Graphical Model

Jiafeng Guo
guojiafeng@ict.ac.cn

Traditional Machine Learning

Markov random field

K-means clustering

RVM

Gaussian mixture

logistic regression

Kalman filter

random forest

HMM

principal components

neural networks

deep networks

support vector machines

kernel PCA

ICA

linear regression

Boltzmann machines

Radial basis functions

Gaussian process

decision trees

factor analysis

Model-based Machine Learning

Traditional:

- “how do I map my problem into standard tools”?

Model-based:

- “what is the model that represents my problem”?

Goal:

A *single* development framework which supports the creation of a wide range of bespoke models

The Fundamental Questions

- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

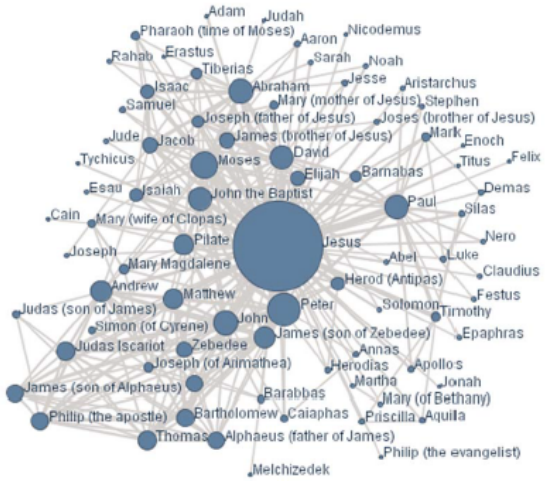
    $e.g.\ P(X_i)$
- Inference
  - How do I answers questions/queries according to my model and/or based given data?

    $e.g.\ P(X_i|D)$
- Learning
  - What model is "right“ for my data?

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

Probabilistic Graphical Models

Graph



Model

$$\mathcal{M}$$

PGM = Probability + Structure

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

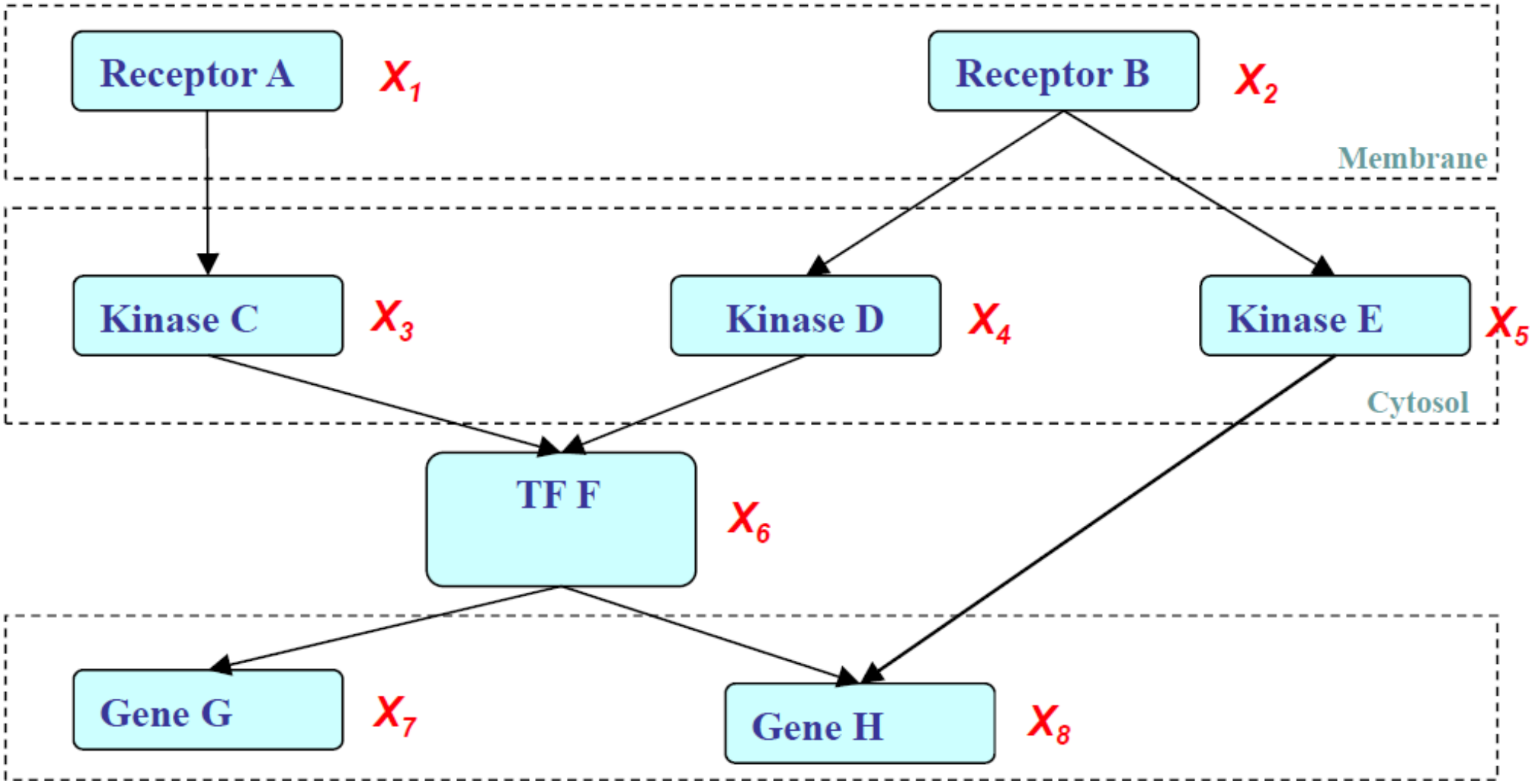
Data

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

Graphical representations of probabilistic models

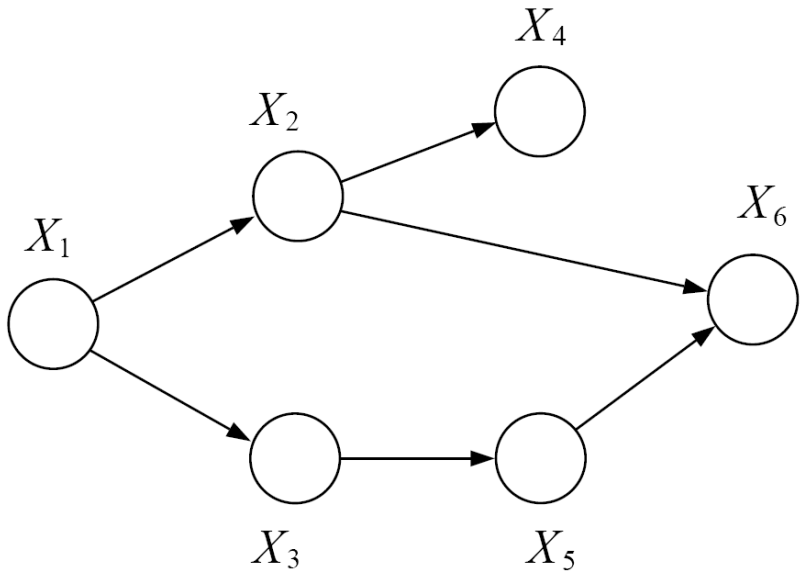
Structure Simplifies Representation

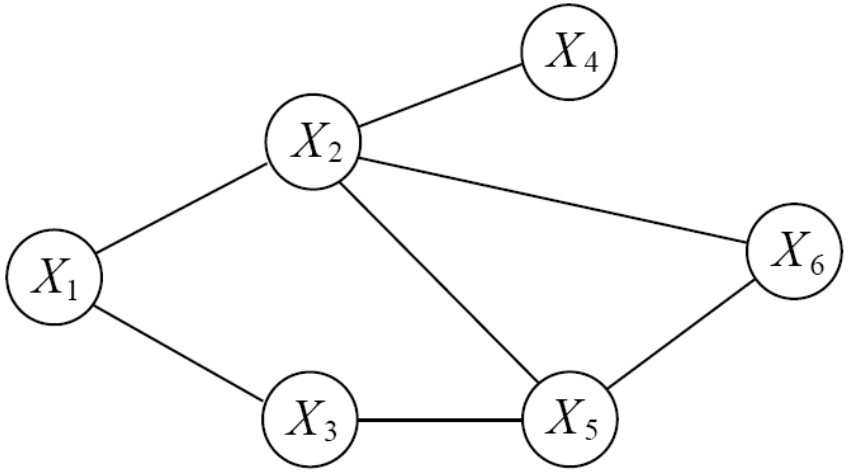
Dependencies among variables



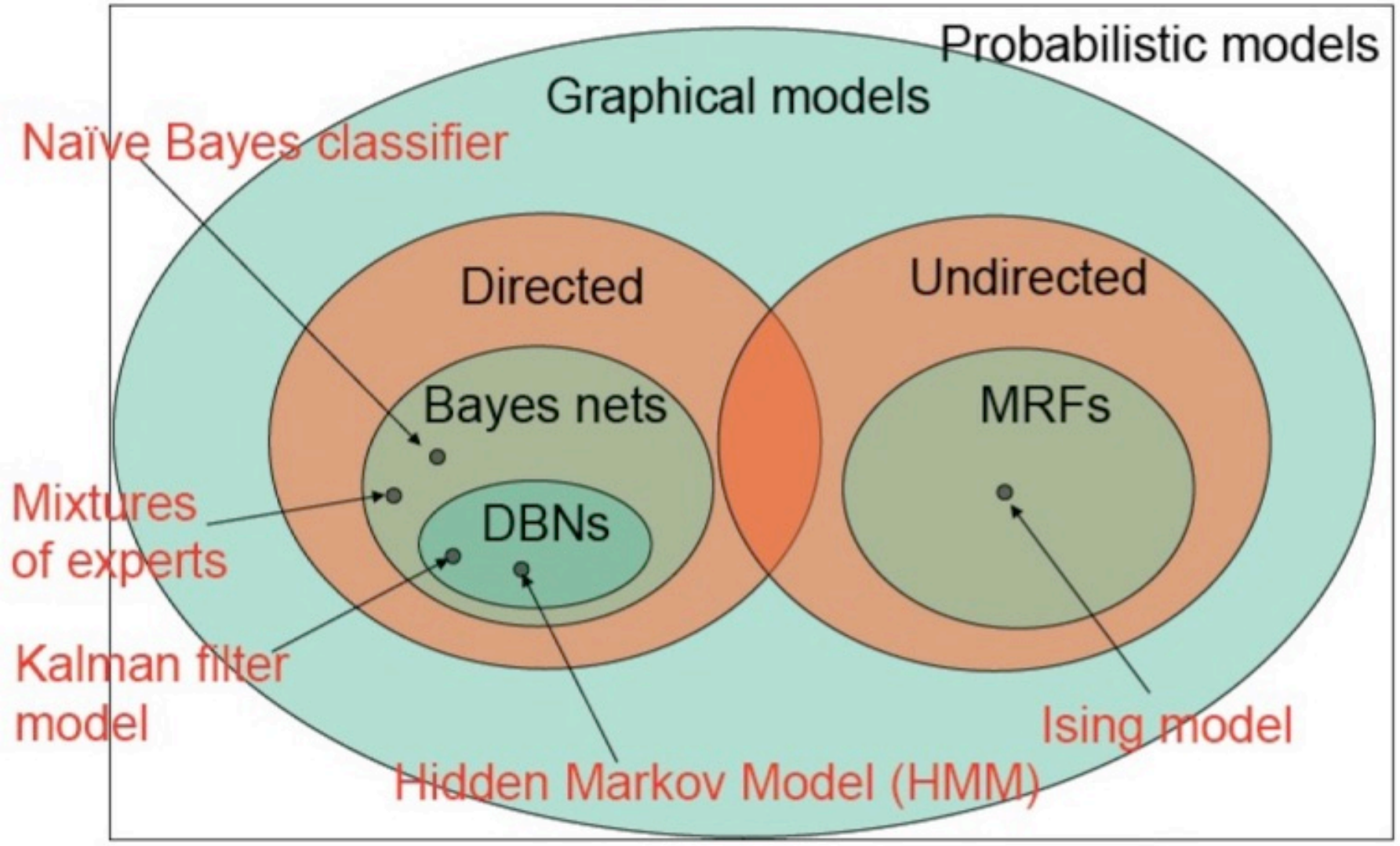
PGM

- Nodes denote random variables/states， Edges denote probabilistic relationship
- Types
  - Directed PGM or Bayesian networks: causality
  - Undirected PGM or Markov random fields: correlation



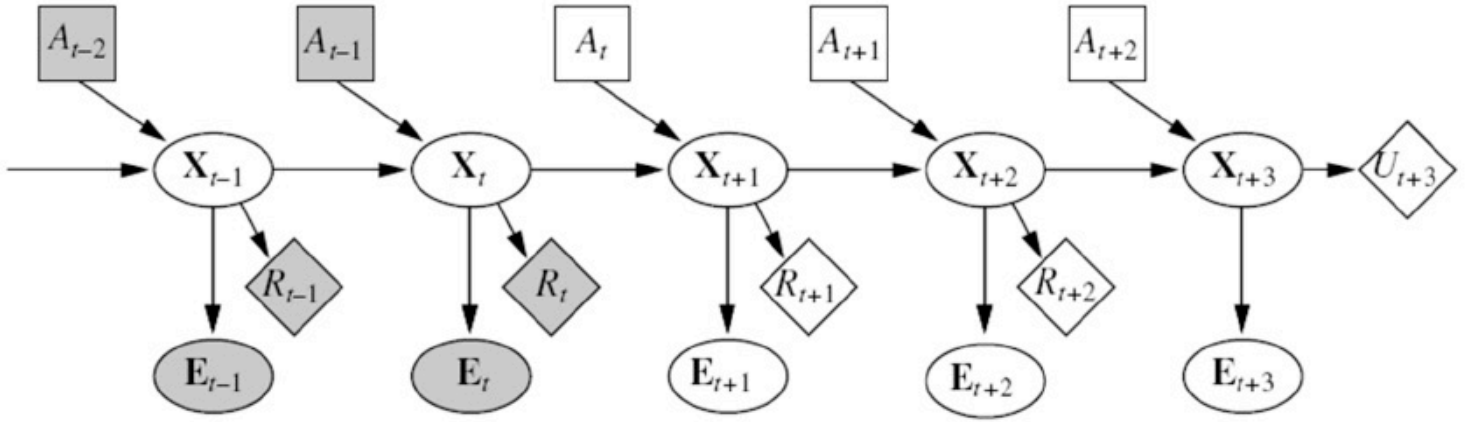


The familiar of PGM

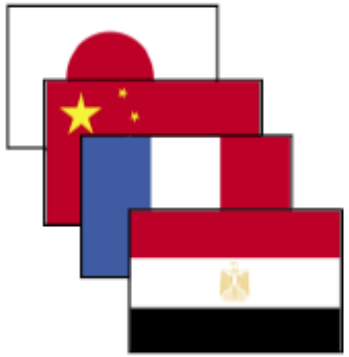


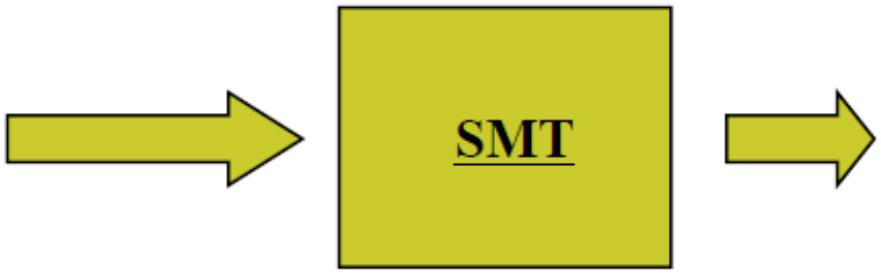
Fancy PGMs: reinforcement learning

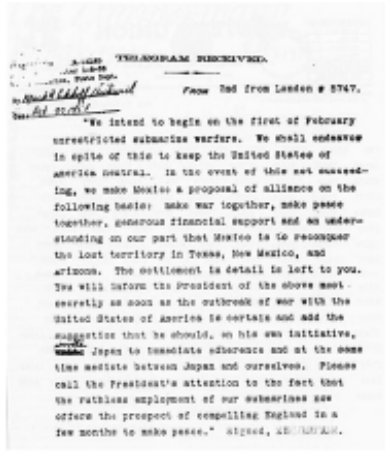
- Partially observed Markov decision processes (POMDP)



