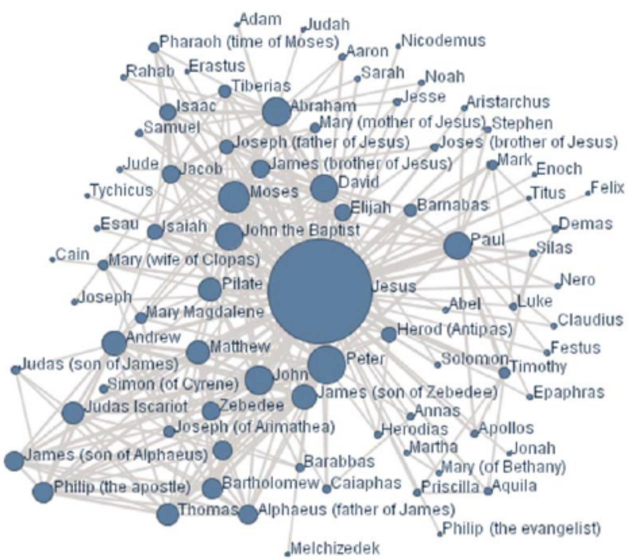
Chapter 10:

Directed Graphical Models (Bayes nets)

张小彬 @CIS Lab

Probabilistic Graph Model

Graph



Model

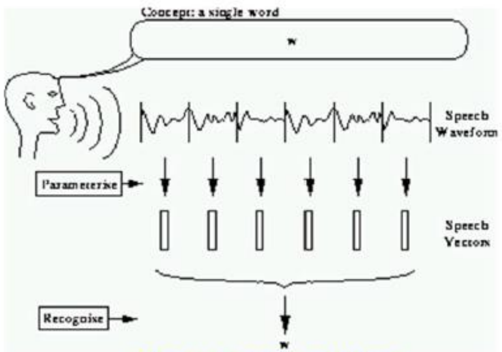
$$\mathcal{M}$$

Probabilistic Theory + Graph Theory!

Data

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

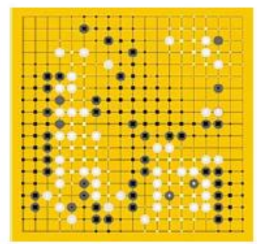
Probabilistic Graph Model



Speech recognition

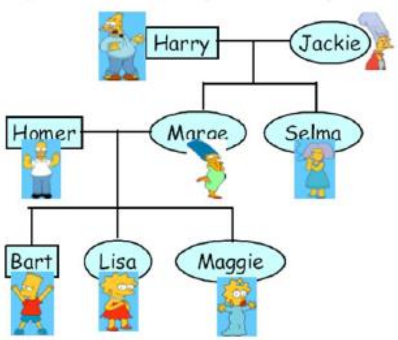
Information retrieval


Computer vision

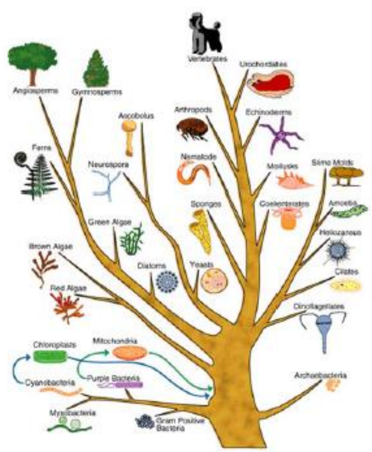

Games


Robotic control


Planning



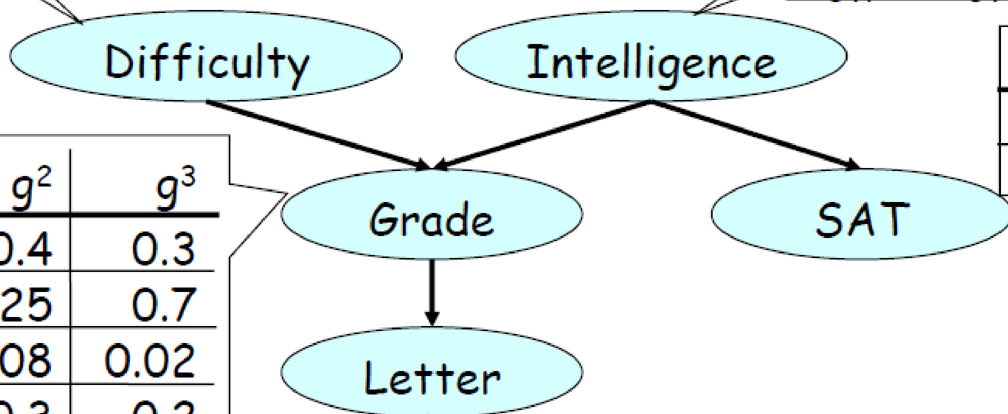
Pedigree


Evolution

The Student Network

d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3



	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Fundamental Questions of PGM

