\ Bo^2$
- *Gaussian Process* by Pro. Xu



Speech recognition

Computer vision

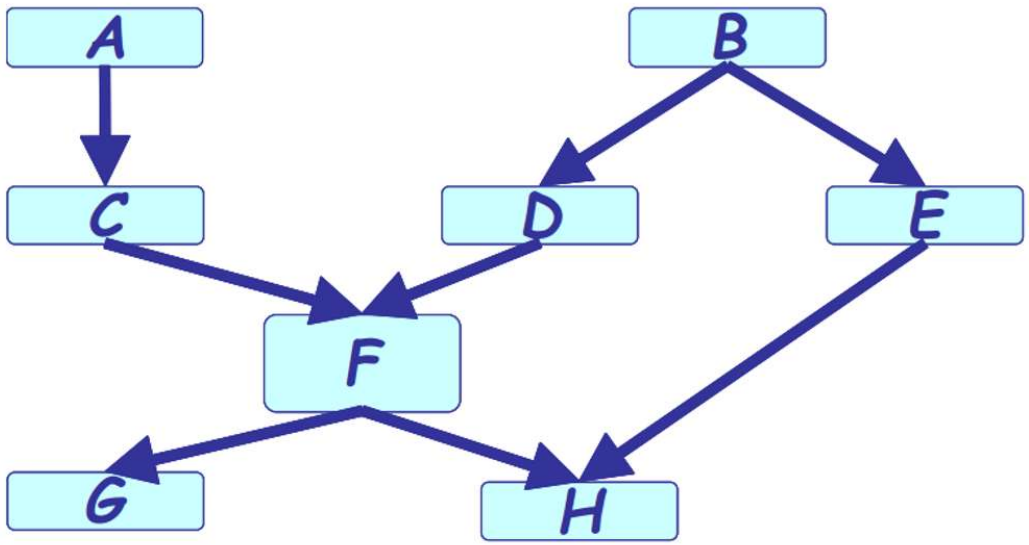
Part Two

Probability Representation

*Warning: The following contains a lot of terms and concepts

Directed Acyclic Graph (DAG)

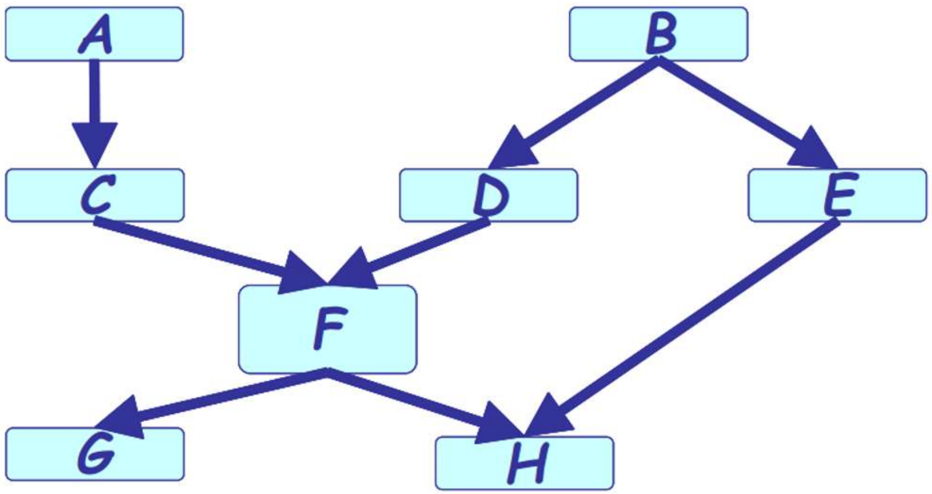
- Represents a **probability distribution** through a DAG that encode **conditional dependency** and **independency relationships** among variables in the model



Factorization Theorem of DAG

- Factors according to "**node given its parents**".
- Use the independencies from graph **G** to represent the probability distribution **P**

$$P(X) = \prod_{i=1:d} P(X_i | Parents(X_i))$$



Why this work?