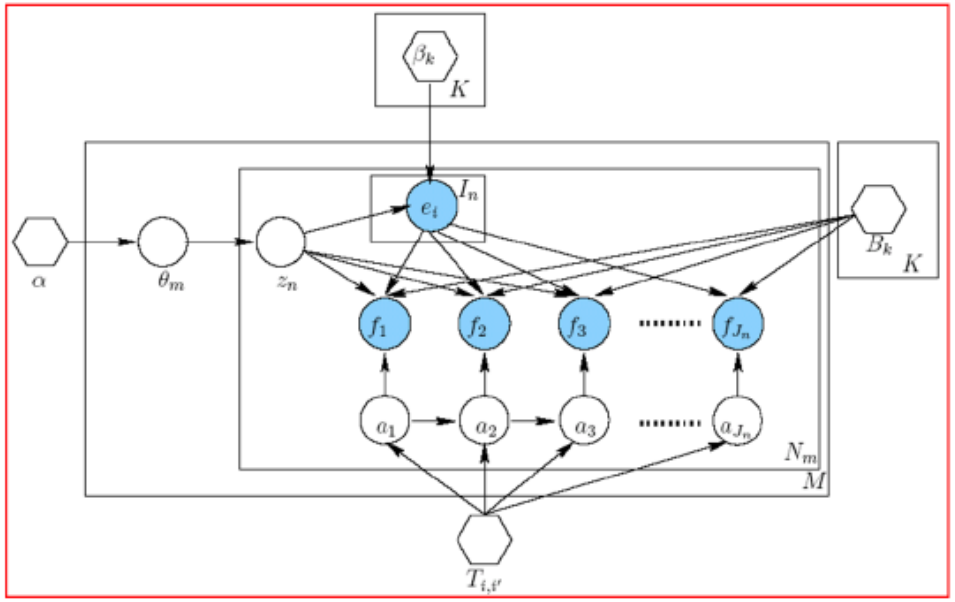
Fancy PGMs: machine translation





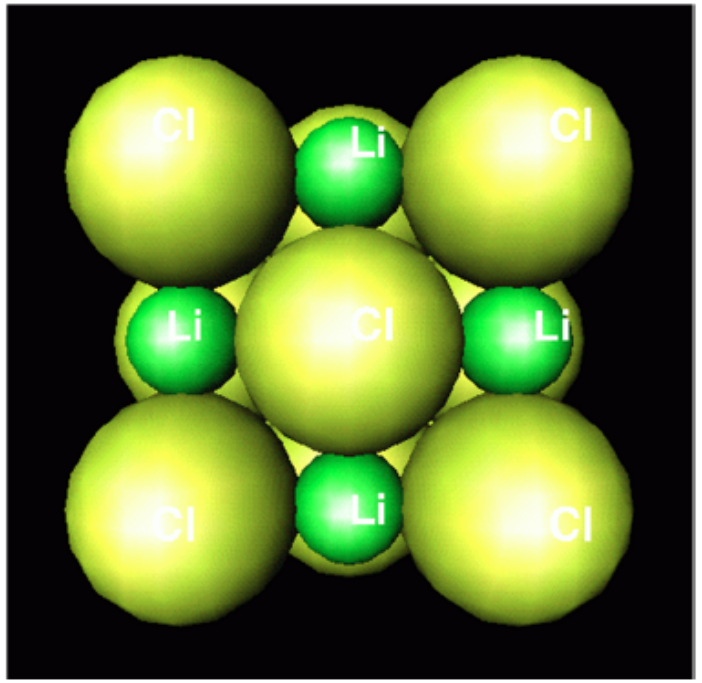


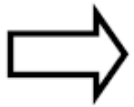
"We intend to begin on the first of February unrestricted submarine warfare. We shall endeavor in spite of this to keep the United States of America neutral. In the event of this not succeeding, we make Mexico a proposal of alliance on the following basis: make war together, make peace together, generous financial support and an understanding on our part that Mexico is to reconquer the lost territory in Texas, New Mexico, and Arizona. The settlement in detail is left to you. You will inform the President of the above most secretly as soon as the outbreak of war with the United States of America is certain and add the suggestion that he should, on his own initiative, invite Japan to immediate adherence and at the same time mediate between Japan and ourselves. Please call the President's attention to the fact that the ruthless employment of our submarines now offers the prospect of compelling England in a few months to make peace." signed,

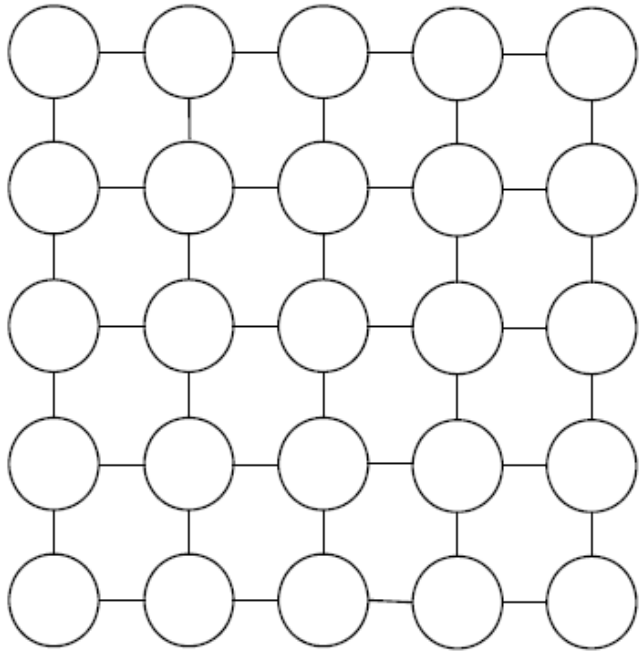


The HM-BiTAM model (B. Zhao and E.P Xing, ACL 2006)

Fancy PGMs: solid state physics





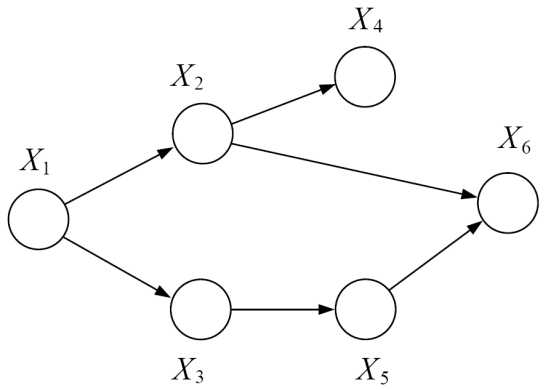


Ising/Potts model

Directed PGM

Directed PGM

- **Definition:** A directed graph $G = (V, E)$ has a finite set of vertices(nodes) V and a set of edges E that consists of ordered pair of vertices

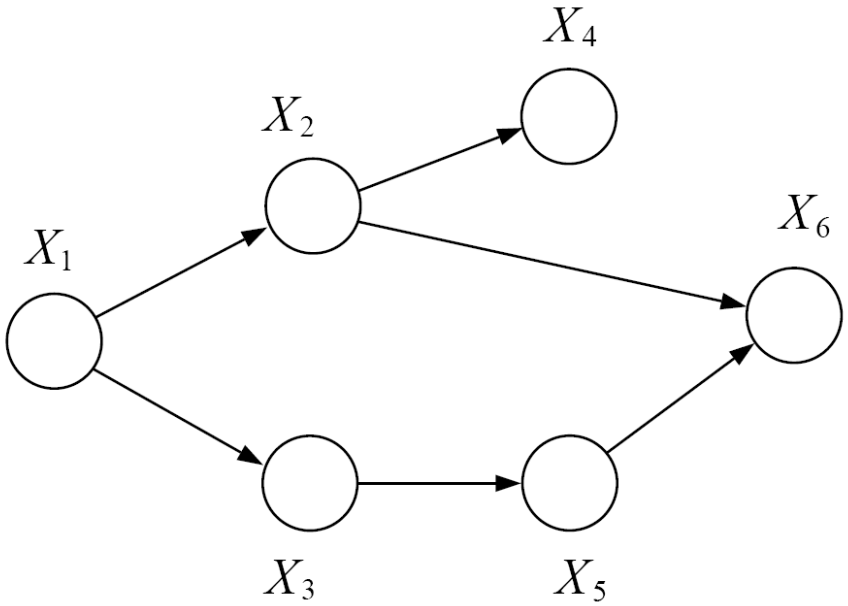


- Directed graphs can express causal relationships
- Often we observe child variables and wish to infer the posterior distribution of parent variables
- Example:

Examples of Directed Graphs

- Hidden Markov models
- Kalman filters
- Factor analysis
- Probabilistic principal component analysis
- Independent component analysis
- Mixtures of Gaussians
- Transformed component analysis
- Probabilistic expert systems
- Sigmoid belief networks
- Hierarchical mixtures of experts
- Etc, etc,…

Directed PGM (BN)



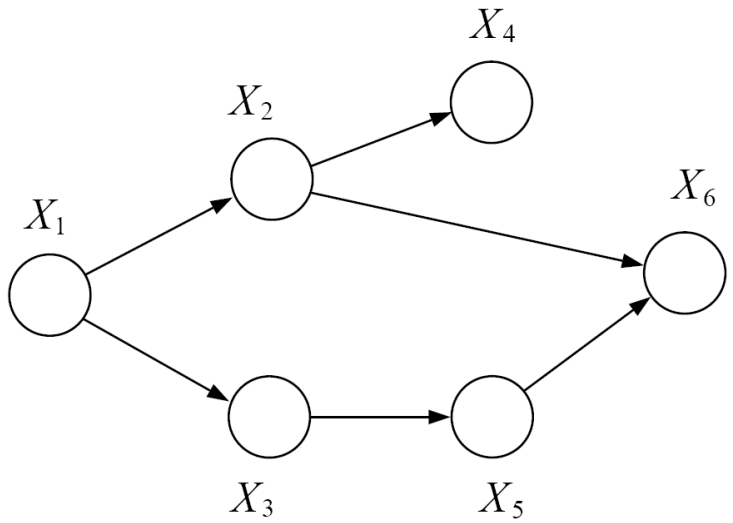
Probability Distribution → Queries

Representation → Implementation

Conditional Independence → Interpretation

Probability Distribution

There is a family of probability distribution that can be represented with this graph



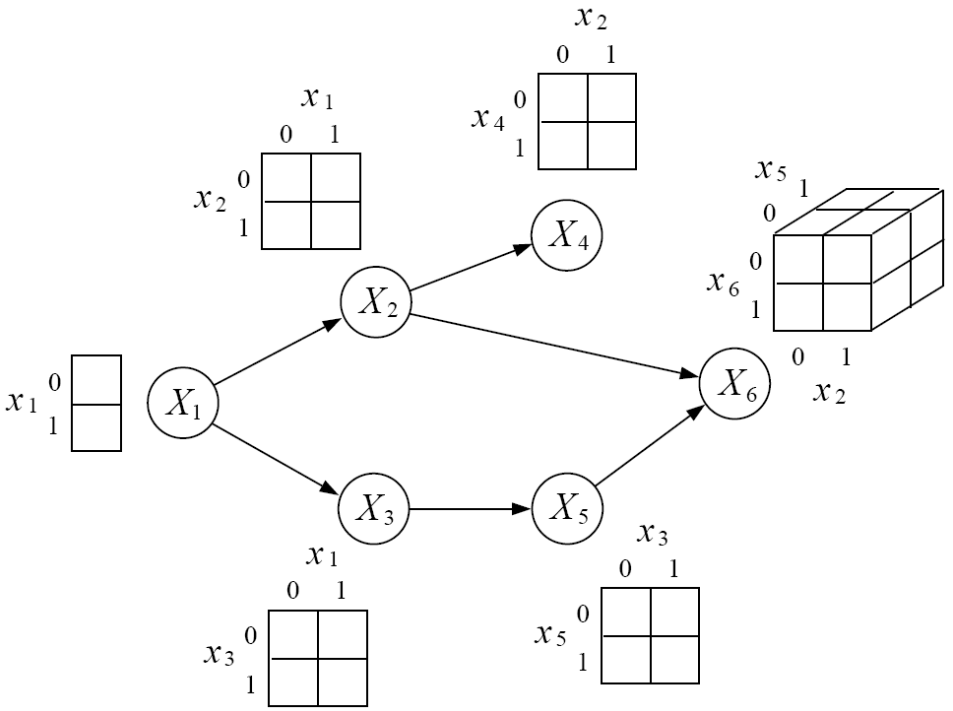
Each node is associated with one conditional probability distribution

Joint Probability Distribution:

$$P(x_1, x_2, \dots x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi i})$$

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2)P(x_5|x_3)P(x_6|x_2, x_5)$$

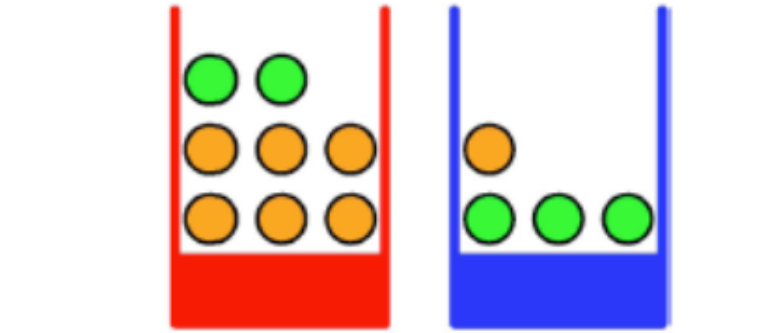
Representation

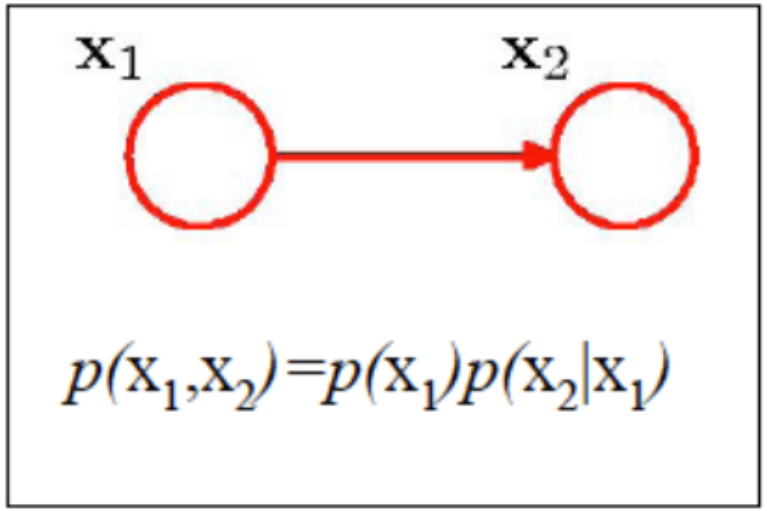


$O(2^n) \longrightarrow O(n*2^k)$

Bayesian networks represent joint probability distributions more economically, using a set of “local” relationships among variables.

Boxes of Fruit Example





- x_1= Box: red/blue
- x_2= Fruit: a/o

$p(x_1)$

$p(x_1=\text{red})= 4/10$

$p(x_1=\text{blue})= 6/10$

$p(x_2|x_1)$

$p(x_2=\text{a}|x_1=\text{red})= 1/4$

$p(x_2=\text{o}|x_1=\text{red})= 3/4$

$p(x_2=\text{a}|x_1=\text{blue})= 3/4$

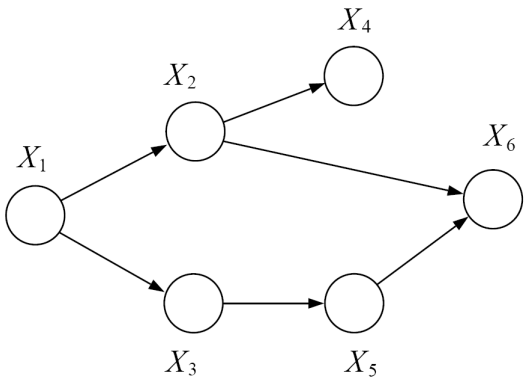
$p(x_2=\text{o}|x_1=\text{blue})= 1/4$

Conditional Independence

Interpret missing edges in terms of conditional independence.