- Representation
  - How to representation a joint distribution?
    - p$(x_1, x_2, \ldots, x_V)$ -> p$(x_{1:V})$
  - Directed Graph Model
    - Bayesian Networks, BNs
  - Undirected Graph Model
    - Markov Random Field, MRF
- Inference
  - Infer Marginal Distribution from Joint Distribution
- Learning
  - Learn parameters and Structures of the PGM

Chain Rule

$$p(x_{1:V}) = p(x_1)p(x_2|x_1)p(x_3|x_2,x_1)p(x_4|x_1,x_2,x_3)\ldots p(x_V|x_{1:V-1})$$

- A direct way to calculate joint distribution
- Language model: Sentence probability
- p(x_1) $O(K)$ parameters
- p$(x_2|x_1)$ $O(K^2)$ parameters
  - stochastic matrix
- p$(x_3|x_1, x_2)$ $O(K^3)$ parameters
  - Conditional probability tables or CPTs
- p$(x_V|x_{1:V-1})$ $O(K^V)$ parameters

Chain Rule (cont.)

- Can we replace CPTs?
- Conditional probability distribution, or CPDs
- $O(K^2V^2)$ parameters, why?
- Each variable depends on all the previous variables

$$p(x_t = k|\mathbf{x}_{1:t-1}) = \mathcal{S}(\mathbf{W}_t\mathbf{x}_{1:t-1})_k$$

Conditional Independence

- Represent large joint distributions
- Conditional independence (CI)

$$X \perp Y|Z \iff p(X,Y|Z) = p(X|Z)p(Y|Z)$$

- Make (first order) Markov assumption

$$x_{t+1} \perp \mathbf{x}_{1:t-1}|x_t$$

- (first order) Markov Chain

$$p(\mathbf{x}_{1:V}) = p(x_1)\prod_{t=1}^{V} p(x_t|x_{t-1})$$

- State transition matrix $p(x_t = j|x_{t-1} = i)$

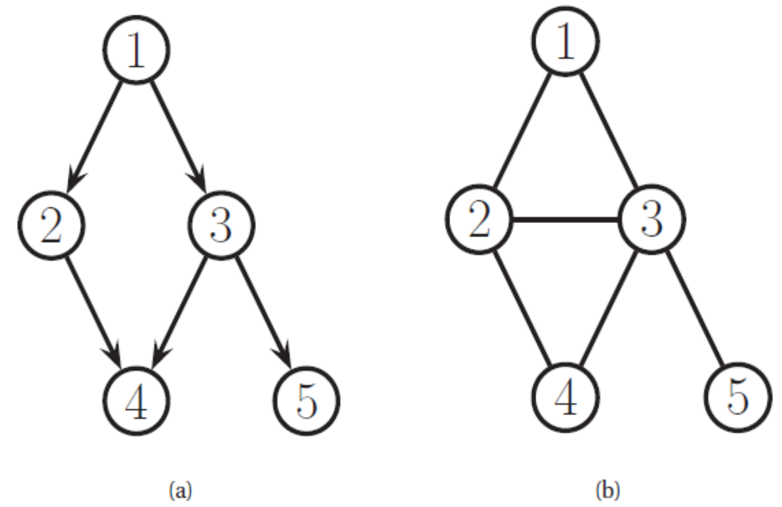
Graphical models

A graphical model

A way to represent a joint distribution
by making CI assumptions.

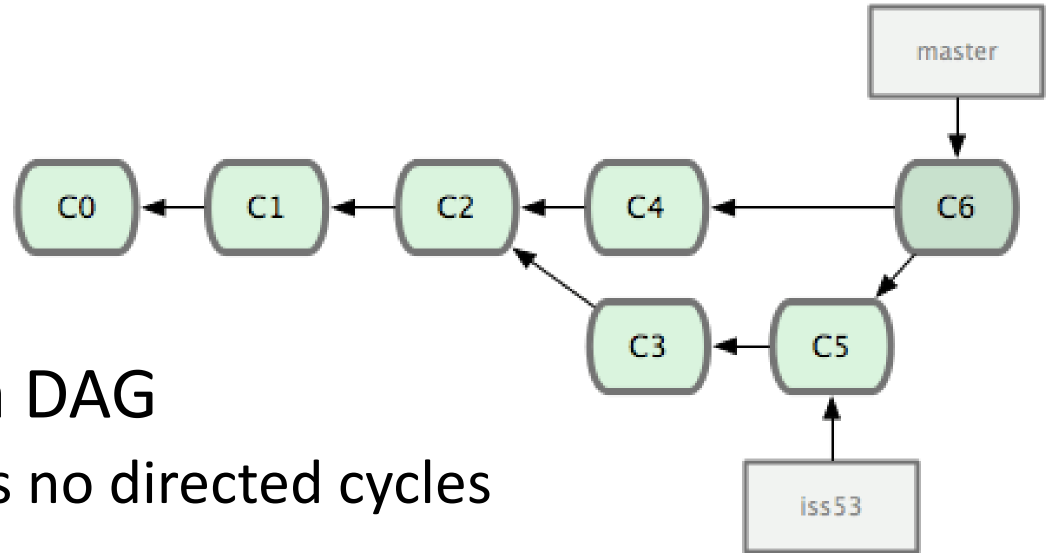
- Study directed graphs in this chapter
- Study undirected graphs in chapter 17