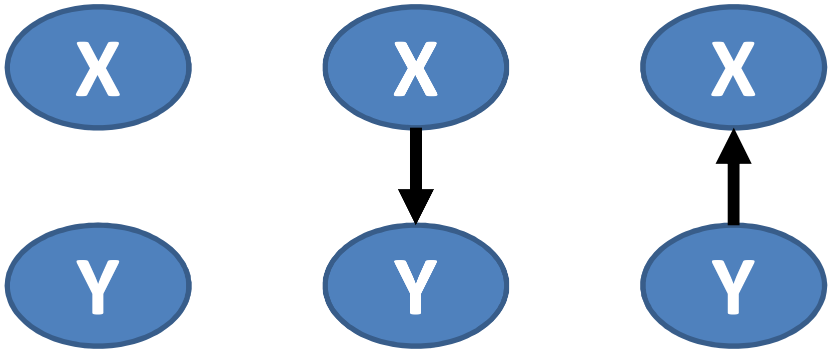
$$P(A\ to\ H) = P(A)P(B)P(C|A)P(D|B)P(E|B)$$
$$P(F|CD)P(G|F)P(H|EF)$$

I-Map

- The link between **P** and **G** on independence assertions
- Definition:
  - I(P) = all independence assertions in form of $(X \perp Y | Z)$ on P
  - I(G) = all independence assertions on G
  - If $I(G) \subseteq I(P)$, then G is the I-Map of P
- If I(G1) = I(G2), then graph G1 and G2 are I-equivalent

I-Map

- Some independency in P may not be in the I-Map



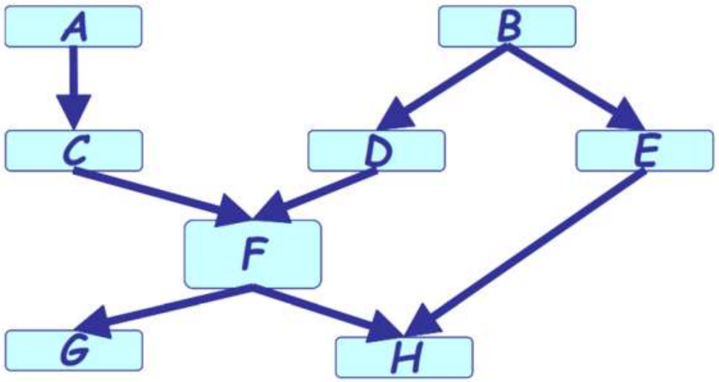
$I(G) = \{X \perp Y\}$ $\quad$ $I(G) = \emptyset$ $\quad$ $I(G) = \emptyset$

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

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

I-Map & Factorization

- G is the I-Map of P, P can be factorized according to G
- P can be factorized according to G, G is the I-Map of P



Target: $P(H|EF) = P(H|ABCDEFG)$

$$P(H|ABCDEFG) = \frac{P(A\ to\ H)}{P(A\ to\ G)} = \frac{P(A\ to\ H)}{\sum_H P(A\ to\ H)}$$

$$= \frac{P(A)P(B)P(C|A)P(D|B)P(E|B)\,P(F|CD)P(G|F)P(H|EF)}{\sum_H P(A)P(B)P(C|A)P(D|B)P(E|B)\,P(F|CD)P(G|F)P(H|EF)}$$

$$= \frac{P(A)P(B)P(C|A)P(D|B)P(E|B)\,P(F|CD)P(G|F)P(H|EF)}{P(A)P(B)P(C|A)P(D|B)P(E|B)\,P(F|CD)P(G|F)\sum_H P(H|EF)}$$

$$= P(H|EF)$$

What we know so far?

Why factorization work

I-Map == Factorization

G is the I-Map of H → $I(G) \subseteq I(P)$,

I(G) don't have to imply every independence in I(P)

Is there any independency in I(G) that are not in I(P)?

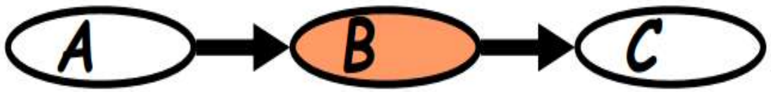
What is in I(G)

What is in I(G)?

➢Local Markov independence

- Given the parents of X_i, X_i is independent with the non-descendants of X_i

$$X_i \perp NonDescendants(X_i) | Parents(X_i)$$

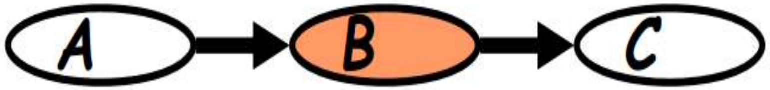


➢*Global Markov Independence

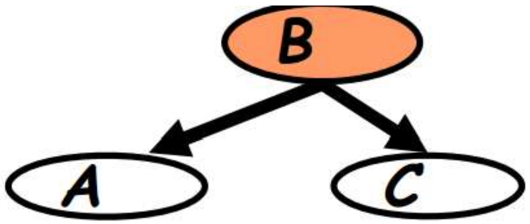
- **D-separation**
- It reveals a concept of Separation among the random variables in graph <u>from a global view</u>