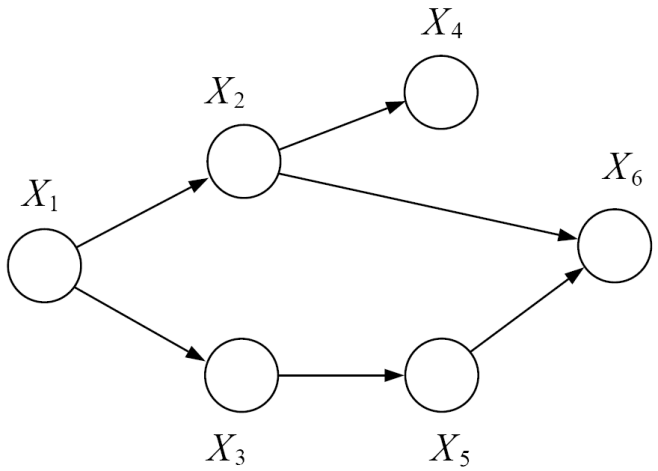
Define an ordering I of the nodes in a graph G to be topological if for every node $i \in V$ the nodes in π_i appear before I in the ordering. For example the ordering $I = \{1,2,3,4,5,6\}$ is a topological ordering for the graph.

Let v_i denote the set of all nodes that appear earlier than I in the ordering I, excluding the parent nodes π_i. Given a topological ordering I for a graph G we associate to the graph the following set of conditional independence statements. $\{X_i \perp X_{v_i} | X_{\pi_i}\}$



Assert the conditional independence of a node from its ancestors, conditional on its parents.

Conditional Independence



$$X_4 \perp \{X_1, X_3\} | X_2$$

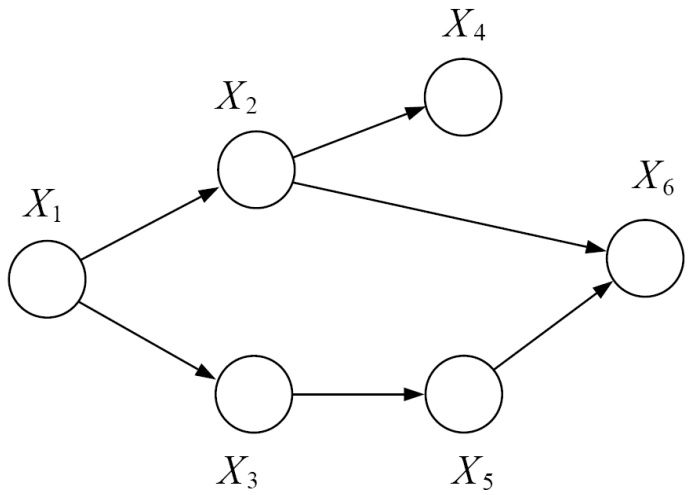
$$P(x_1, x_2, x_3, x_4) = \sum_{x_5}\sum_{x_6} P(x_1, x_2, x_3, x_4, x_5, x_6) = \sum_{x_5}\sum_{x_6} P(x_1)P(x_2|x_2)\, P(x_3|x_1)P(x_4|x_2)P(x_5|x_3)P(x_6|x_2, x_5)$$

$$= P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2)\sum_{x_5} P(x_5|x_3)\sum_{x_6} P(x_6|x_2, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2)$$

$$P(x_1, x_2, x_3) = \sum_{x_4} P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2) = P(x_1)P(x_2|x_1)P(x_3|x_1)$$

$$P(x_4|x_1, x_2, x_3) = P(x_4|x_2)$$

Conditional Independence and the Bayes ball algorithm



Whether there are other conditional independence statements that are true of such joint probability distributions?
Whether these statements also have a graphical interpretation?

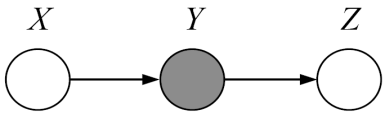
Graph Separation?

X1 is independent of X6 given X2 and X3

X2 is not necessarily independent of X3 given X1 and X6

Make the notion of “blocking” precise

3 canonical graphs

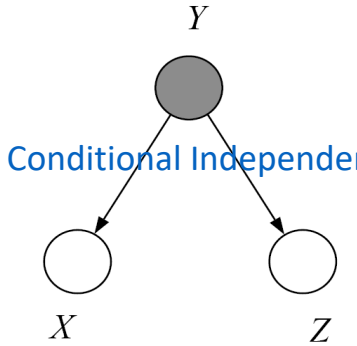


$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

$$P(z|x,y) = \frac{P(x,y,x)}{P(x,y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$X \perp\!\!\!\perp Z \mid Y$$

Classical Markov chain
"Past", "present", "future"



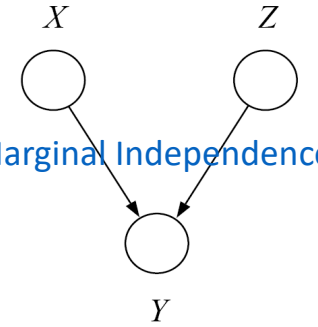
Conditional Independence

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

$$P(x,z|y) = \frac{P(y)P(x|y)P(z|y)}{P(y)} = P(x|y)P(z|y)$$

$$X \perp\!\!\!\perp Z \mid Y$$

Common cause
Y "explains" all the dependencies between X and Z



Marginal Independence

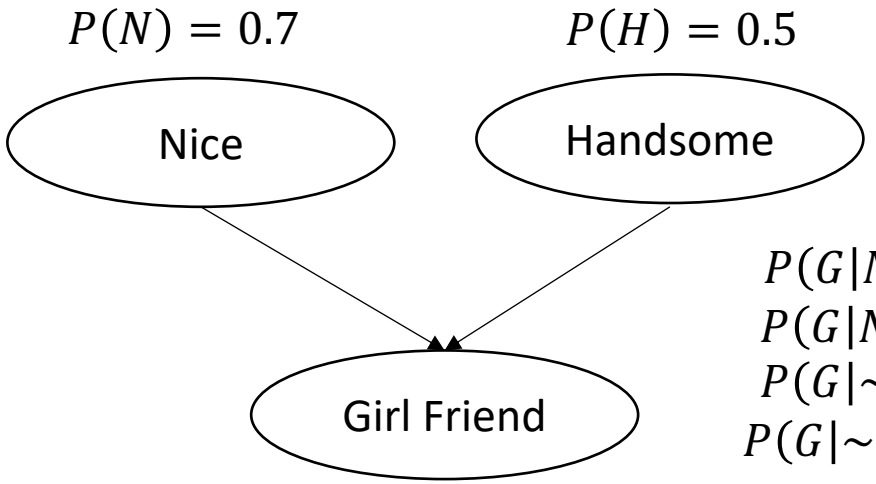
$$P(x,y,z) = P(x)P(z)P(y|x,z) = P(x)P(z)\frac{P(x,y,z)}{P(x,z)}$$

$$X \perp\!\!\!\perp Z$$

Common effect
Multiple, competing explanation

Explaining Away

- Illustration:



$$P(G|N,H) = 0.95$$
$$P(G|N,\sim H) = 0.8$$
$$P(G|\sim N,H) = 0.6$$
$$P(G|\sim N,\sim H) = 0.1$$

$$P(N|G,H) = \frac{P(G|N,H)P(N)}{P(G|H)} = \frac{0.95 * 0.7}{0.845} \approx 0.787$$

$$P(N|G) = \frac{P(G|N)P(N)}{P(G)} = \frac{0.875 * 0.7}{0.7175} \approx 0.854$$

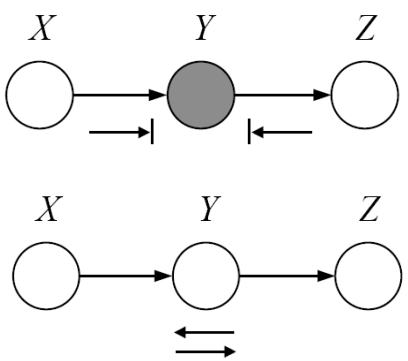
$$P(N|G,H) < P(N|G)$$

- Knowing that a man is handsome decreases the probability that he is a nice guy.
- Knowing that the a man has a girl friend, nice and handsome become dependent:

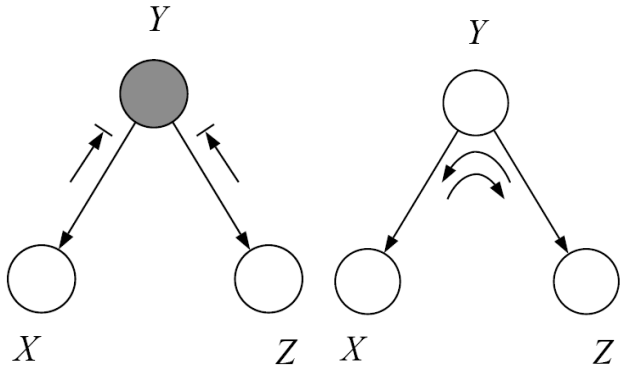
$$P(N|G,H) \neq P(N|G)$$

Conditional Independence (check)

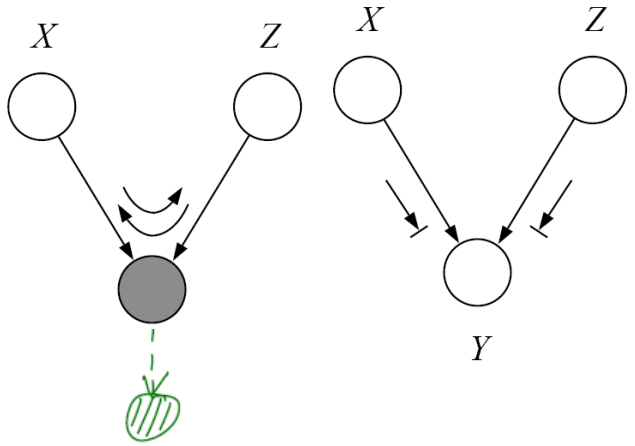
Bayes ball algorithm (rules)



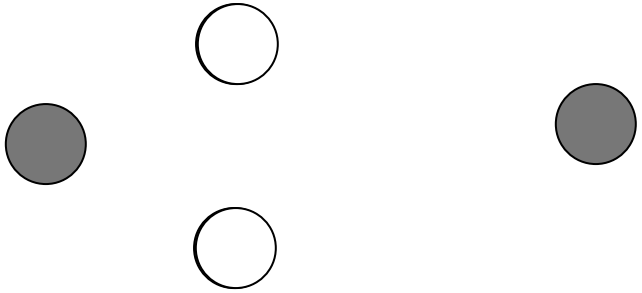
One incoming arrow
and one outgoing arrow



Two outgoing arrows



Two incoming arrows



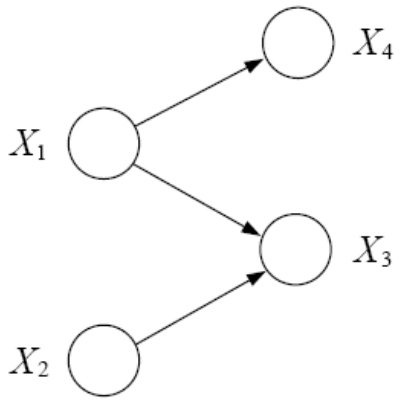
Check through reachability

$$X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$$

$$X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$$

Characterization of directed graphical models

- Graphical model is associated with a family of probability distributions
- Defined in two equivalent ways:



$$p(x_1, \ldots, x_n) \triangleq \prod_{i=1}^{n} p(x_i \mid x_{\pi_i}).$$

$$X_1 \perp\!\!\!\perp X_2$$

$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$

$$X_3 \perp\!\!\!\perp X_4 \mid X_1$$

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

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

The characterizations of probability distributions via numerical parameterization and conditional independence statements are one and the same !

2011 Turing Award was for Bayesian networks

A.M. TURING CENTENARY CELEBRATION WEBCAST

MORE ACM AWARDS

A.M. TURING AWARD

Search TYPE HERE

A.M. TURING AWARD WINNERS BY...

ALPHABETICAL LISTING | YEAR OF THE AWARD | RESEARCH SUBJECT

Photo-Essay

BIRTH:

September 4, 1936, Tel Aviv.

EDUCATION:

B.S., Electrical Engineering (Technion, 1960); M.S., Electronics (Newark College of Engineering, 1961); M.S., Physics (Rutgers University, 1965); Ph.D., Electrical Engineering (Polytechnic Institute of Brooklyn, 1965).

EXPERIENCE:

Research Engineer, New York University Medical School (1960–1961); Instructor,

JUDEA PEARL

United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

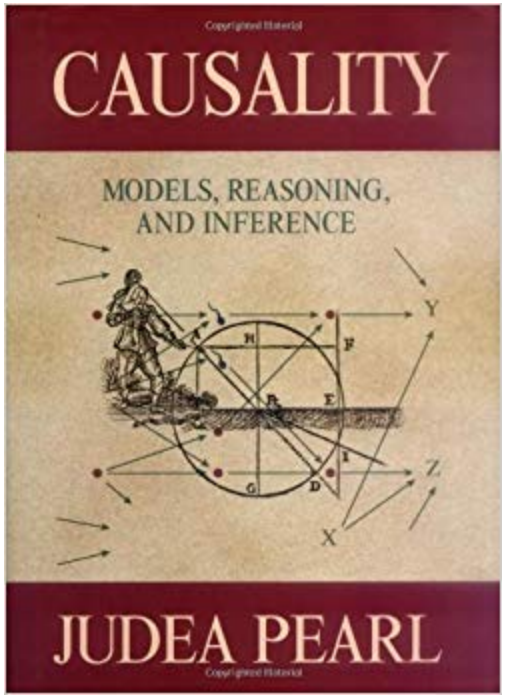
SHORT ANNOTATED BIBLIOGRAPHY | ACM DL AUTHOR PROFILE | ACM TURING AWARD LECTURE VIDEO | RESEARCH SUBJECTS | ADDITIONAL MATERIALS

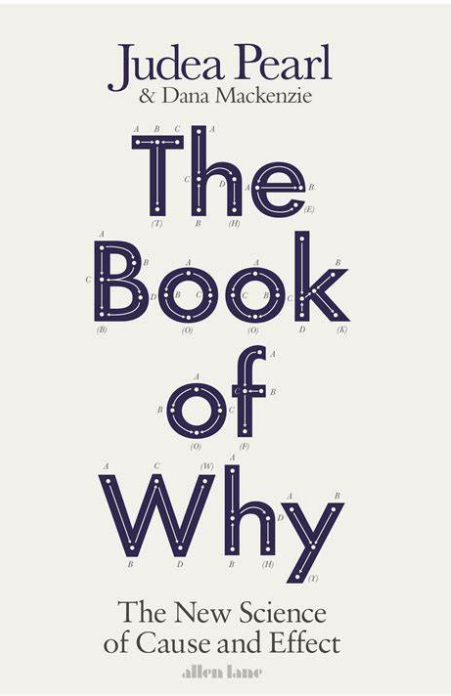
Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

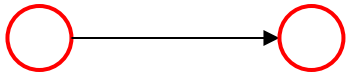
He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for causal inference that has had significant impact in the social sciences.