Graph terminology



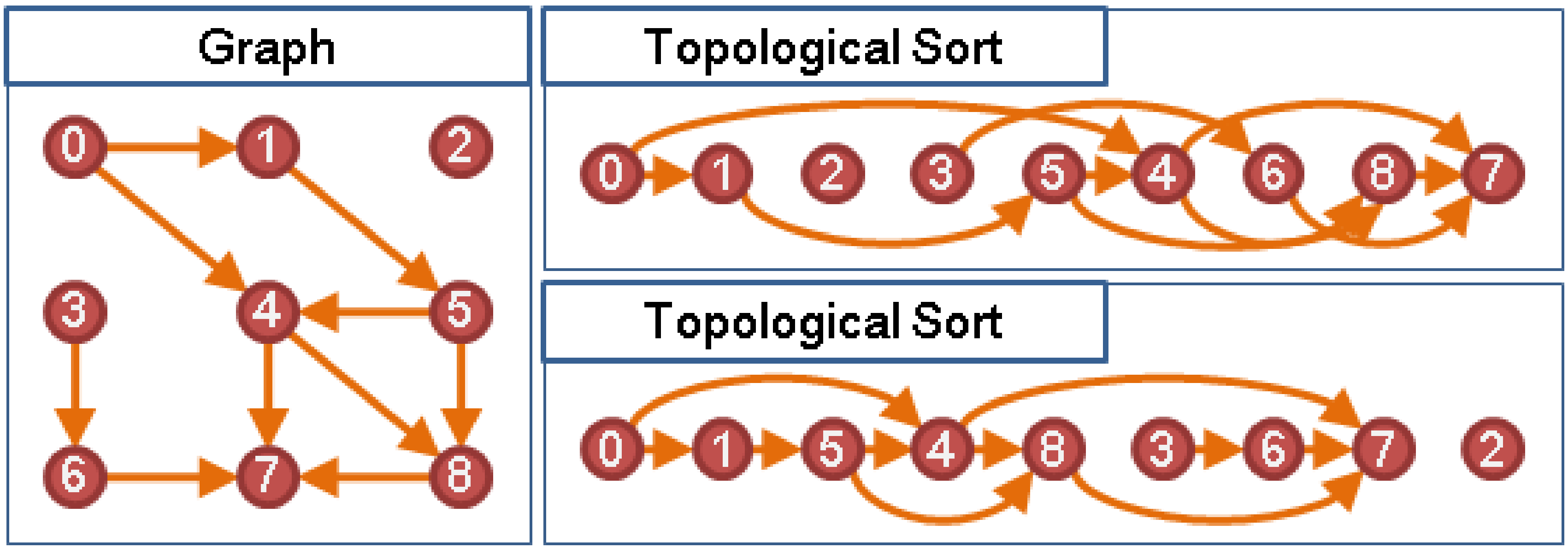
- Graph G
- Parent, Child, Family
- Ancestors, descendants, neighbors
- Degree, in-degree, out-degree, cycle or loop
- DAG, directed acyclic graph
- Topological ordering
- Path or trail
- Subgraph, clique, maximal clique, maximum clique