Local Structures on Graph

- **Cascade** $(A \perp C \mid B)$

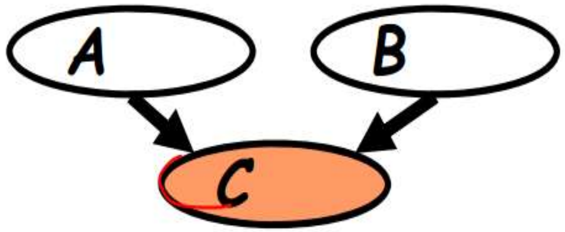


- **Common Parent** $(A \perp C \mid B)$



- **V-Structure**

  If C has two causes A & B observation of one of them would "explain away" the other(less likely to be observed)



Active Trails

数据挖掘实验室
Data Mining Lab

- Causal Trail
  - X → Z → Y
- Evidential Trail
  - X ← Z ← Y
- Common Cause
  - X ← Z → Y

The trail is active if and only if Z is not observed

- Common Effect
  - X → Z ← Y

The trail is active if and only if Z is observed or one of Z's descendants is observed

D-separation

- 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(X;Y|Z)**, if there is no **active trail** between any node $x \in X$ and $y \in Y$ given Z

- Define I(G) to be all the independence properties that corresponds to D-separation

$$\mathrm{I(G)} = \{\mathrm{X} \perp \mathrm{Y} | \mathrm{Z} : \mathrm{d} - \mathrm{sep}(\mathrm{X}; \mathrm{Y} | \mathrm{Z})\}$$

*What D-separation do?

➢ Soundness(可靠性)

- Given Z, if x and y are d-separated, $(x \perp y) | z$
- P implies any independences in D-separation (proof passed)

➢ Completeness(完备性)

- **D-separation contains all independence assertions**
- If X and Y given Z are not D-separated in G, then X and Y are dependent in some(not all) distribution P that factorizes over G

Thus, Factorization works

DAG Summary

➢ Three equivalences

- P factorizes over G

$\downarrow$ I-map

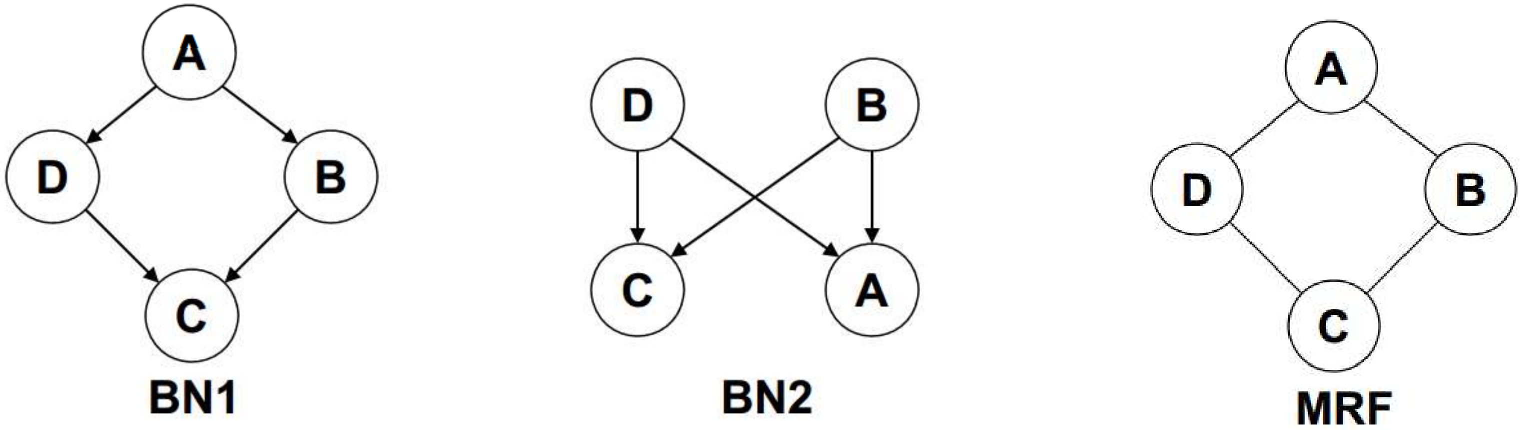
- P satisfies the local independence of G

- P satisfies the global independence(D-separation) of G

DAG Test

➢ Example

- Suppose we have a model
- where $(A \perp C) \mid \{B,D\}$ and $(B \perp D) \mid \{A,C\}$
- Can you write down a DAG to represent it for me?