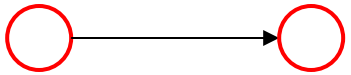
Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in Bnei Brak, a Biblical town his grandfather went to reestablish in 1924. In 1956, after serving in the Israeli army and joining a Kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a B.S. degree in Electrical Engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the *Technion Magazine*, he emphasized the thrill of discovery:

From correlation to causality





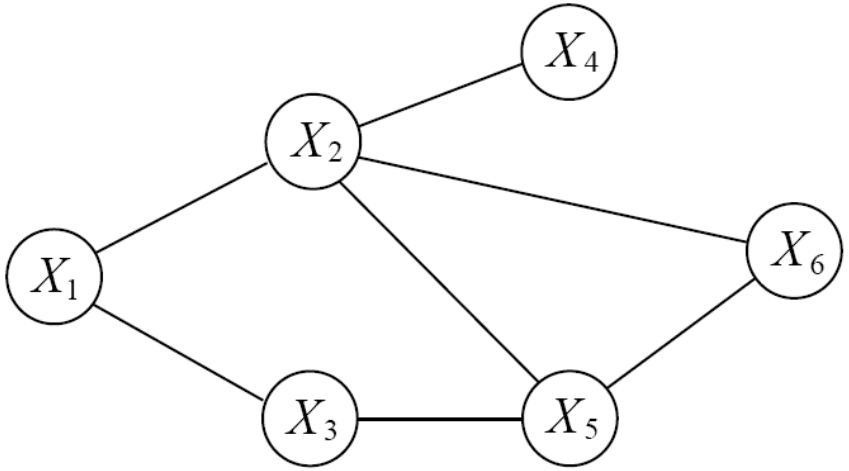




Undirected PGM

Undirected PGM (MRF)

Definition An undirected graph $G = (V, E)$has a finite set of vertices(nodes) V and a set of edges E that consists of a pair of vertices



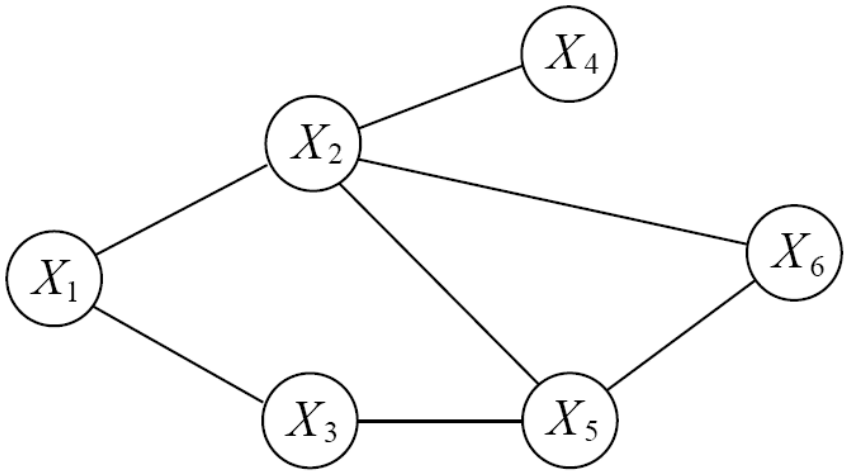
Probability Distribution → Queries

Representation → Implementation

Conditional Independence → Interpretation

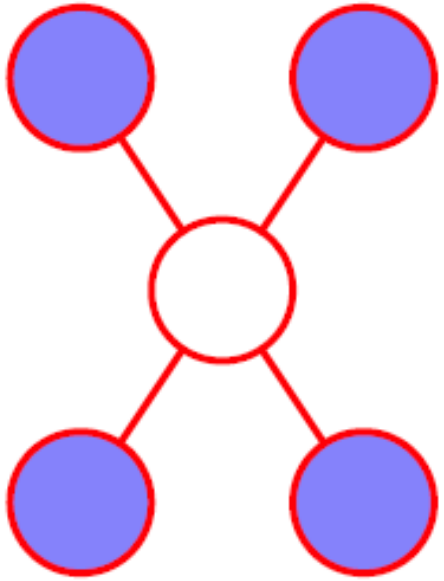
Conditional Independence

naive graph-theoretic separation



$$X_A \perp X_C | X_B$$

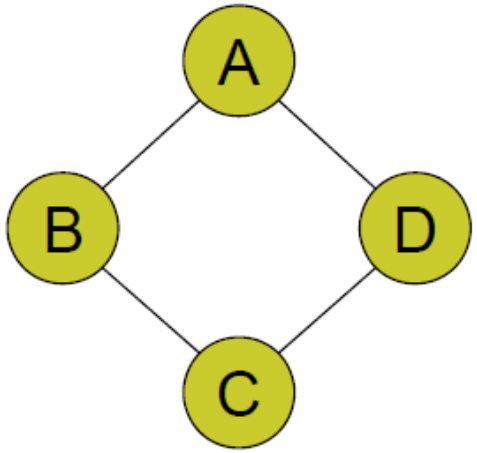
It's a "reachability" problem in graph theory.



For an undirected graph, the Markov blanket of a node consists of the set of neighboring nodes.

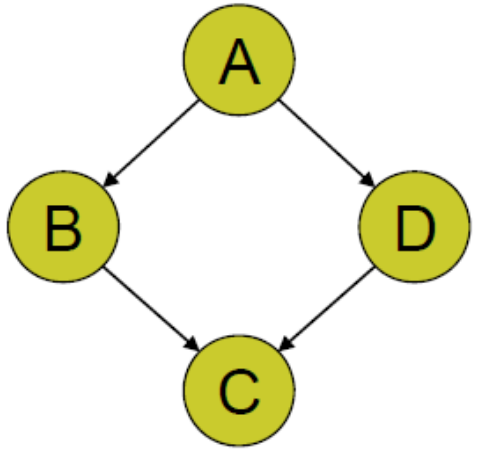
Directed versus Undirected

Is it possible to reduce undirected graph to directed graph?



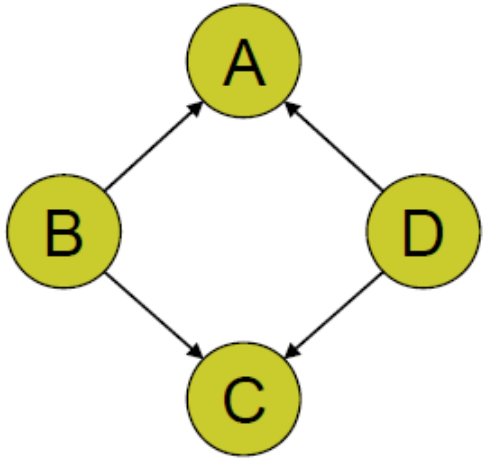
$A \perp C \mid \{B,D\}$

$B \perp D \mid \{A,C\}$



$A \perp C \mid \{B,D\}$

$B \perp D \mid A$ ✗

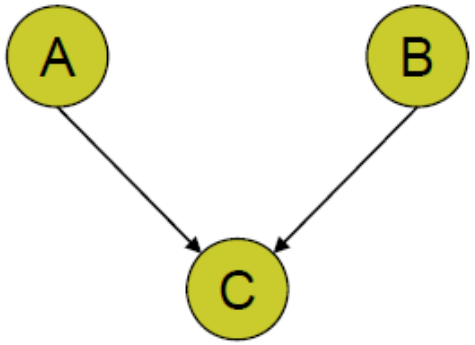


$A \perp C \mid \{B,D\}$

$B \perp D$ ✗

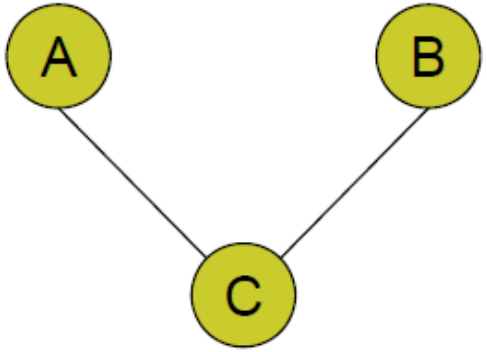
Directed versus Undirected

Is it possible to reduce directed graph to undirected graph?



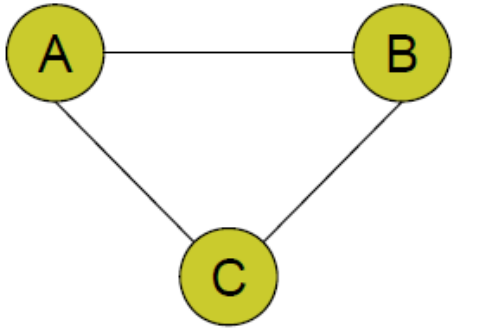
$A \perp B$

$\neg (A \perp B \mid C)$



$A \perp B \mid C$

$\neg (A \perp B)$

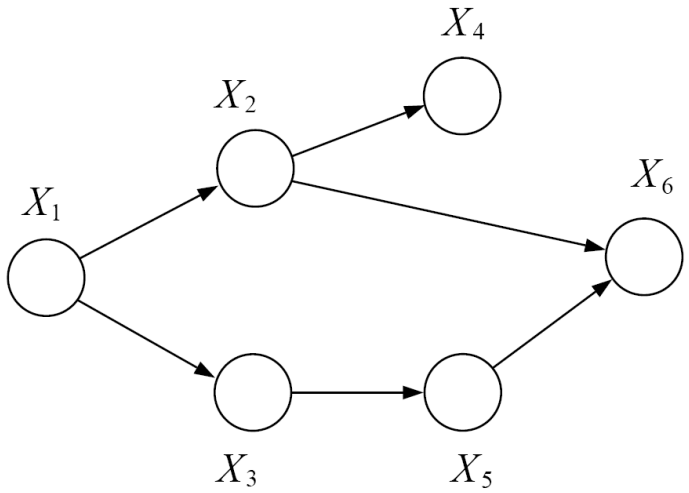


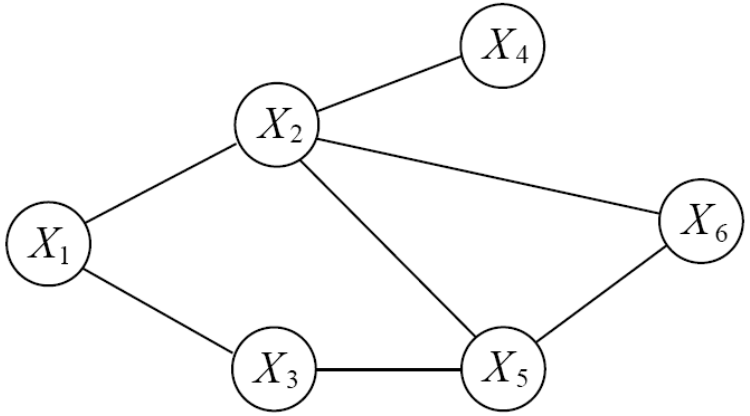
$\neg (A \perp B \mid C)$

$\neg (A \perp B)$

Probability Distribution(1)

- Directed Graphs: utilize “local” parameterization to obtain the joint probabilities
- Undirected Graphs: utilize conditional probabilities to represent the joint ?
  - Consistency problem
- Abandon conditional probabilities
  - Lose local probabilistic interpretation
  - retain the ability to choose these functions independently and arbitrarily
  - retain the all-important representation of the joint as a product of local functions





Probability Distribution(2)

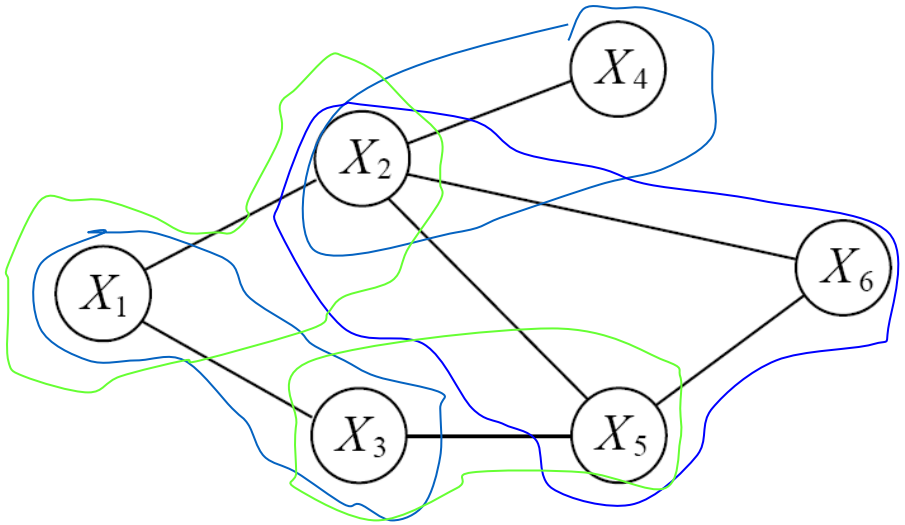
- Key problem: decide the domain of the local functions
  - Conditional Independence: Graph Separation

- Clique
  - A clique of a graph is a fully-connected subset of nodes.
  - Local functions should not be defined on domains of nodes that extend beyond the boundaries of cliques.
- Maximal cliques
  - The maximal cliques of a graph are the cliques that cannot be extended to include additional nodes without losing the probability of being fully connected.
  - We restrict ourselves to maximal cliques without loss of generality, as it captures all possible dependencies.

- Potential function (local parameterization)
  - $\varphi_{X_c}(x_c)$: potential function on the possible realizations x_c of the maximal clique X_c
  - (Strictly) positive, real-valued function

Probability Distribution(3)

Maximal cliques



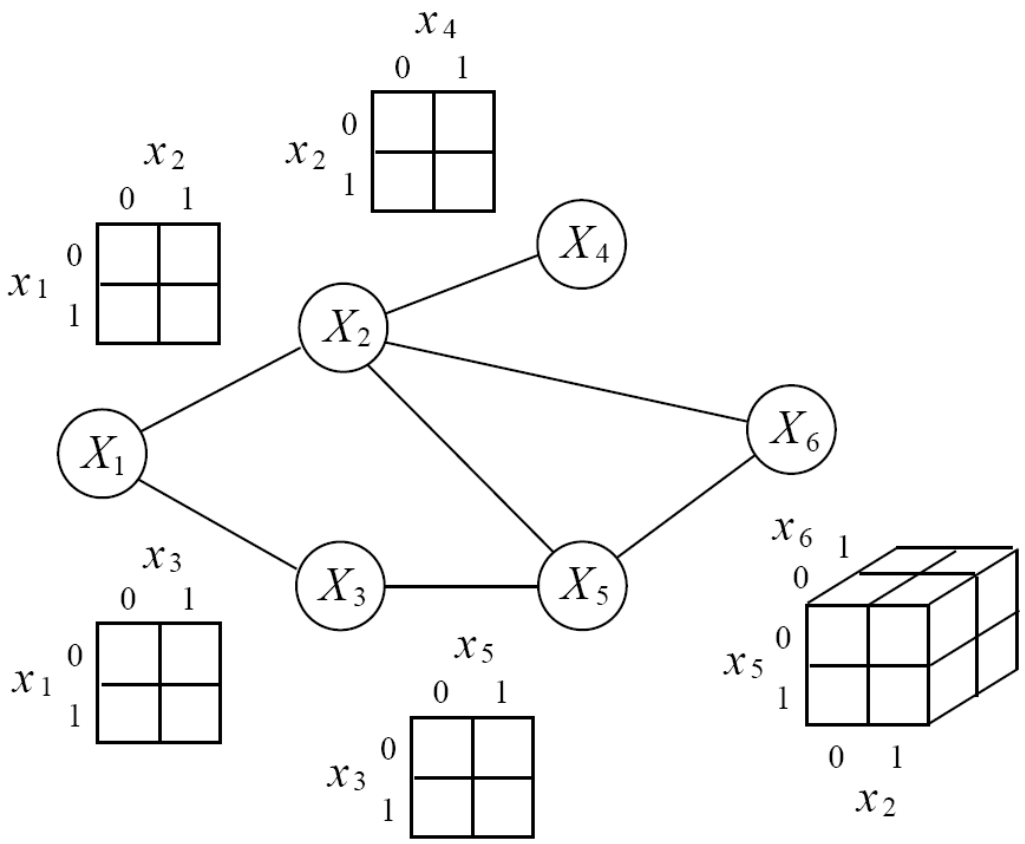
Joint probability distribution

$$P(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}}\varphi_{X_c}(x_c)$$

Normalization factor

$$Z = \sum_{x}\prod_{C\in\mathcal{C}}\varphi_{X_c}(x_c)$$

Representation

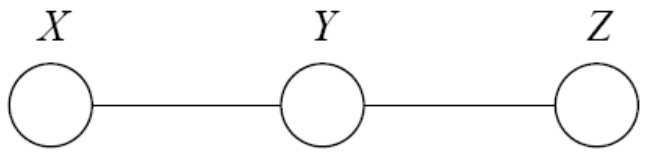


$O(2^n) \longrightarrow O(r*2^k)$

The interpretation of potential functions (1)

- Why not replace the potential functions *with marginal* probabilities $p(x_c)$ *?*

$$P(x) = \frac{1}{Z}\prod_{C \in \mathcal{C}} \varphi_{X_c}(x_c)$$



$X \perp Z | Y$

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

$$P(x,y,z) \neq P(x,y)P(z)$$

$$P(x,y,z) = P(x,y)P(y,z) implies\ P(y) = 0\ or\ P(y) = 1$$

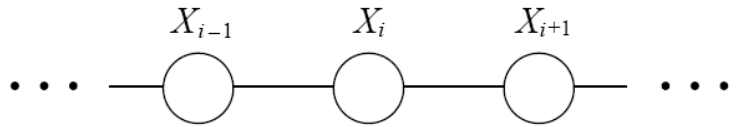
The interpretation of potential functions (2)

- In general, potential functions are neither conditional probabilities nor marginal probabilities
- Natural interpretation
  - agreement, constraint, or energy
- To represent potentials in an unconstrained form

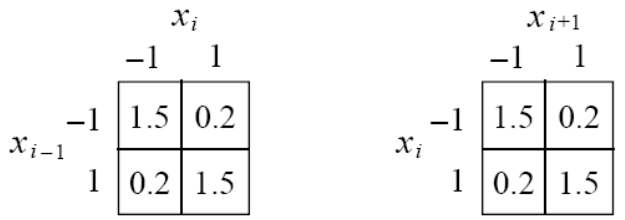
$$P(x) = \frac{1}{Z}\prod_C \varphi_{X_c}(X_C) = \frac{1}{Z}\prod_C \exp\{-H_C(X_C)\} = \frac{1}{Z}\exp\left\{-\sum_c H_C(X_C)\right\} = \frac{1}{Z}\exp\{-H(x)\}$$

$$Z = \sum_x \prod_C \varphi_{X_c}(X_c) = \sum_x \exp\{-H(x)\}$$

Boltzman distribution



(a)



(b)

Magnetic behavior of crystals

Hidden Markov Model

Introduction

- Hidden Markov Model (HMM) is a graphical model for modeling sequential data.