Tree and DAG



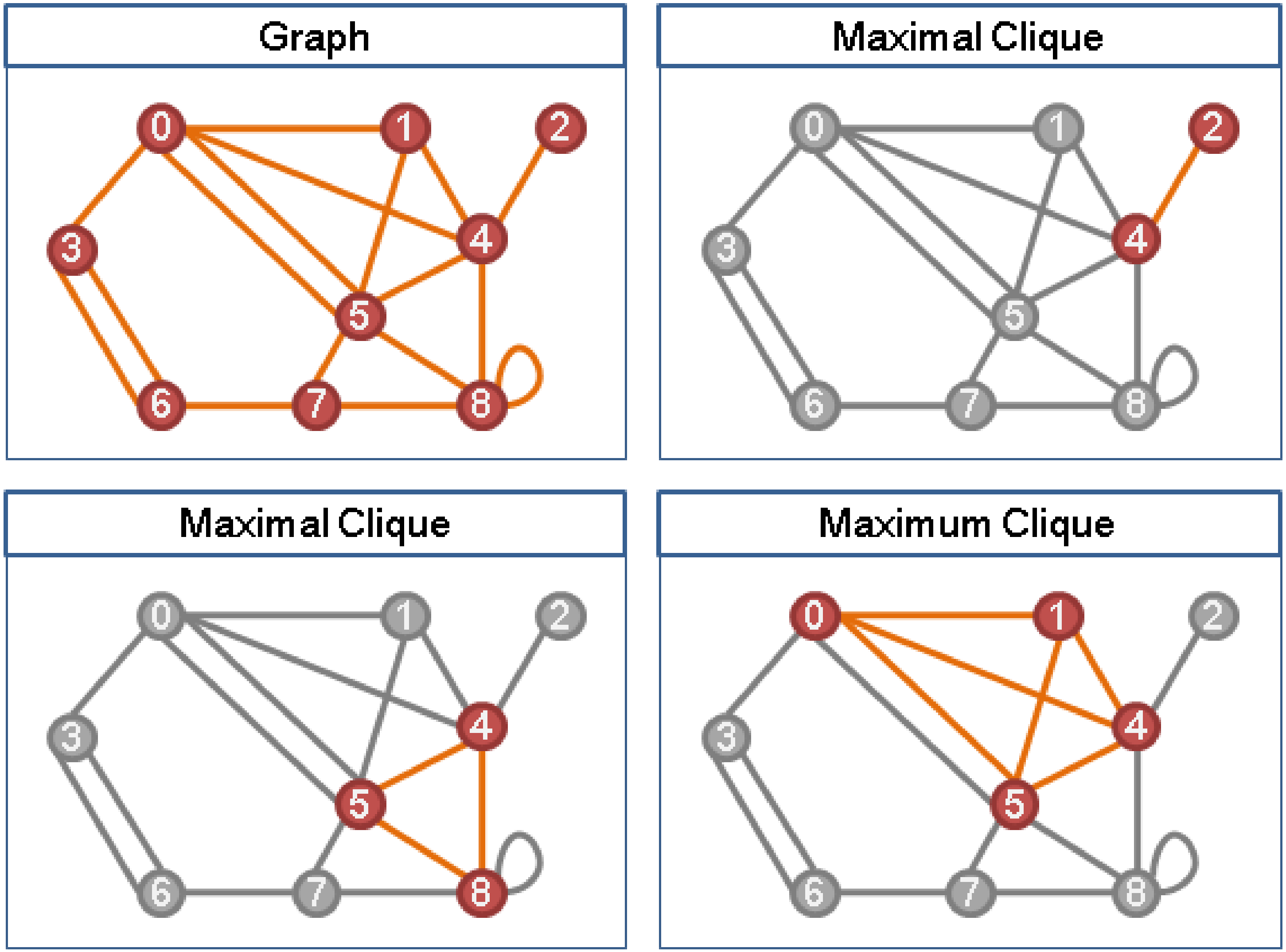
- A Directed Tree is a DAG
  - A directed tree has no directed cycles
- A DAG is not necessarily a tree
- A DAG can have multiply parents
  - Also called poly tree
  - Otherwise called moral directed tree
  - *More than one path between two nodes*
  - *May have loops(cycles) if turned undirected*

Topological ordering



- Only DAG has topological ordering
- Parents have lower numbers than their children
- Figure 10.1 ?

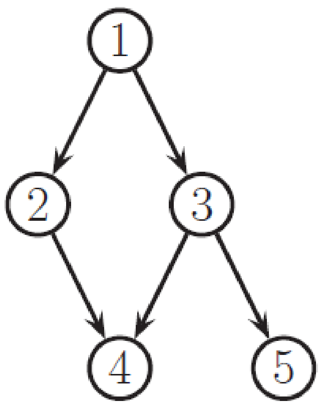
Subgraph & Clique



Directed Graphical Models

- Ordered Markov property

$$x_s \perp \mathbf{x}_{\text{pred}(s) \setminus \text{pa}(s)} | \mathbf{x}_{\text{pa}(s)}$$



- **Bayesian Networks**
- Belief Networks
- Causal Networks

For example, the DAG in Figure 10.1(a) encodes the following joint distribution:

$$\begin{aligned} p(\mathbf{x}_{1:5}) &= p(x_1)p(x_2|x_1)p(x_3|x_1, \cancel{x_2})p(x_4|\cancel{x_1}, x_2, x_3)p(x_5|\cancel{x_1}, \cancel{x_2}, x_3, \cancel{x_4}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3) \end{aligned}$$