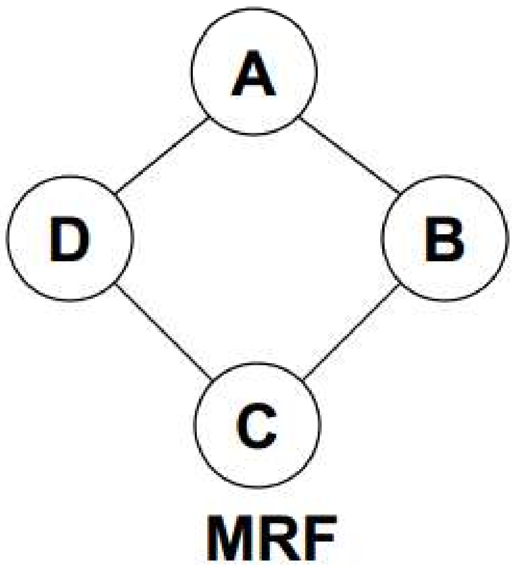
For some distribution, we can not use DAG to represent!

Markov Random Field (UGM)

数据挖掘实验室
Data Mining Lab

➢Perceptual knowledge

- An <u>undirected graphical model</u> that explicitly expresses the relationships between nodes in a undirected way.
- Thus independence definition of UGM is different from DAG



Markov Property on UGM

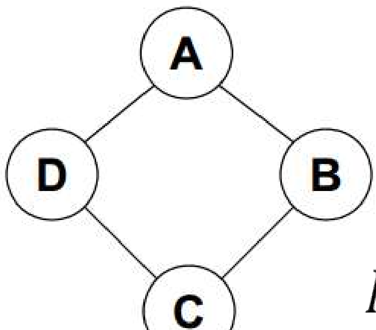
➢ Local Markov Independence

- The Local Markov independencies associated with H is

$$\mathrm{I}_{\iota}(H):\{X_i \perp V - \{X_i\} - \underline{\boldsymbol{MarkovBlanket}}(X_i)\}$$

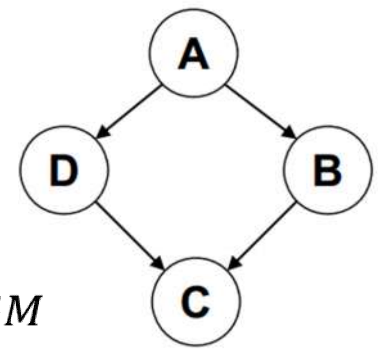
➢ *Markov Blanket*

- UGM={all neighbors of X}
- *DAG={Parents(X), Children(X),Parents(Children(X))}



MRF

$$MB(D) = \{A, C\}_{UGM} \\ = \{A, C, B\}_{DAG}$$



BN1

Markov Property on UGM

数据挖掘实验室
Data Mining Lab

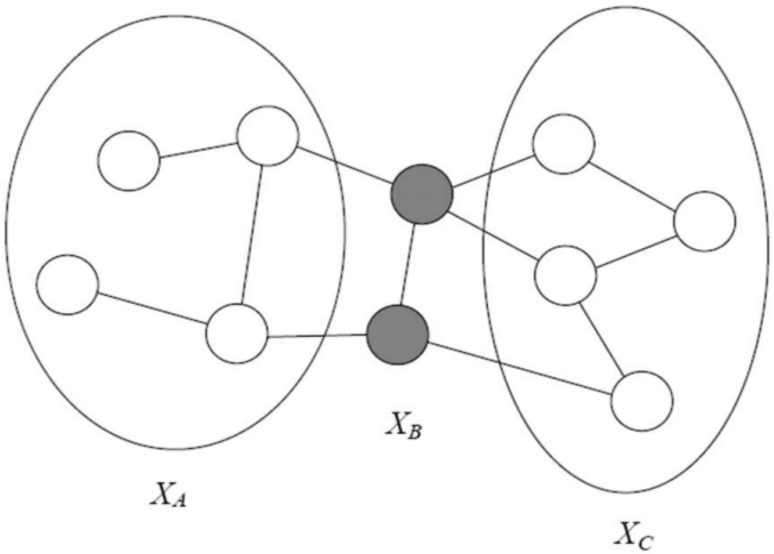
➢Global Markov Independence

- **B separate A and C** if every path from A to C pass B, namely

$$Sep(A; C|B)$$

- For any set A, B and C, such that B separates A and C, A is independent of C given B:

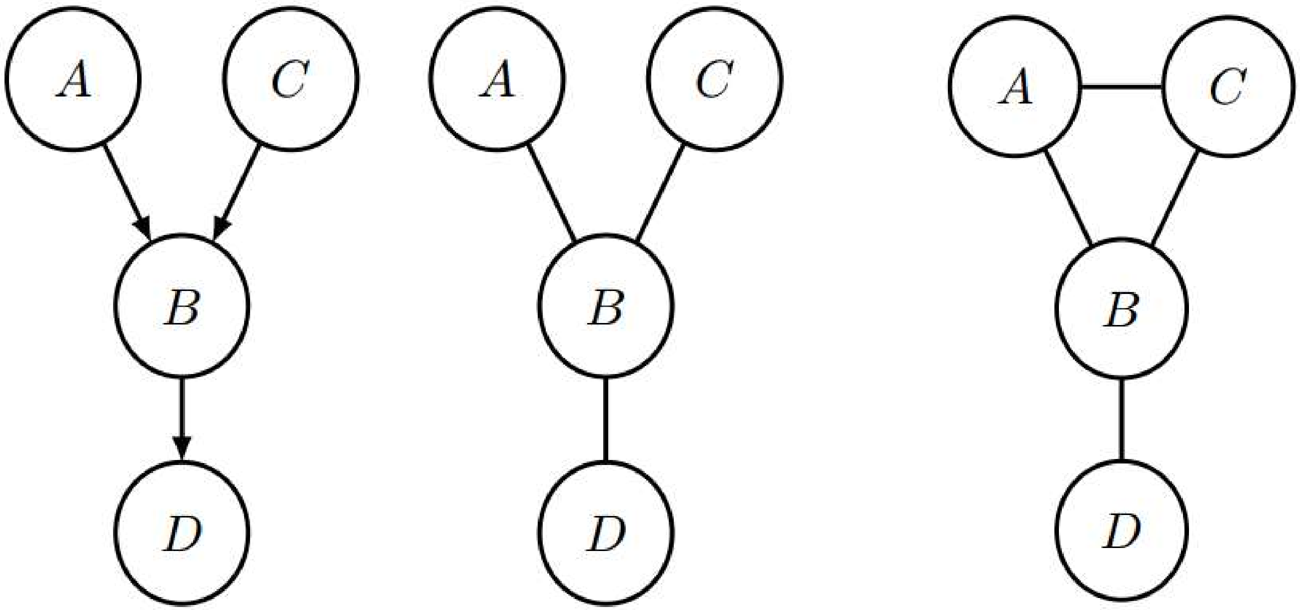
$$I(H) = \{A \perp C|B: Sep(A; C|B)\}$$



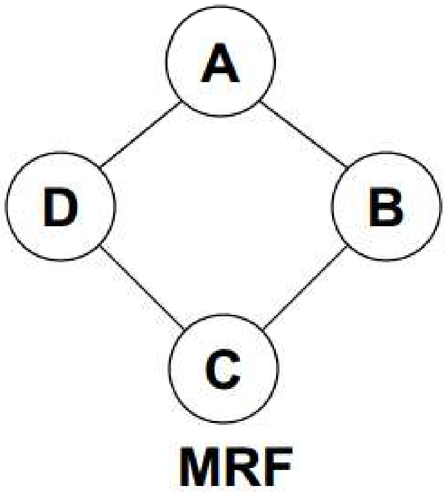
Markov Random Field (UGM)

- DAG is used for causality
- UGM blurs the causality and is used for characterizing mutual relationships
- How to convert DAG to UGM ?