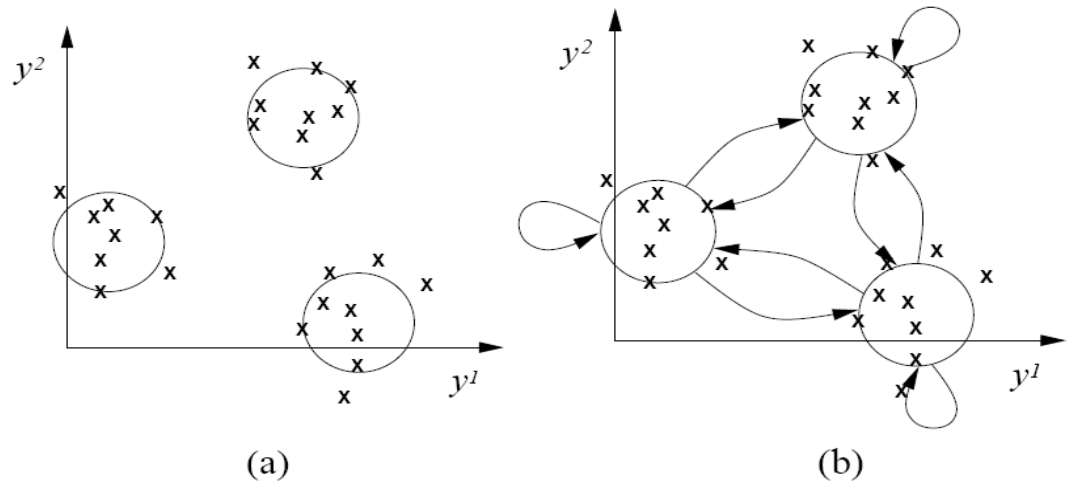
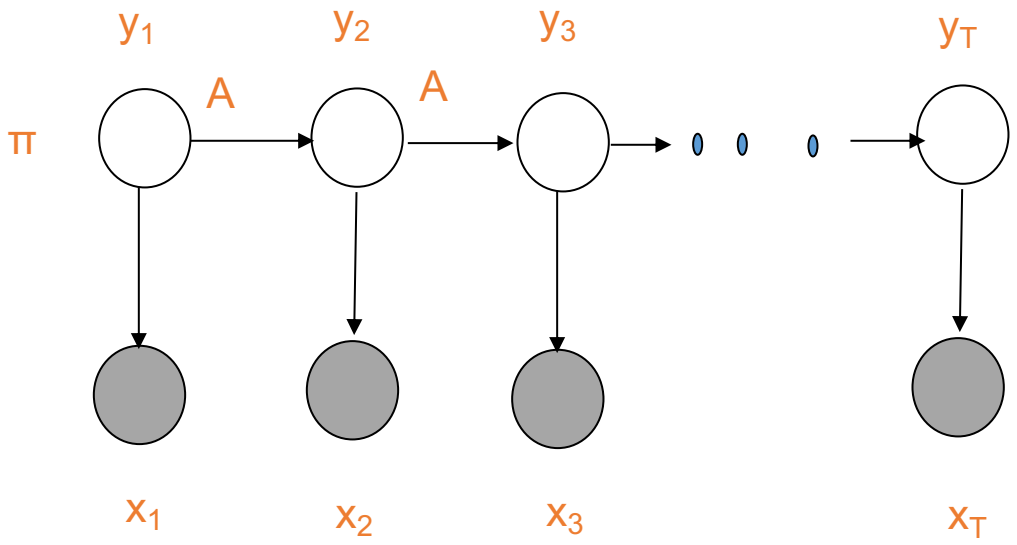


Figure 12.1: (a). A sample point is generated from a mixture model by first selecting a mixture component and then generating a data point from that mixture component. (b) An HMM generalizes the mixture model by allowing the choice of the mixture component at a given step to depend on the choice of the mixture component at the previous step. The arrows in the diagram represent these transitions between the mixture components.

- A generalization of mixture model

HMM



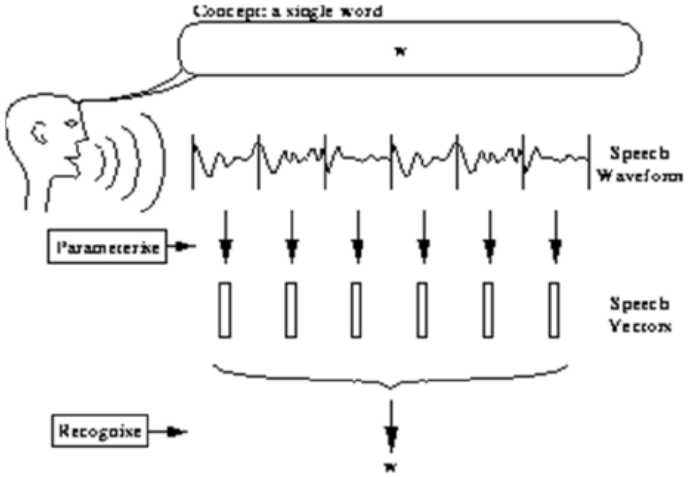
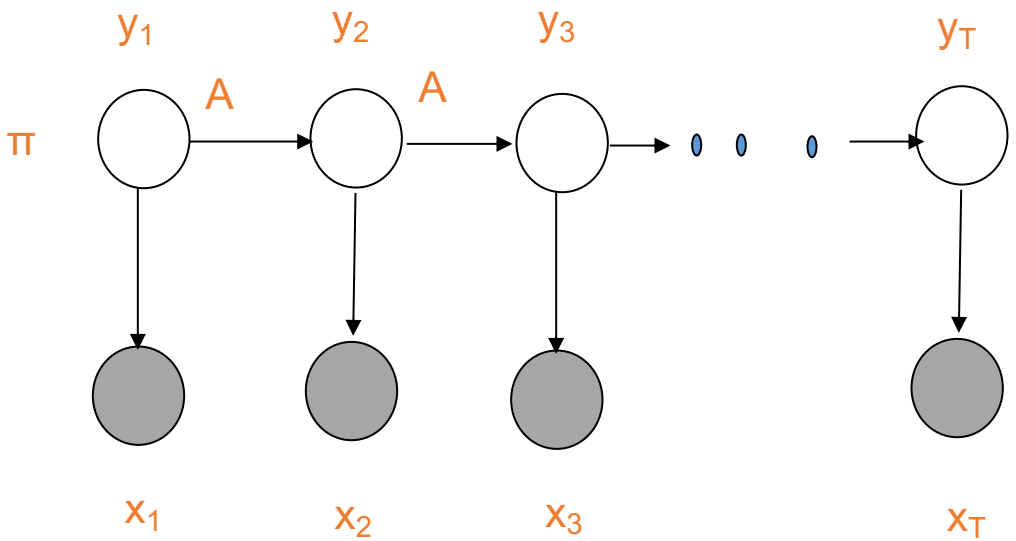


Fig. 1.2 Isolated Word Problem

Top node in each slice represents the multinomial y_t variable and the bottom node represents the observable x_t variable

HMM

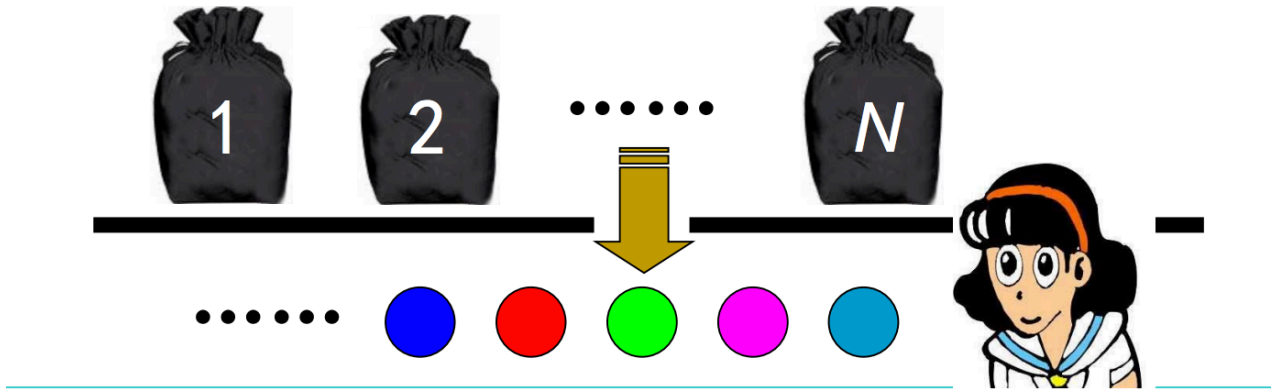
- Conditioning on state y_t renders y_{t-1} and y_{t+1} independent.
- Generally, y_s is independent of y_u, for s<t and t<u.
- This is also for output nodes x_S and x_u, when conditioned on state node y_t.
- Conditioning on output node, does not yield any conditional independence.



Example

- Example: N bags, each with M balls of different colors. An experimenter selects a bag according to a probability distribution, then randomly takes a ball according to the probability distribution of different colored balls in the bag, and reports the color of the ball.

- To outsiders: the observable process is a sequence of balls of different colors, while the sequence of bags is unobservable. Each bag corresponds to a state in the HMM; the color of the ball corresponds to the output of the state in the HMM.



Composition of Hidden Markov Model

1. The number of states in the model is N (the number of bags)

2. The number of different symbols M (number of different colored balls) that can be output from each state

3. State transition probability matrix $A = a_{ij}$, a_{ij} is the probability that the experimenter takes the ball from one bag (state S_i) to another bag (state S_j).

4. From the state S_j, the probability distribution matrix of a certain symbol v_k is observed as: $B = b_j(k)$ where $b_j(k)$ is the probability that the experimenter takes the k-th color ball from the j-th bag.

5. The probability distribution of the initial state is: $\pi = \pi_i$

For convenience, the HMM is generally written as: $\mu = (A, B, \pi)$
Or $\mu = (S, O, A, B, \pi)$ to indicate the parameter set of the model

Three Fundamental Problems

- Likelihood evaluation problem: Given the model $\mu = (A, B, \pi)$ and the observation sequence $O = O_1 O_2 \ldots O_T$, how to quickly calculate the probability $P(O|\mu)$?

  ➡ Forward Algorithm

- State sequence decoding (or inference) problem: Given the model $\mu = (A, B, \pi)$ and the observation sequence $O = O_1 O_2 \ldots O_T$, how to choose the "optimal" state sequence $S = s_1 s_2 \ldots s_T$ in a certain sense so that State sequence "best explained" observation sequence?

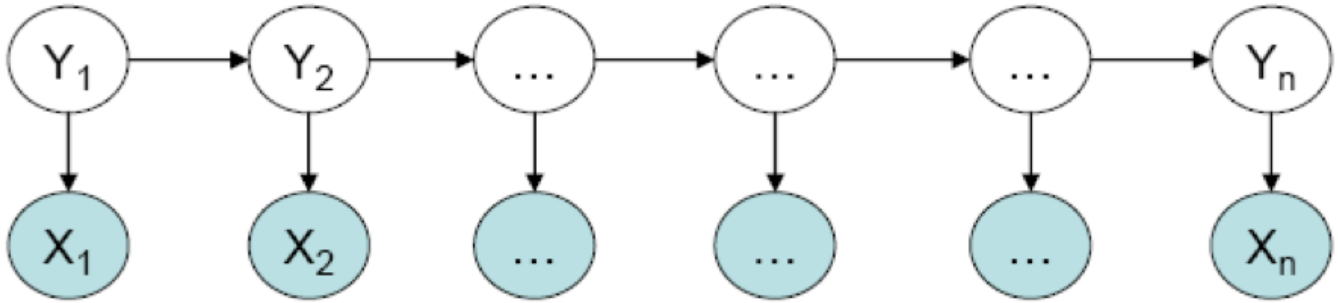
  ➡ Viterbi Algorithm

- Parameter estimation (or learning) problem: Given an observation sequence $O = O_1 O_2 \ldots O_T$, how to find the parameter values of the model according to the maximum likelihood estimation? That is, how to adjust the parameters of the model so that $P(O|\mu)$ is the largest?

  ➡ Baum-Welch or forward-backward procedure

Conditional Random Fields

Shortcomings of Hidden Markov Model



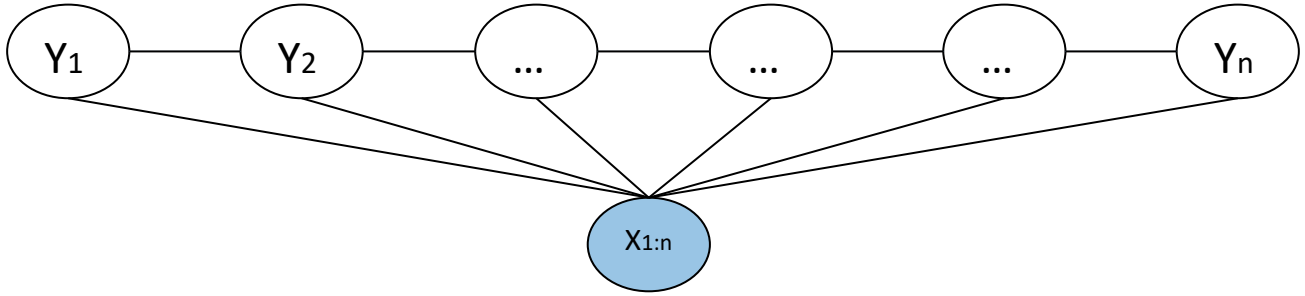
- HMM models capture dependences between each state and only its corresponding observation
  - NLP examples : In a sentence segmentation task, each segmental state may depend not just on a single word (and the adjacent segmental stages), but also on the (non-local) features of the whole line such as line length, indentation , amount of white space, etc
- Mismatch between learning objective function and prediction objective function
  - HMM learns a joint distribution of states and observations P(Y,X) , but in a prediction task, we need the conditional probability P(Y|X)

CRF

- Conditional random fields (CRFs), proposed by J. Lafferty and others in 2001, are probabilistic structural models used to label and divide sequence structure data, and have been widely used in NLP and image processing.

- Basic idea: Given the observation sequence X, output the identification sequence Y, and calculate the optimal labeled sequence by calculating P(Y|X).