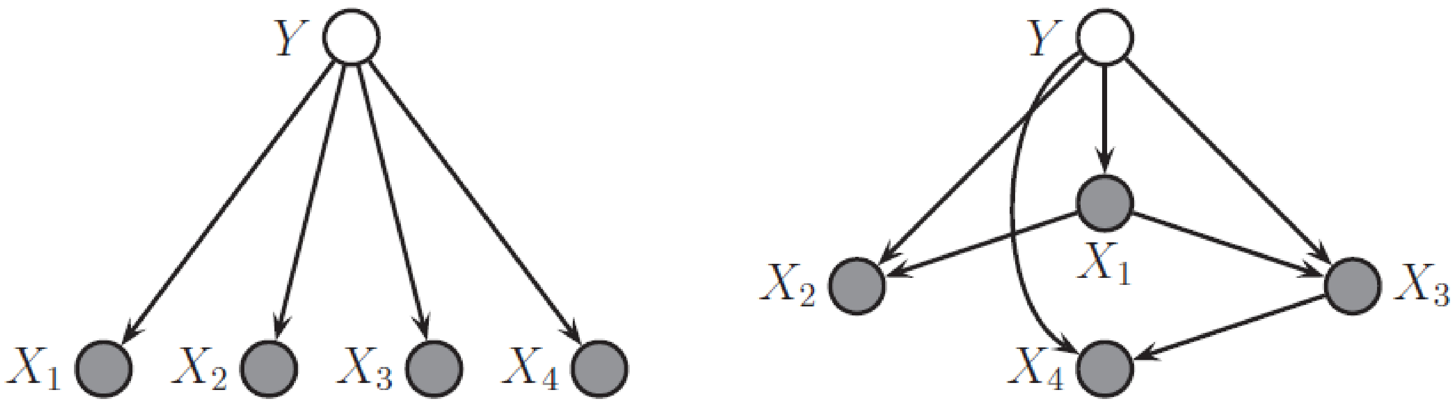
Directed Graphical Models (cont.)

- In general, we have

$$p(\mathbf{x}_{1:V}|G) = \prod_{t=1}^{V} p(x_t|\mathbf{x}_{\text{pa}(t)})$$

- This equation holds only if
  - the CI assumptions encoded in DAG in G are correct
- If each node has O(F) parents and K states
  - Model has $O(VK^F)$ parameters

Example: Naïve Bayes classifiers



(a) (b)

$$p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j|y)$$

- NBC assumes the features
  - are conditionally independent given class labels
- Tree-augmented naïve Bayes classifier
  - Find the optimal tree structure using the Chow-Liu algorithm

Example: Markov Chain

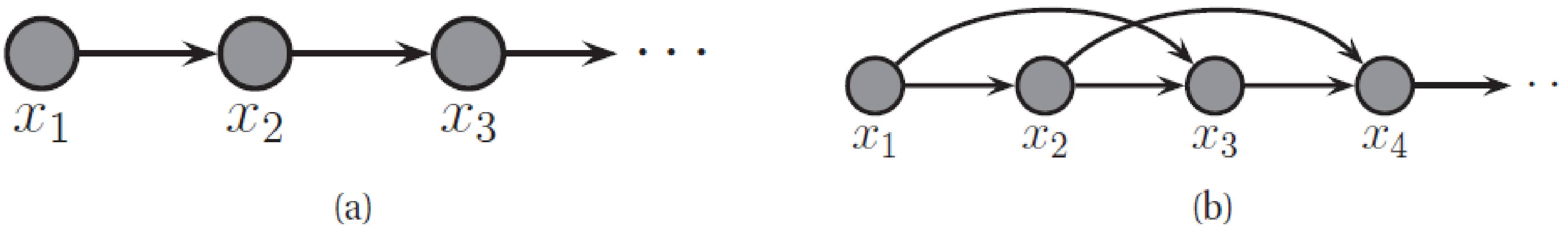


Figure 10.3 A first and second order Markov chain.

- Second order Markov chain:

$$p(\mathbf{x}_{1:T}) = p(x_1, x_2)p(x_3|x_1, x_2)p(x_4|x_2, x_3) \ldots = p(x_1, x_2) \prod_{t=3}^{T} p(x_t|x_{t-1}, x_{t-2})$$

Example: Hidden Markov Model

- Hidden variable z_t
- Observed variable x_t
- $p(z_t|x_t)$?

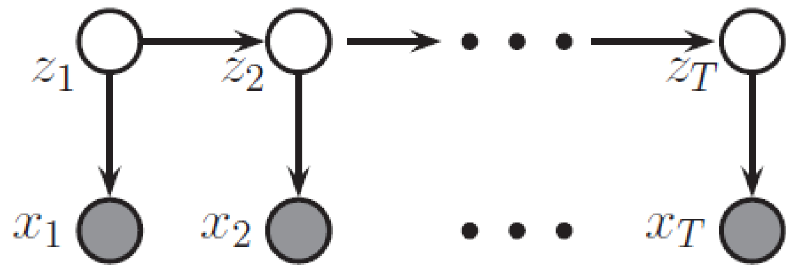


Figure 10.4 A first-order HMM.

- **齐次马尔可夫假设**
  - 当前的隐变量只和前一个隐变量有关，得到转移模型（transition model）
  - 即 $p(z_t|z_{t-1}) = p(z_t|z_{1:T}, \mathbf{x}_{1:T})$，可以对 CPD $p(z_t|z_{t-1})$ 进行建模
- **观测独立性假设**
  - 当前的观察变量只和当前的隐变量有关，得到观察模型（observation model）
  - 即 $p(\mathbf{x}_t|z_t) = p(\mathbf{x}_t|z_{1:T}, \mathbf{x}_{1:T})$，可以对 CPD $p(\mathbf{x}_t|z_t)$ 进行建模

Example: HMM (cont.)