→ Moralization → Graph Elimination, Junction Tree Inference



How to write the joint probability of the random variables in UGM?



$$P(A, B, C, D) = ?$$

Undirected Graph Model

数据挖掘实验室
Data Mining Lab

➢ Definition

- An UGM represents a distribution P defined by an undirected **graph H** and a set of positive-valued **potential function ψ_C** associated with the **clique** of H, s.t

$$\mathbf{P}(X_1, \dots, X_n) = \frac{1}{Z}\prod \psi_C(X_c)$$

$$Z = \sum \prod \psi_C(X_c)$$

Too many "New words"

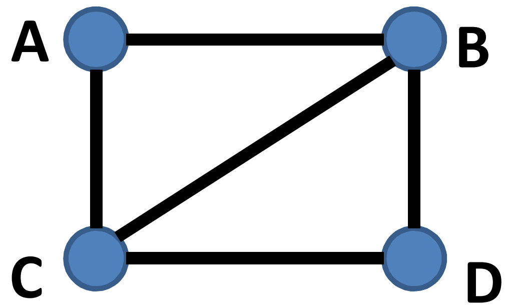
Clique

➢Definition

- A set of nodes is a clique, if any nodes in that set are connected with an edge

➢Max Clique

- Clique is a max clique, if we cannot add another node to make it a bigger clique



Max clique= ABC, BCD

Potential Function

➢ Info

- Positive-valued function (Why?)
- Represents the coupling strength of the clique, which indicates how much the nodes within that clique covary
- In most cases, the Exponential Function

$$\varphi_c = \exp(-f(c))$$

- Where f(c) is called **Energy Function** with a higher energy configuration having lower probability

*Why it work?

- **Gibbs Distribution**

$$Q\{X\} = \prod_{A} V_A(X)$$

*Why it work?

数据挖掘实验室
Data Mining Lab

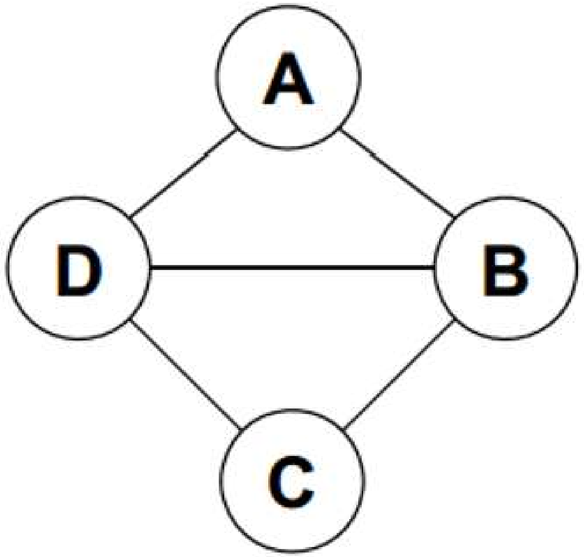
➢ The *Hammersley-Clifford theorem* proves that a MRF and Gibbs Field are equivalent with regard to the same graph

- Given any MRF, all joint probability distributions that satisfy the conditional independencies can be written as clique potentials over the maximal cliques of the corresponding Gibbs Field

- Given any Gibbs Field, all of its joint probability distributions satisfy the conditional independence relationships specified by the corresponding MRF

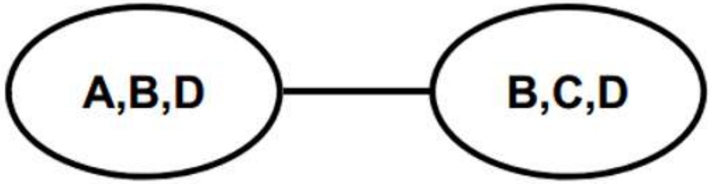
Calculation Of Clique

- Given a UGM, How to choose the size of clique to calculate our factorization?



Calculation Of Clique

- Using max-clique

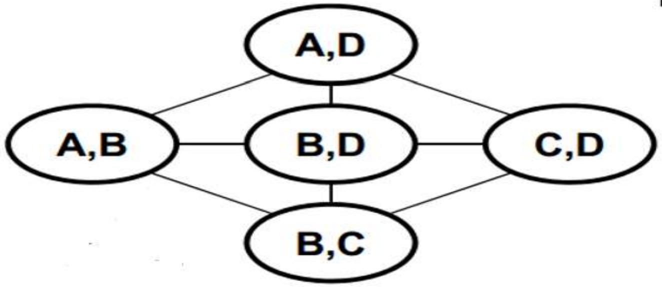


$\psi_c(\mathbf{x}_{124})$ $\psi_c(\mathbf{x}_{234})$

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

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

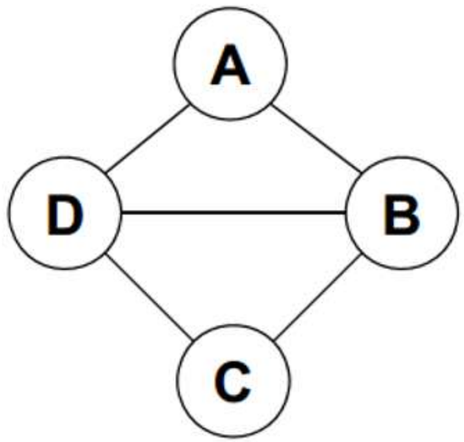
- Using sub-clique



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

- Canonical Representation(Using all terms)

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



Why Concern?