From HMM to CRF

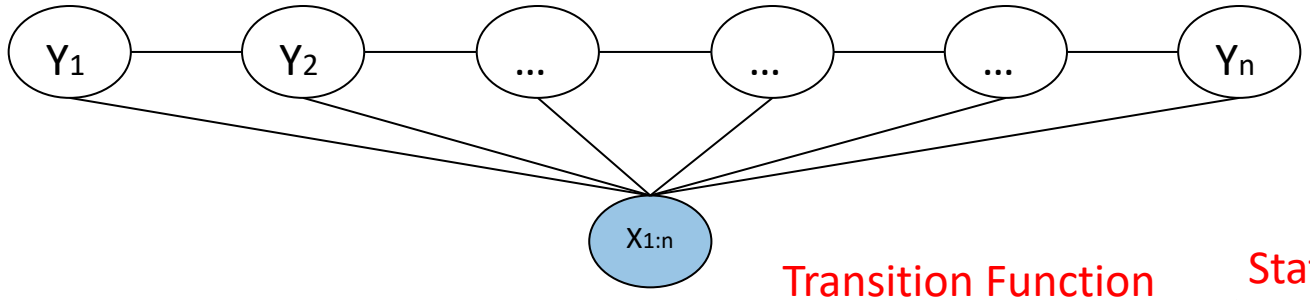


$$P(y_{1:n}|x_{1:n}) = \frac{1}{Z(x_{1:n})}\prod_{i=1}^{n}\phi(y_i, y_{i-1}, x_{1:n}) = \frac{1}{Z(x_{1:n})}\prod_{i=1}^{n}\exp(w^T f(y_i, y_{i-1}, x_{1:n}))$$

- CRF is an undirected graph model
  - Discriminative model
  - Models the dependence between each state and the entire observation sequence

Conditional Random Fields

- General parametric form:



Transition Function

State Function

$$P(y|x) = \frac{1}{Z(x,\lambda,\mu)} \exp\left(\sum_{i=1}^{n}\left(\sum_k \lambda_k t_k(y_i, y_{i-1}, x) + \sum_l \mu_l s_l(y_i, x)\right)\right)$$

$$= \frac{1}{Z(x,\lambda,\mu)} \exp\left(\sum_{i=1}^{n}\left(\lambda^T t(y_i, y_{i-1}, x) + \mu^T s(y_i, x)\right)\right)$$

$$where\ Z(x,\lambda,\mu) = \sum_y \exp\left(\sum_{i=1}^{n}\left(\lambda^T t(y_i, y_{i-1}, x) + \mu^T s(y_i, x)\right)\right)$$

CRF

- Probability calculation: Given P(**y**|**x**), X and Y, calculate conditional probability $P(Y_i = y_i|x)P(Y_{i-1} = y_{i-1}, Y_i = y_i|x)$ and corresponding expectations

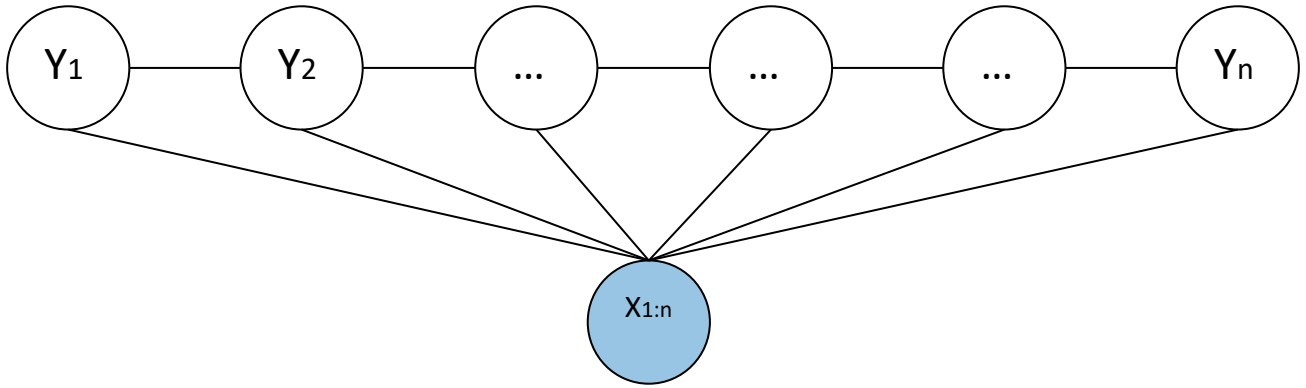
  ➡ Forward-backward Algorithm

- Inference: Given CRF parameters λ and μ, find the $\boldsymbol{y}^*$ that maximizes P(**y**|**x**)

  ➡ Viterbi Algorithm

- Parameter Learning: Given $\{(x_d, y_d)\}_{d=1}^{N}$, find λ^*, μ^*

  ➡ Gradient Descent Algorithms or Quasi-Newton Methods



Application: Word Segmentation

Basic idea:

- Think of the word segmentation process as the classification of words: each word occupies a certain word-forming position (ie, the lexeme) when constructing a specific word.

- In general, each word has only 4 positions: the beginning (B), the middle (M), the ending (E), and the individual formation (S).

Application: Word Segmentation

- 乒乓球拍卖完了。

(1)乒乓球/ 拍/ 卖/ 完/ 了/ 。/

(2)乒乓球/ 拍卖/ 完/ 了/ 。/

(3)乒/B 乓/M 球/E 拍/S 卖/S 完/S 了/S 。/S

乒/B 乓/M 球/E 拍/S 卖完了。

B, E, M, S ？

特征：

- 当前字的前后 n 个字
- 当前字左边字的标记
- 当前字在词中的位置

Word Segmentation: State Function

- 一元特征：当前字、 当前字的前一个字、当前字的后一个字
- 二元特征：各标记间的转移特征

$$s_1(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”， 当前字的标记}y_i\text{是M} \\ 0 & \text{否则} \end{cases}$$

$$s_2(y_i, X, i) = \begin{cases} 1 & \text{如果当前字是“拍”， 当前字}y_i\text{的标记是E} \\ 0 & \text{否则} \end{cases}$$

……

乒/B 乓/M 球/E 拍/S 卖？ 完了。

Word Segmentation: Transition Function

对应转移函数的特征：

$$t_1(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标记}y_{i-1}\text{是B，当前字的标记}y_i\text{是M} \\ 0 & \text{否则} \end{cases}$$

$$t_2(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{如果前一个字的标记}y_{i-1}\text{是M，当前字的标记}y_i\text{是M} \\ 0 & \text{否则} \end{cases}$$

……

乒/B 乓/M 球/E 拍/S 卖？ 完了。

Word Segmentation

- Parameter training: Training corpus to estimate the feature weights λ_j so that it can find a most likely label sequence Y given an observation sequence X, that is, the conditional probability P(Y|X) is the largest
- Negative log-likelihood function:

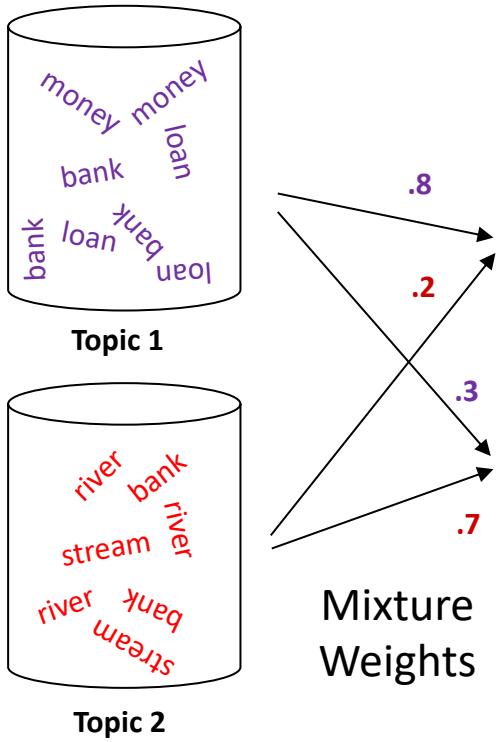
$$L(\boldsymbol{\lambda}) = -\log p(Y \mid X, \boldsymbol{\lambda}) + \frac{\varepsilon}{2}\boldsymbol{\lambda}^2$$

- After determining the parameters, decode them to find the optimal sequence

PLSA

Probabilistic Latent Semantic Analysis

- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- Each document is represented as a list of mixing proportions for the mixture components.



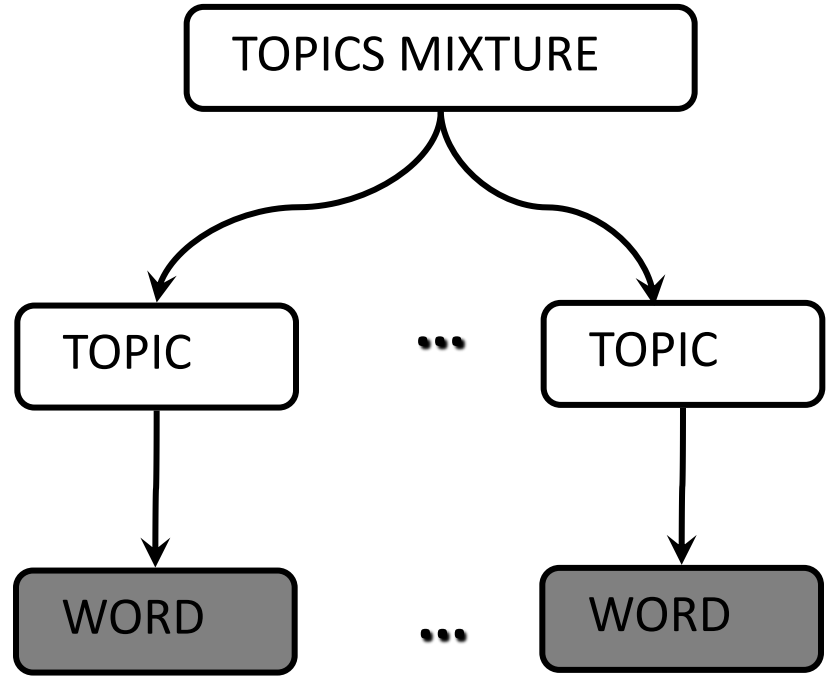
DOCUMENT 1: money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] money[1] stream[2] bank[1] money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] bank[1] money[1] stream[2]

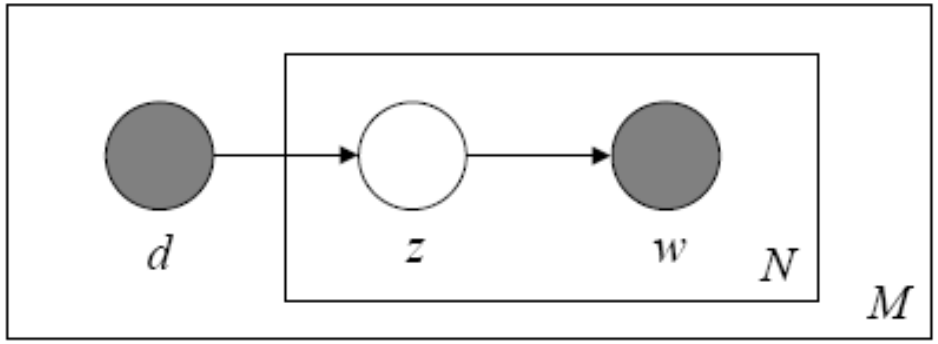
DOCUMENT 2: river[2] stream[2] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] bank[1] stream[2] river[2] loan[1] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] bank[2] stream[2] bank[2] money[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] money[1] bank[1] stream[2] river[2] bank[2] stream[2] bank[2] money[1]

Document Generation as a Probabilistic Process

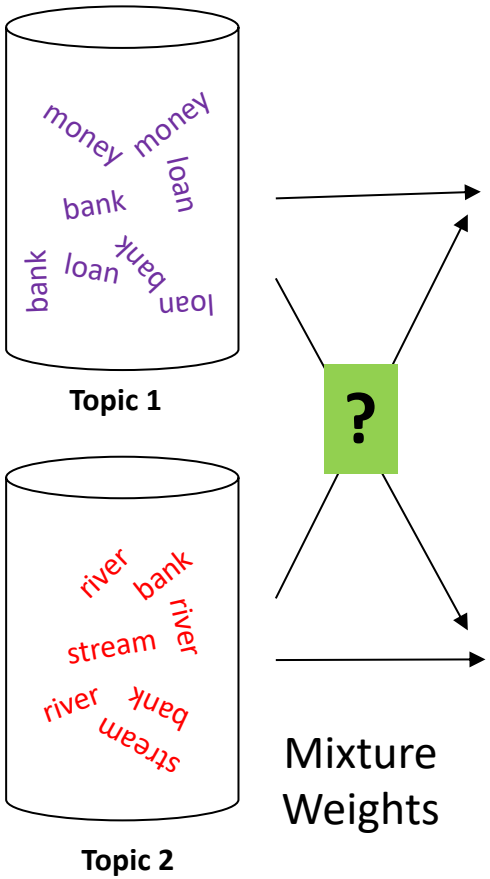
- For each document, choose a mixture of topics
- For every word slot, sample a topic [1..T] from the mixture

$$p(d, w_n) = p(d)p(w|d)$$
$$= p(d) \sum_z p(w_n|z)p(z|d)$$
$$= \sum_{z \in Z} P(z)P(d|z)P(w|z)$$





Inference



文档1: money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] money[1] stream[2] bank[1] money[1] bank[1] bank[1] loan[1] river[2] stream[2] bank[1] money[1] river[2] bank[1] money[1] bank[1] loan[1] bank[1] money[1] stream[2]

文档2: river[2] stream[2] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] bank[1] stream[2] river[2] loan[1] bank[2] stream[2] bank[2] money[1] loan[1] river[2] stream[2] bank[2] stream[2] bank[2] money[1] river[2] stream[2] loan[1] bank[2] river[2] bank[2] money[1] bank[1] stream[2] river[2] bank[2] stream[2] bank[2] money[1]

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

EM

- E-step: Calculate the posterior probability of latent variable

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

- M-step: Update parameters

$$P(w|z) \propto \sum_{d \in D} n(d,w)P(z|d,w)$$

$$P(d|z) \propto \sum_{w \in W} n(d,w)P(z|d,w)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w)P(z|d,w)$$

Example: topics found from a Science Magazine papers collection

universe	0.0439
galaxies	0.0375
clusters	0.0279
matter	0.0233
galaxy	0.0232
cluster	0.0214
cosmic	0.0137
dark	0.0131
light	0.0109
density	0.01

drug	0.0672
patients	0.0493
drugs	0.0444
clinical	0.0346
treatment	0.028
trials	0.0277
therapy	0.0213
trial	0.0164
disease	0.0157
medical	0.00997

cells	0.0675
stem	0.0478
human	0.0421
cell	0.0309
gene	0.025
tissue	0.0185
cloning	0.0169
transfer	0.0155
blood	0.0113
embryos	0.0111

sequence	0.0818
sequences	0.0493
genome	0.033
dna	0.0257
sequencing	0.0172
map	0.0123
genes	0.0122
chromosome	0.0119
regions	0.0119
human	0.0111

years	0.156
million	0.0556
ago	0.045
time	0.0317
age	0.0243
year	0.024
record	0.0238
early	0.0233
billion	0.0177
history	0.0148

bacteria	0.0983
bacterial	0.0561
resistance	0.0431
coli	0.0381
strains	0.025
microbiol	0.0214
microbial	0.0196
strain	0.0165
salmonella	0.0163
resistant	0.0145

male	0.0558
females	0.0541
female	0.0529
males	0.0477
sex	0.0339
reproductive	0.0172
offspring	0.0168
sexual	0.0166
reproduction	0.0143
eggs	0.0138

theory	0.0811
physics	0.0782
physicists	0.0146
einstein	0.0142
university	0.013
gravity	0.013
black	0.0127
theories	0.01
aps	0.00987
matter	0.00954

immune	0.0909
response	0.0375
system	0.0358
responses	0.0322
antigen	0.0263
antigens	0.0184
immunity	0.0176
immunology	0.0145
antibody	0.014
autoimmune	0.0128

stars	0.0524
star	0.0458
astrophys	0.0237
mass	0.021
disk	0.0173
black	0.0161
gas	0.0149
stellar	0.0127
astron	0.0125
hole	0.00824

Polysemy

PRINTING
PAPER
PRINT
PRINTED
TYPE
PROCESS
INK
PRESS
IMAGE
PRINTER
PRINTS
PRINTERS
COPY
COPIES
FORM
OFFSET
GRAPHIC
SURFACE
PRODUCED
CHARACTERS

PLAY
PLAYS
STAGE
AUDIENCE
THEATER
ACTORS
DRAMA
SHAKESPEARE
ACTOR
THEATRE
PLAYWRIGHT
PERFORMANCE
DRAMATIC
COSTUMES
COMEDY
TRAGEDY
CHARACTERS
SCENES
OPERA
PERFORMED

TEAM
GAME
BASKETBALL
PLAYERS
PLAYER
PLAY
PLAYING
SOCCER
PLAYED
BALL
TEAMS
BASKET
FOOTBALL
SCORE
COURT
GAMES
TRY
COACH
GYM
SHOT

JUDGE
TRIAL
COURT
CASE
JURY
ACCUSED
GUILTY
DEFENDANT
JUSTICE
EVIDENCE
WITNESSES
CRIME
LAWYER
WITNESS
ATTORNEY
HEARING
INNOCENT
DEFENSE
CHARGE
CRIMINAL

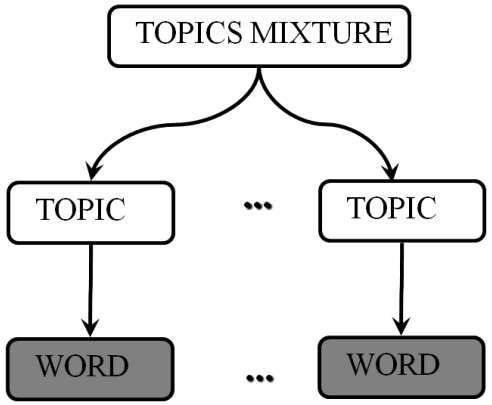
HYPOTHESIS
EXPERIMENT
SCIENTIFIC
OBSERVATIONS
SCIENTISTS
EXPERIMENTS
SCIENTIST
EXPERIMENTAL
TEST
METHOD
HYPOTHESES
TESTED
EVIDENCE
BASED
OBSERVATION
SCIENCE
FACTS
DATA
RESULTS
EXPLANATION

STUDY
TEST
STUDYING
HOMEWORK
NEED
CLASS
MATH
TRY
TEACHER
WRITE
PLAN
ARITHMETIC
ASSIGNMENT
PLACE
STUDIED
CAREFULLY
DECIDE
IMPORTANT
NOTEBOOK
REVIEW

LDA

From PLSA to LDA

- Problem of PLSA:
  - Incomplete: Provide no probabilistic model at the level of documents
  - The number of parameters in the model grows linear with the size of the corpus
  - It is not clear how to assign probability to a document outside of the training data



Latent Dirichlet allocation

- Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus.