- Dynamic Bayesian Network
- MRF, CRF, RNN ?

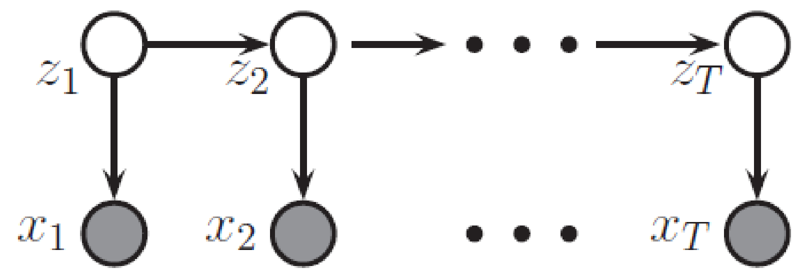


Figure 10.4 A first-order HMM.

- Part of speech tagging
  - x_t represent a word
  - z_t is part of speech
- Automatic speech recognition
  - x_t speech signal features, z_t is the word
  - $p(z_t|z_{t-1})$ is language model
  - $p(x_t|z_t)$ is acoustic model

Inference

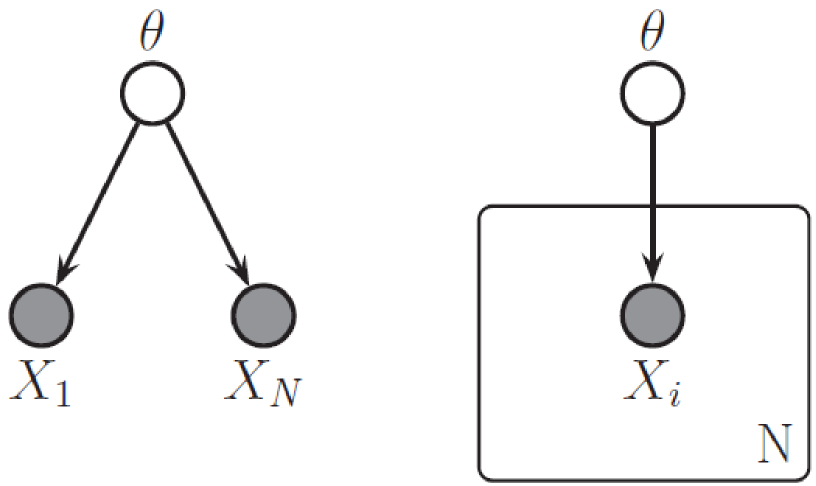
- Joint distribution $p(x_{1:V}|\theta)$
  - visible variables x_v
  - hidden variables x_h
- Infer the unknowns

$$p(\mathbf{x}_h|\mathbf{x}_v, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{p(\mathbf{x}_v|\boldsymbol{\theta})} = \frac{p(\mathbf{x}_h, \mathbf{x}_v|\boldsymbol{\theta})}{\sum_{\mathbf{x}'_h} p(\mathbf{x}'_h, \mathbf{x}_v|\boldsymbol{\theta})}$$

- If x_h is x_q and x_n, how to get x_q?
  - marginalizing out!!

$$p(\mathbf{x}_q|\mathbf{x}_v, \boldsymbol{\theta}) = \sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_n|\mathbf{x}_v, \boldsymbol{\theta})$$

Plate notation



- A form of syntactic sugar called plates
    - draw a box around the repeated variables
- How to represent iid ?