- Using max-clique
  - Loss more local structure information
  - The space of values of max-clique is larger
  - Represent graph with less terms

- Using sub-clique
  - Partition functions are much easier to compute

- Using all terms
  - …

These three clique configurations are equivalent?

Interpretation of Clique Potentials

- Independence statement implies (by definition) that the joint below must factorize as

$$\begin{aligned} \mathrm{P}(X,Y,Z) &= P(Y)P(X|Y)P(Z|Y) \\ &= P(X,Y)P(Z|Y) \\ &= P(X|Y)P(Z,Y) \\ &= \varphi(x,y)\varphi(y,z) \end{aligned}$$

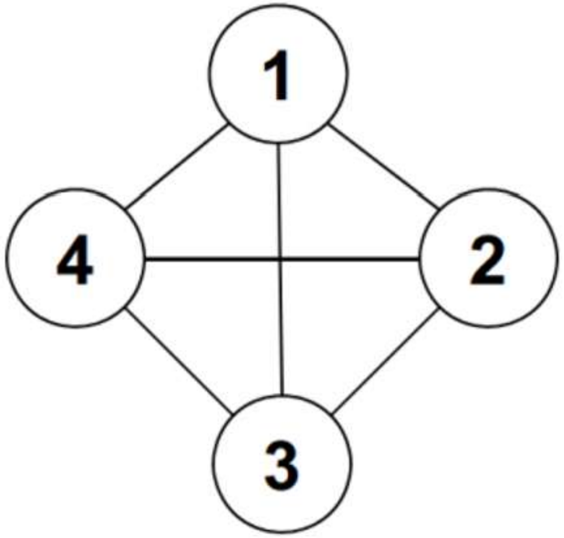
- Potential function on some clique can't all be marginal or conditionals

Part Three

Example Models Quick View

Boltzmann Machine

- A fully connected graph with pairwise potentials on binary-valued nodes, the energy function is expressed in sub-clique form



$$P(X_1, X_2, X_3, X_4) = \frac{\exp(\sum_{i,j} \varphi(X_i X_j))}{Z}$$

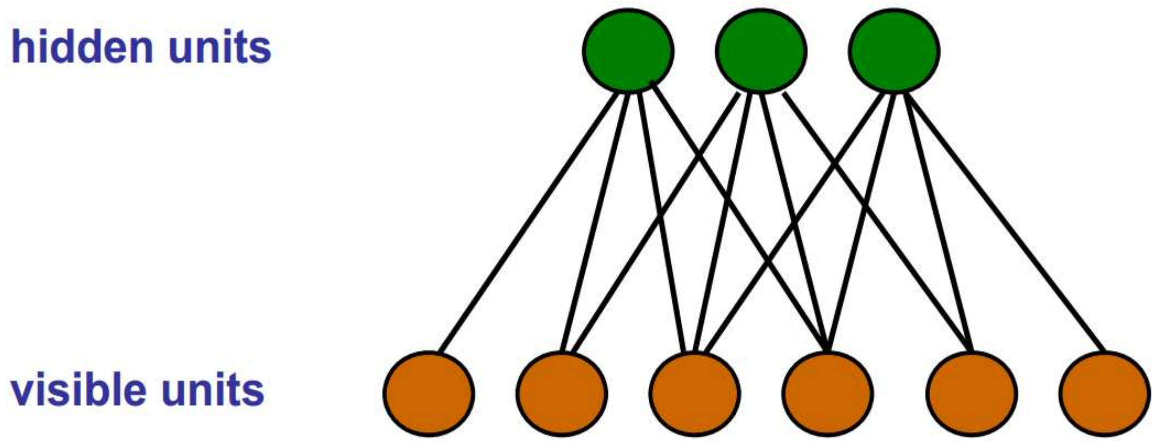
$$= \frac{\exp(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \alpha_i X_i + C)}{Z}$$

Figure 4: An example Boltzmann machine.