- Generative process for each document **w** in a corpus *D*:

1. Choose $N \sim \text{Poisson}(\xi)$
2. Choose $\theta \sim \text{Dir}(\alpha)$ $\quad p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 = 1} \ldots \theta_k^{\alpha_k - 1}$
3. For each of the *N* words w_n
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

β_{ij} is a a element of $k \times V$ matrix, $\beta_{ij} = p(w^j = 1 | z^i = 1)$

Latent Dirichlet allocation

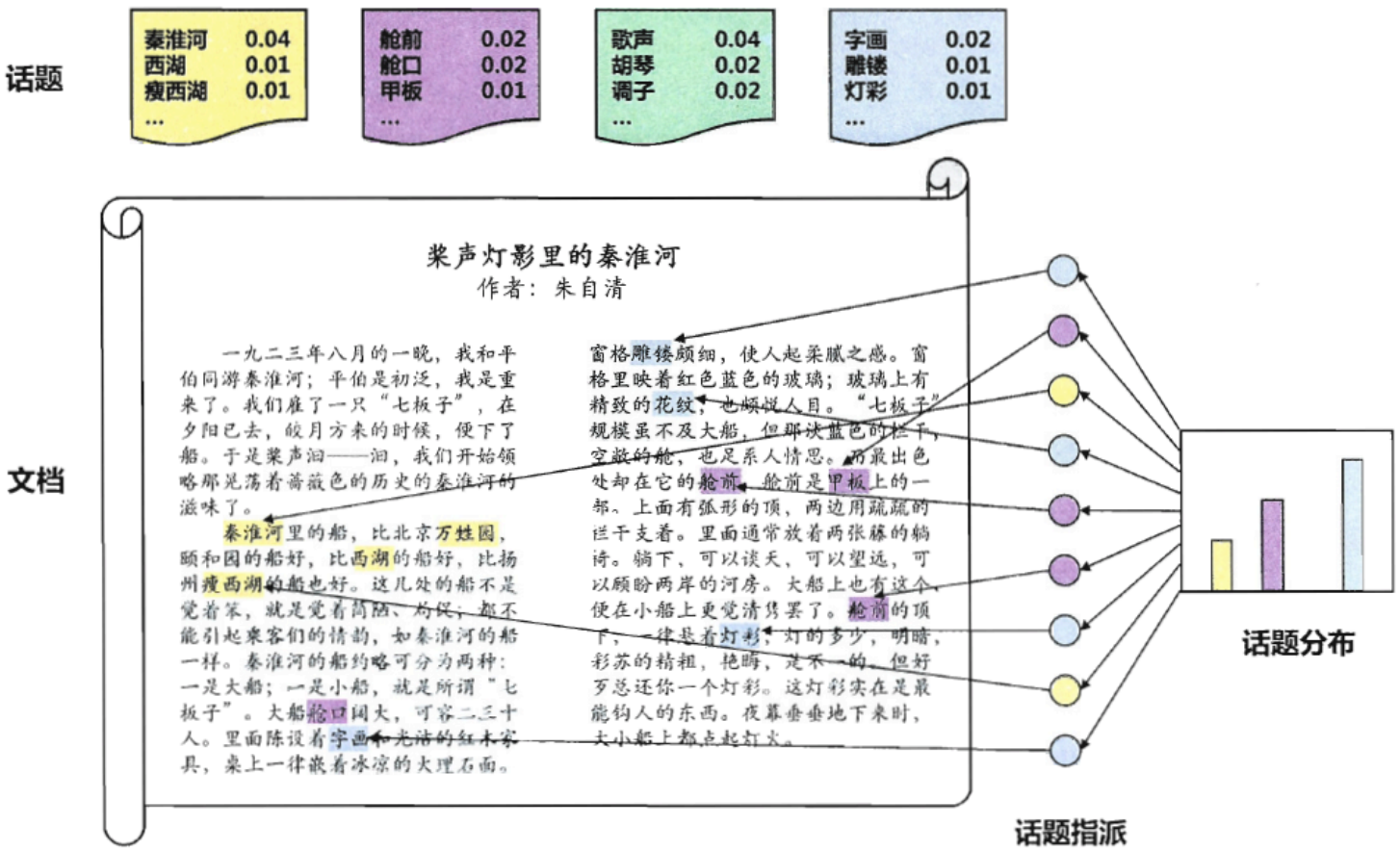
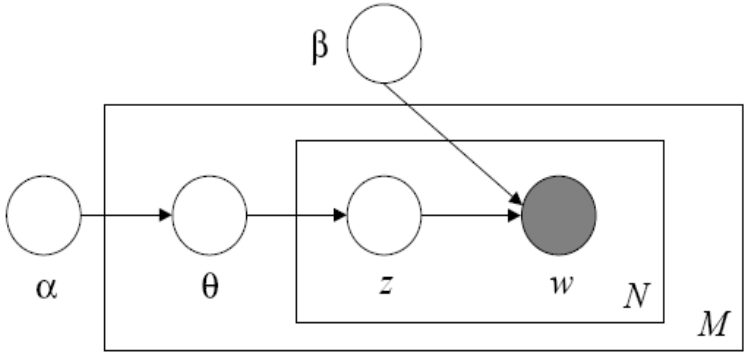


图 14.11 LDA 的文档生成过程示意图

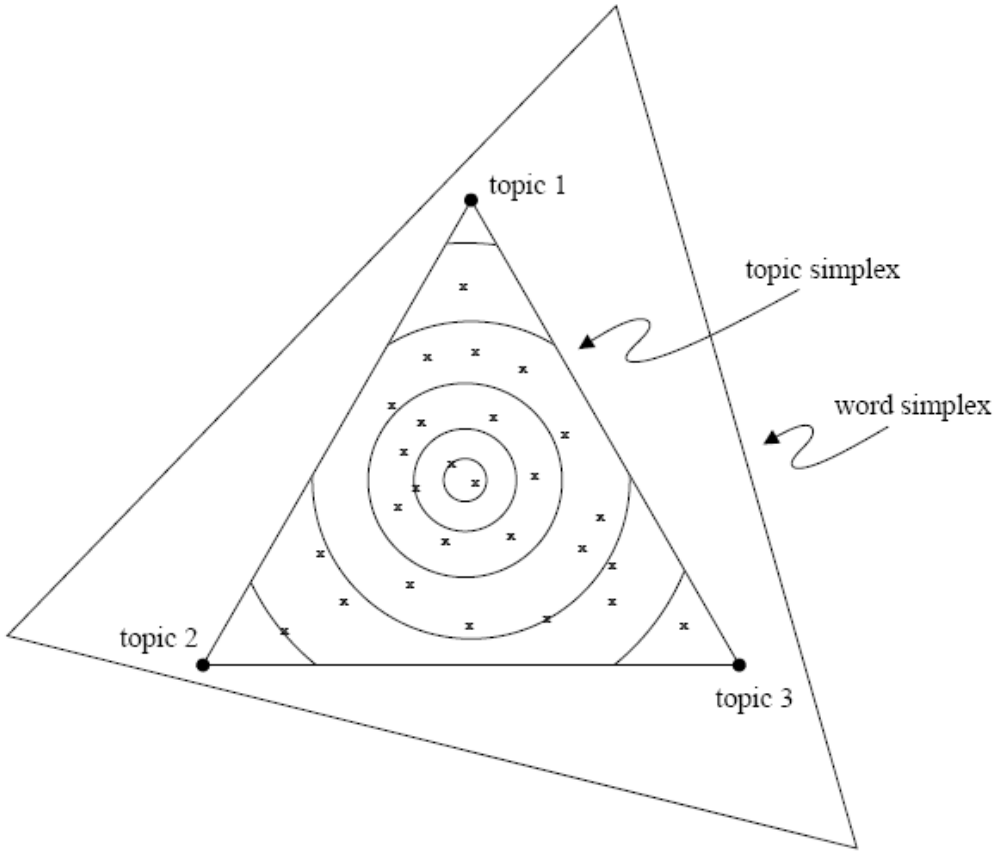


- There are three levels to LDA representation
  - α, β are corpus level parameters
  - θ_d is a document-level variables
  - z_{dn}, w_{dn} are word-level variables

Geometric Interpretation

Three topics and three words

- The difference between LDA and PLSA is that LDA has a smoother distribution on topic simplex.



Distributions

- The joint distribution of a topic θ, and a set of *N* topic **z**, and a set of *N* words **w**:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

- Marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta$$

- Probability of a corpus:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

Inference and parameter estimation

- The key inferential problem is that of computing the posteriori distribution of the hidden variable given a document

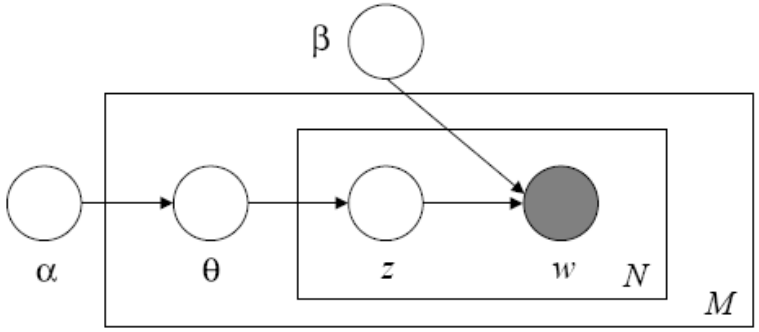
$$p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)}$$

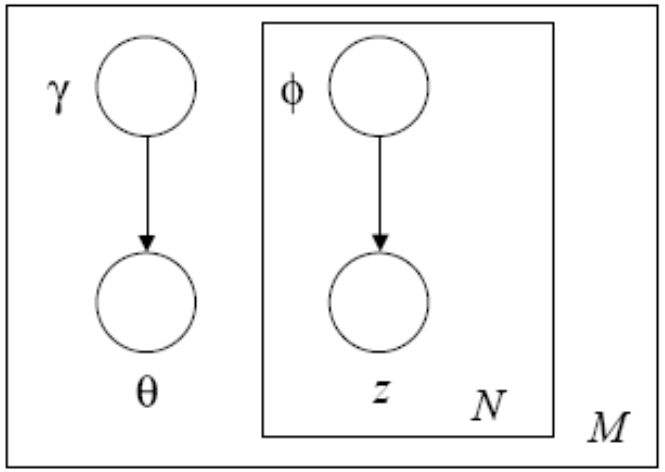
$$p(\boldsymbol{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

- Unfortunately, this distribution is intractable to compute in general. A function which is intractable due to the coupling between θ and β in the summation over latent topics

Inference and parameter estimation

- Drop some edges and the **w** nodes





$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}$$

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n)$$

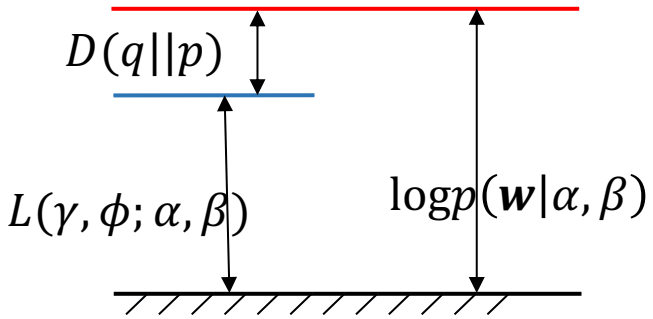
Inference and parameter estimation

- Lower bound on Log-likelihood

$$\log p(\boldsymbol{w}|\alpha,\beta) = \log \int \sum_{z} p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)d\theta = \log \int \sum_{\boldsymbol{z}} q(\theta,\boldsymbol{z}|\gamma,\phi)\frac{p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)}{q(\theta,\boldsymbol{z}|\gamma,\phi)}d\theta$$

$$\geq \int \sum_{\boldsymbol{z}} q(\theta,\boldsymbol{z}|\gamma,\phi)\log\frac{p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)}{q(\theta,\boldsymbol{z}|\gamma,\phi)}d\theta = E_q[\log p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)] - E_q[p(\theta,\boldsymbol{z}|\gamma,\phi)]$$

$$D(q(\theta,\boldsymbol{z}|\gamma,\phi)||p(\theta,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta))$$
$$= \int \sum_z q(\theta,\boldsymbol{z}|\gamma,\phi)\log q(\theta,\boldsymbol{z}|\gamma,\phi)d\theta - \int \sum_z q(\theta,\boldsymbol{z}|\gamma,\phi)\log p(\theta,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta)d\theta$$
$$= \int \sum_z q(\theta,\boldsymbol{z}|\gamma,\phi)\log q(\theta,\boldsymbol{z}|\gamma,\phi)d\theta - \int \sum_z q(\theta,\boldsymbol{z}|\gamma,\phi)\log \frac{p(\theta,\boldsymbol{z},\boldsymbol{w},\alpha,\beta)}{p(\boldsymbol{w},\alpha,\beta)}d\theta$$
$$= E_q[q(\theta,\boldsymbol{z}|\gamma,\phi)] - E_q[p(\theta,\boldsymbol{z},\boldsymbol{w},\alpha,\beta)] + E_q[p(\boldsymbol{w},\alpha,\beta)]$$



KL between variational posteriori and true posteriori

$$\log p(\boldsymbol{w}|\alpha,\beta) = E_q[\log p(\theta,\boldsymbol{z},\boldsymbol{w}|\alpha,\beta)] - E_q[\log q(\theta,\boldsymbol{z}|\gamma,\phi)] + D(q(\theta,\boldsymbol{z}|\gamma,\phi)||p(\theta,\boldsymbol{z}|\boldsymbol{w},\alpha,\beta))$$

$$L(\gamma,\phi;\alpha,\beta)$$

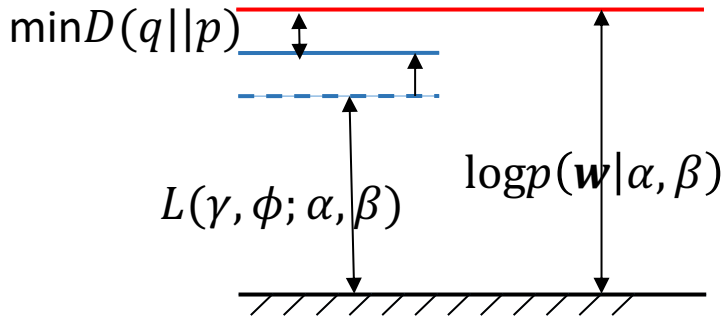
Inference and parameter estimation

- (E-step) For each document, find the optimizing values of the variational parameters $\{\gamma, \phi\}$

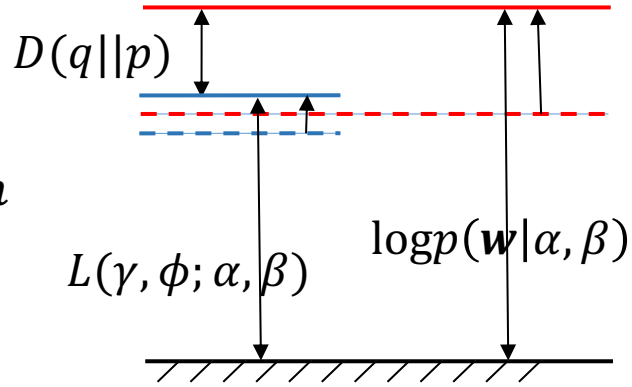
$$(\gamma^*, \phi^*) = \arg min_{(\gamma,\phi)} D(q(\theta, \boldsymbol{z}|\gamma, \phi)||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta))$$

$$\phi_{ni} \propto \beta_{iv} \exp\left(\psi(\gamma_i) - \psi\left(\sum_{j=1}^{k} \gamma_j\right)\right)$$

$$\gamma_i \propto \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$



- (M-step) Maximize the result lower bound on the log likelihood with respect to the model parameters α and β
  - Apply coordinate descent method to $L(\gamma, \phi; \alpha, \beta)$
  - Analytical solution of parameter β, $\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$
  - Parameter α is solved iteratively using linear time Newton method