Plate notation of NBC

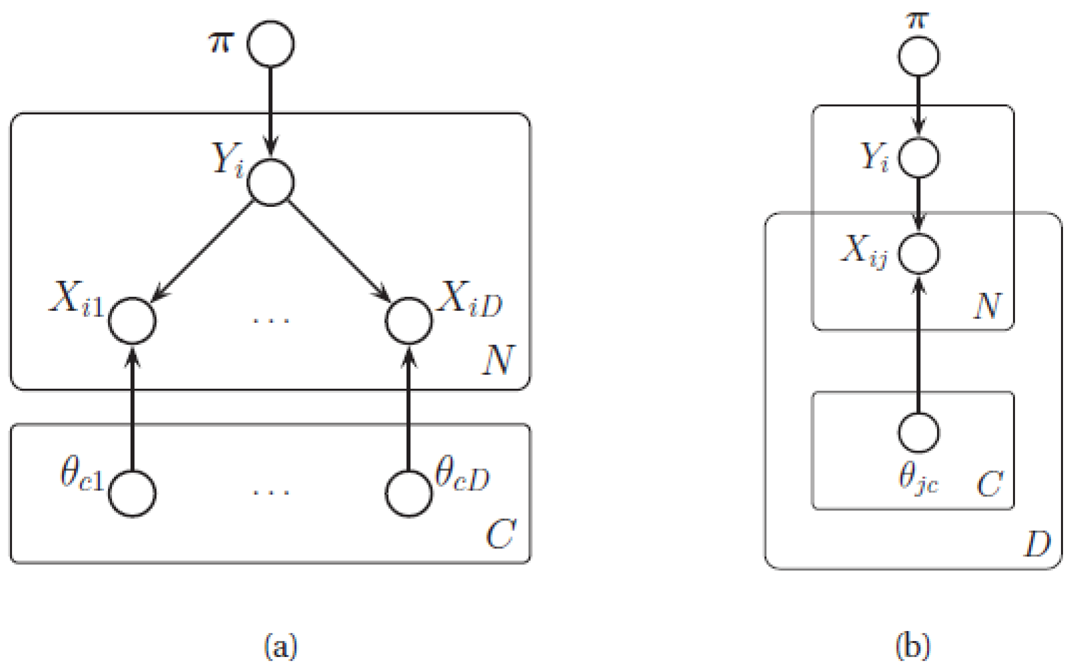
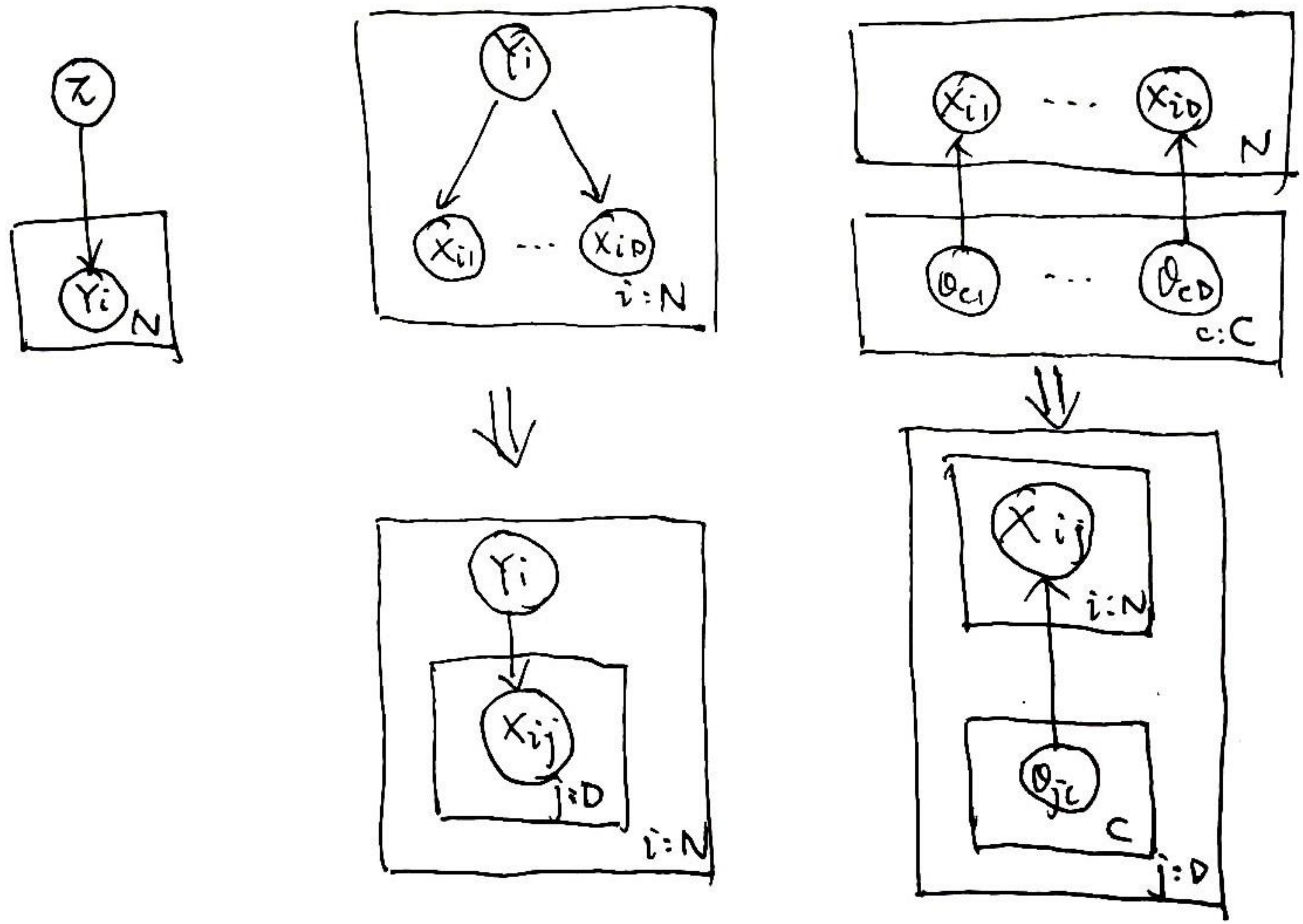


Figure 10.8 Naive Bayes classifier as a DGM. (a) With single plates. (b) WIth nested plates.

Plate notation of NBC (cont.)