Restricted Boltzmann Machine

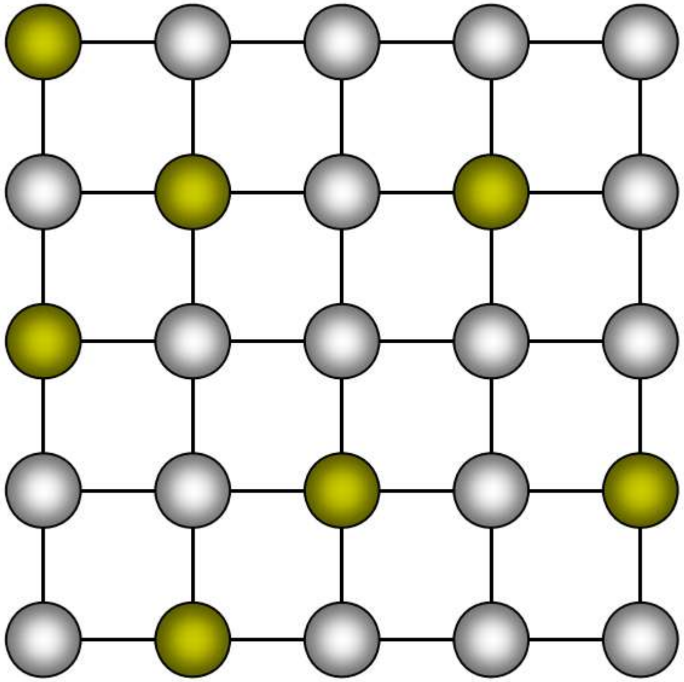
- Consists of many layers, each layers has two sub-layers: one for hidden units h_i and one for visible units x_i, the probability function for RBM is:

$$P(X,H|\theta) \propto \exp(\sum_i \theta_i \varphi(x_i) + \sum_i \theta_i \varphi(h_i) + \sum_{i,j} \theta_{ij} \varphi(x_i, h_j) + A(\theta))$$



Ising Model

- 2D-Ising Model won 1968 Nobel Prize in Chemistry



$$P(X) = \frac{\exp(\sum_{i,j} \varphi(x_i x_j))}{Z}$$

$$= \frac{\exp(\sum_{ij} \theta_{ij} x_i x_j + \sum_i \alpha_i x_i)}{Z}$$

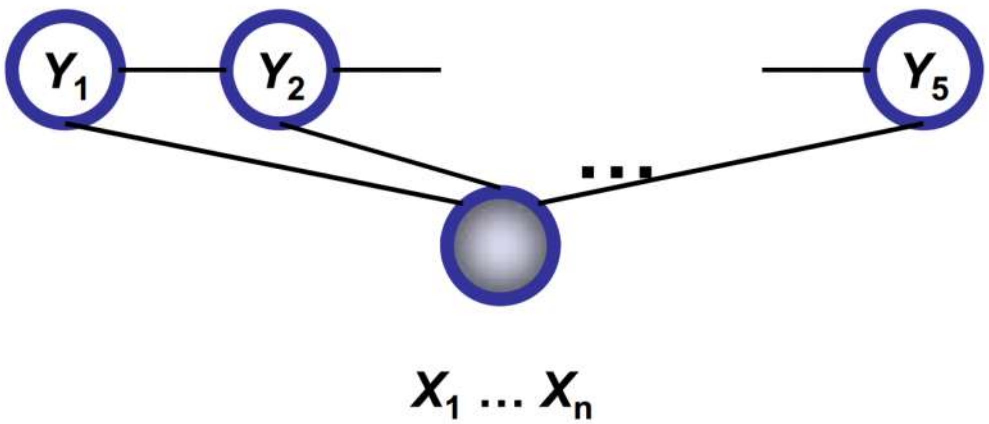
Conditional Random Field

- A **discriminate** UGM models the conditional probability of a label sequence y (hidden) given an observation sequence x.

$$P(Y|X) \propto \exp\Big(\sum_{i,k} \lambda_k\, t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l\, s_l(y_i, x, i)\Big)$$

Transition feature function on edge

State feature function on node



Take Home Message

- GM = Multivariate statistics + Structure
- Graph structure helps to represent a probability distribution in a compact factorized way
- I-map, D-separation
- Clique, Potential Function
- Local Markov & Global Markov
  - → Equivalence (positive distribution P on UGM)
- Factorization
  - Node given its parents
  - Clique
- Fancy Models

Thanks

By HC