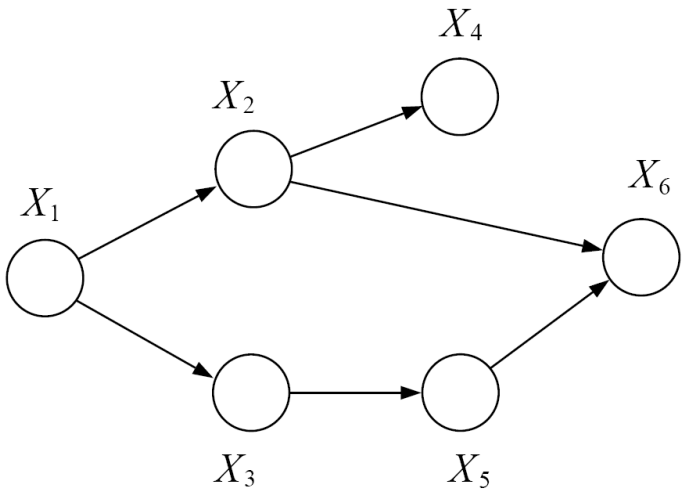
Learning and Inference

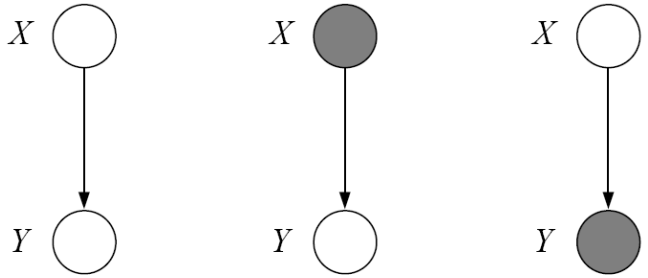
Probabilistic Inference and Learning

- We now have compact representations of probability distributions: **PGMs**
- A PGM M describes a unique probability distribution P
- Typical tasks:
  - Task 1: How do we answer **queries** about P_M, *e.g.*, $P_M(X|Y)$?
    - ➢ We use **inference** as a name for the process of computing answers to such queries
  - Task 2: How do we estimate a **plausible model** M from data D?
    1. We use **learning** as a name for the process of obtaining point estimate of M.
    2. But for *Bayesian*, they seek $p(M|D)$, which is actually an **inference** problem.
    3. When not all variables are observable, even computing point estimate of M need to do **inference** to impute the *missing data*.

Inference

- Possible Queries:





Marginal distribution e.g. $P(X_6)$

Posterior distribution e.g. $P(X_2|X_6 = 1)$

Marginal distribution $P(y) = \sum_x P(y|x)P(x)$

Posterior distribution $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$

Inference methods

- Exact inference:
  - variables elimination
  - belief propagation

  higher computational complex

- Approximate inference
  - sampling
  - variational inference

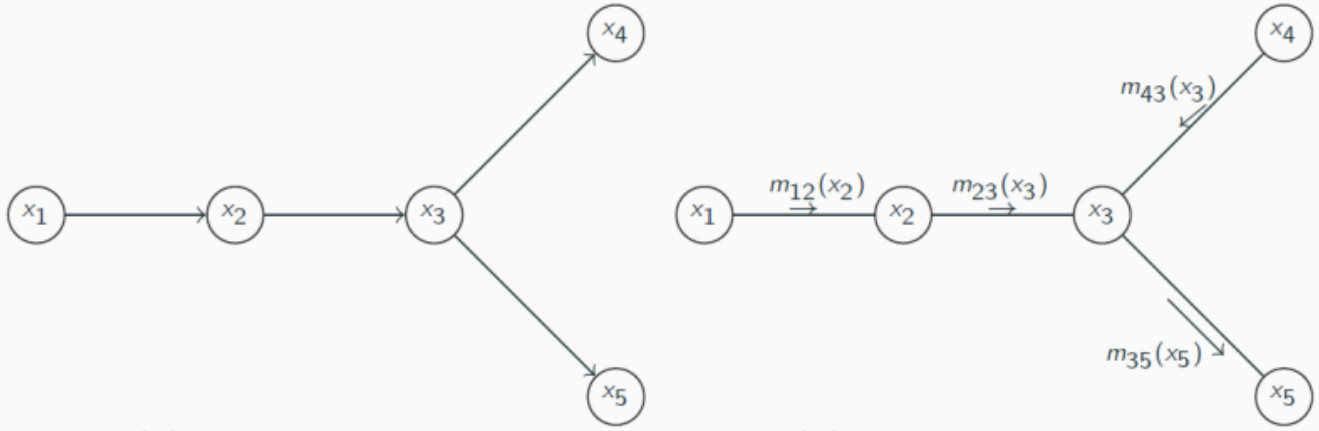
  lower computational complex

Variables Elimination

- Given $P(x_1, \dots, x_5)$, the goal is to calculate $P(x_5)$:

$$P(x_5) = \sum_{x_1, x_2, x_3, x_4} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_3)$$

- Eliminate other variables $\{x_1, \dots, x_4\}$ with an order



(a) Bayesian Network (b) Message Propagation

Figure 9: Variables Elimination and Message Propagation

Variables Elimination

- Eliminate other variables $\{x_1, \dots, x_4\}$ with an order $\{x_1, x_2, x_4, x_3\}$

$$P(x_5) = \sum_{x_1,x_2,x_3,x_4} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_3)$$

$$= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2) \sum_{x_1} P(x_1)P(x_2|x_1)$$

$$= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2)\, m_{12}(x_2)$$

$$= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3)\, m_{23}(x_3)$$

$$= \sum_{x_3} P(x_5|x_3)\, m_{23}(x_3) m_{43}(x_3)$$

$$= m_{35}(x_5)$$

Variables Elimination

- The Sum-Product Algorithm: works for Bayesian network and Markov network

$$m_{i\to j}(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k\in n(i)\backslash \mathrm{j}} m_{ki}(x_i)$$

- There is repeated computation for multiple marginal distributions
  - Blief Propagation: take $m_{i\to j}(x_j)$ as a message that x_i transforms to x_j
  - marginal distribution:

$$P(x_i) \propto \prod_{k\in n(i)} m_{k\to i}(x_i)$$

Blief Propagation

1. Leaf-to-root: assign one root node, transform information from all leaf nodes until root node accepts all neighbor message
2. Root-to-leaf: transform information from root node, until all leaf nodes accept information

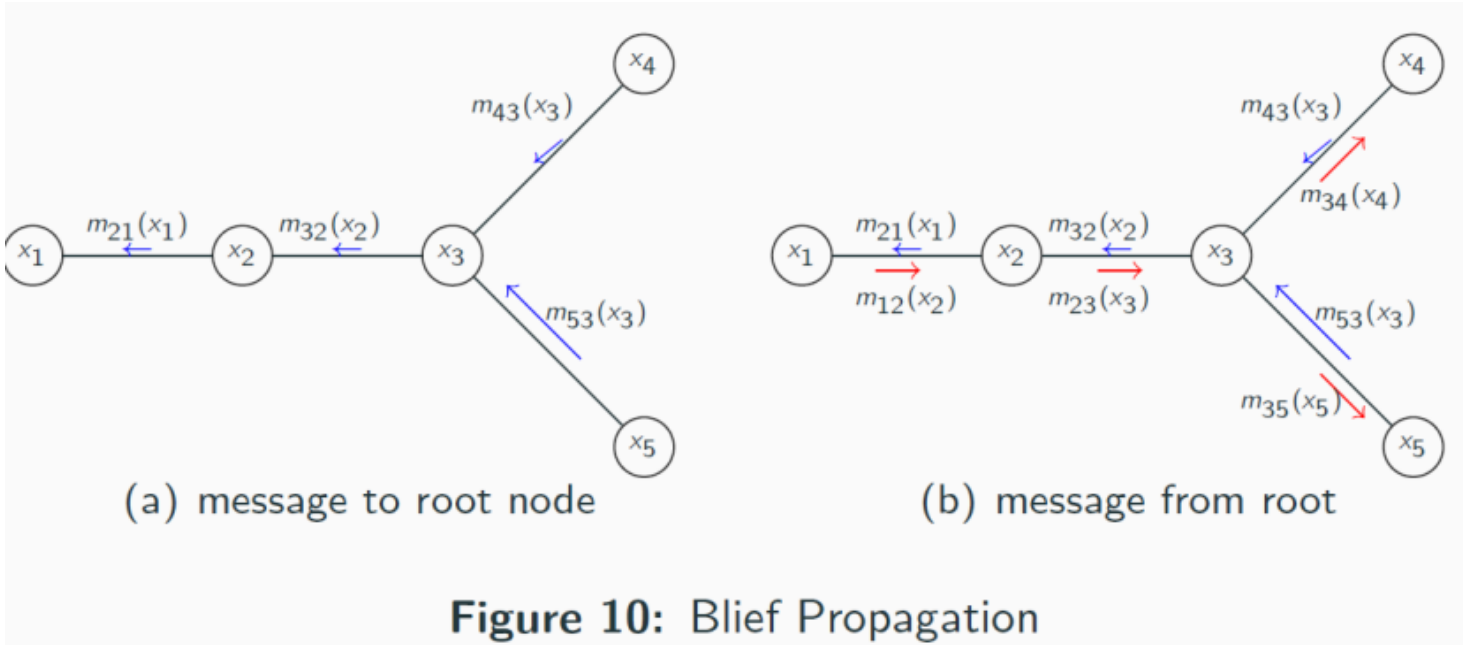


Figure 10: Blief Propagation

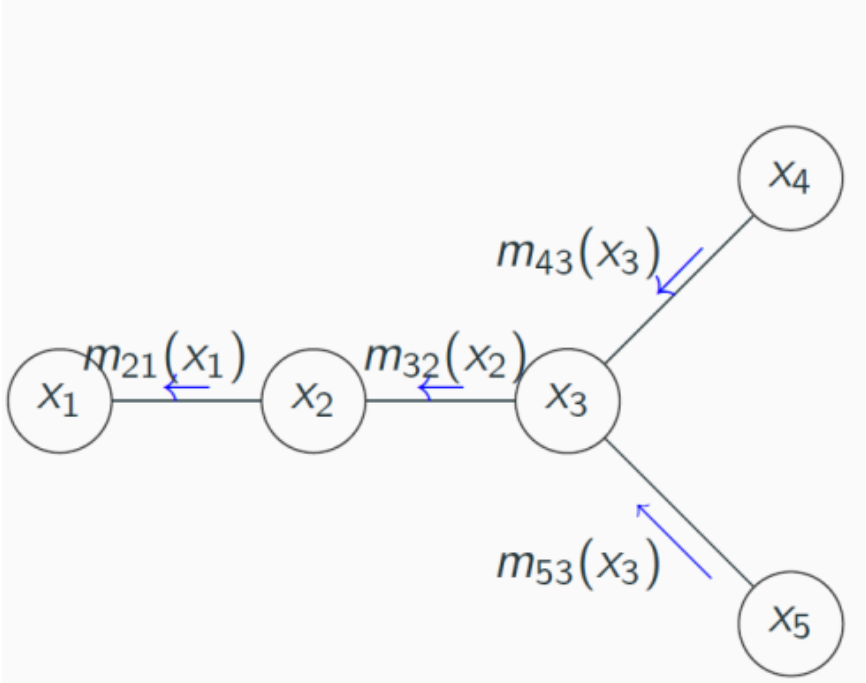
Blief Propagation

- Joint probability

$$P(x_1, x_2, \dots, x_5) = \frac{1}{Z}\psi_{12}(x_1, x_2)\ \psi_{23}(x_2, x_3)\ \psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$$

where, $Z = \sum_x \psi_{12}(x_1, x_2)\ \psi_{23}(x_2, x_3)\ \psi_{34}(x_3, x_4)\psi_{35}(x_3, x_5)$

- Leaf → root:



$$m_{43}(x_3) = \sum_{x_4} \psi(x_4, x_3)$$

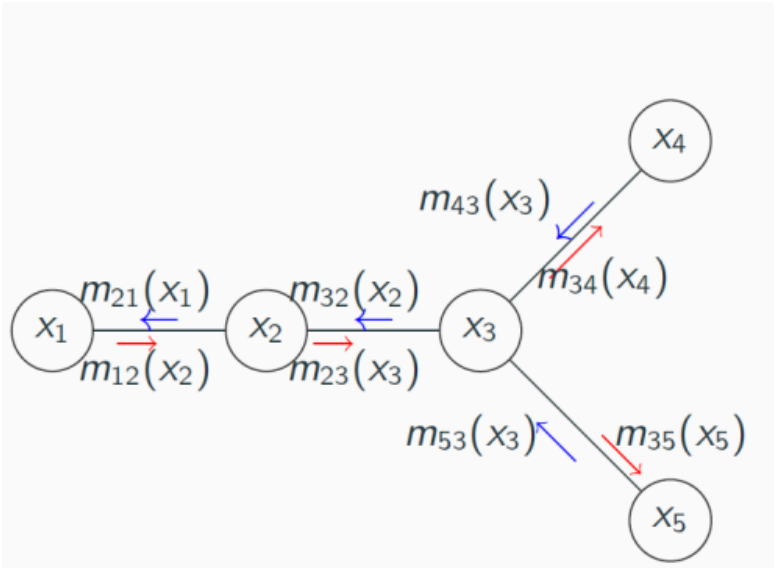
$$m_{53}(x_3) = \sum_{x_5} \psi(x_5, x_3)$$

$$m_{32}(x_2) = \sum_{x_3} \psi(x_3, x_2) m_{43} m_{53}$$

$$m_{21}(x_1) = \sum_{x_2} \psi(x_2, x_1) m_{32}$$

Blief Propagation

- Root → leaf :



$$m_{12}(x_2) = \sum_{x_1} \psi(x_1, x_2)$$

$$m_{23}(x_3) = \sum_{x_2} \psi(x_2, x_3) m_{12}(x_2)$$

$$m_{34}(x_4) = \sum_{x_3} \psi(x_3, x_4) m_{23}(x_3) m_{53}(x_3)$$

$$m_{35}(x_5) = \sum_{x_3} \psi(x_3, x_4) m_{23}(x_3)\, m_{43}(x_3)$$

- Marginal distribution

$$P(x_1) \propto m_{21}(x_1) \qquad P(x_2) \propto m_{12}(x_2) m_{32}(x_2)$$

$$P(x_3) \propto m_{23}(x_3) m_{43}(x_3) m_{53}(x_3)$$

$$P(x_4) \propto m_{34} \qquad P(x_5) \propto m_{35}$$

Learning

- learning parameters for a BN with given structure and is completely observable

$$\ell(\theta; D) = logp(D|\theta) = log\prod_n(\prod_i p(x_{n,i}|x_{n,\pi_i},\theta_i)) = \sum_i(\sum_n logp(x_{n,i}|x_{n,\pi_i},\theta_i))$$

- MLE: Count
- MAP: Add pseudo-counts

- Partially observed: Expectation-Maximization

Learning

- For directed graphical models, the log-likelihood decomposes into a sum of terms, one per family (node plus parents).
- For undirected graphical models, the log-likelihood does not decompose, because the normalization constant Z is a function of **all** the parameters

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \qquad Z = \sum_{x_{1,..,}x_n} \prod_{c \in C} \psi_c(x_c)$$

- In general, we will need to do inference (i.e., marginalization) to learn parameters for undirected models, even in the fully observed case.

Summary

- A PGM is a natural/perfect tool for
  - Representation (数据结构) and
  - Inference (算法).
- Two types of PGM
  - Bayesian Network
  - Markov Random Fields
- Typical PGMs
  - Hidden Markov Model is a directed generative models
    - Three canonical problems
    - Forward-backward algorithm
  - Conditional Random Fields are undirected discriminative models
    - They overcome the label bias problem of MEMMs by using a global normalizer