Learning

- Just Regular MAP

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\mathbf{x}_{i,v} | \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- If we use uniform prior, MAP -> MLE

- From Bayesian view
  - Parameters are also unknown variables
  - No difference between Inference and Learning

Learning from complete data

- Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{t=1}^{V} p(x_{it}|\mathbf{x}_{i,\mathrm{pa}(t)}, \boldsymbol{\theta}_t) = \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t)$$

- If prior factorizes,

$$p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\boldsymbol{\theta}_t)$$

- Posterior also factorizes.

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_{t=1}^{V} p(\mathcal{D}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t)$$

Learning with missing and/or latent variables

- What is missing data?
  - Consider an image with occluder
  - A broken sensor
  - Sparse matrix, like user dictionary
- Likelihood is no longer convex
- ML or MAP estimate is locally optimal
- How to deal with missing and/or latent variables?
  - EM, Expectation-Maximum Algorithm
  - Structure-EM Algorithm

CI Properties of DGMs & I-map

- Consider the independence of any pair in DGMs.
- CI assumptions in graphical model is like this,

$$\mathbf{x}_A \perp_G \mathbf{x}_B | \mathbf{x}_C$$

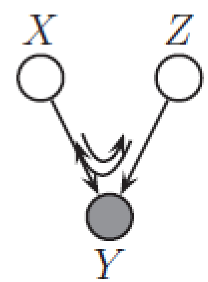
- $I(G)$ is a set
  - All CI statements encoded by the graph G
- $I(p)$ is a set
  - All CI statements encoded by distribution p
- $I(G) \subseteq I(p)$ iff
  - G is an I-map (independent map) for p
  - P is Markov wrt G
- Full connected graph; minimal I-map ?

Active Trail

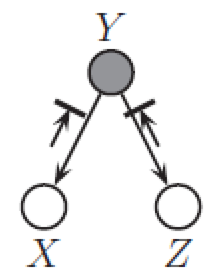
A trial (path) $X_1 - \cdots - X_n$ is **active** if

It has no v-structures $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$

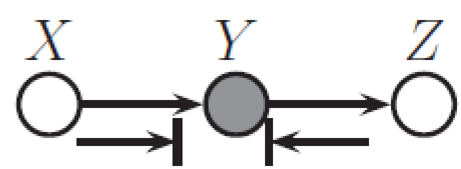
- A trial $X - Y - Z$ is active?
- A trial $X - Y - Z$ is active given Y ?



(c)



(b)



(a)

D-separation

Undirected path P is **d-separated** by a set of nodes E(containing the evidence) iff at least

1. P contains a chain

$$s \rightarrow m \rightarrow t \text{ or } s \leftarrow m \leftarrow t, \text{ where } m \in E$$

2. P contains a tent of fork

$$s \swarrow^{m} \searrow t, \text{ where } m \in E$$

3. P contains a v-structures

$$s \searrow_{m} \swarrow t, \text{ where } m \text{ is not in } E \text{ and nor is any descendant of } m.$$

Global Markov properties (G)

- A set of nodes A, B and third observed set E
- We say A is d-separate from B given E, iff
  - Every node $a \in A, b \in B$ is separated given E

- Define CI properties for BNs

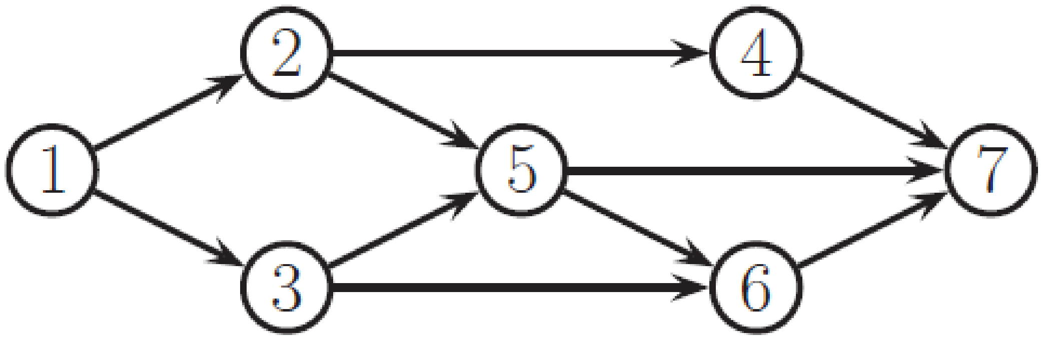
$$\mathbf{x}_A \perp_G \mathbf{x}_B | \mathbf{x}_E \iff \text{A is d-separated from B given E}$$

- This is called directed Global Markov property (G)

Global Markov properties (cont.)

- $X \rightarrow Y \rightarrow Z$
  - $p(x,y,z) = p(x)p(y|x)p(z|y)$
  - $x \perp z|y$
- $X \leftarrow Y \rightarrow Z$
  - $p(x,y,z) = p(y)p(x|y)p(z|y)$
  - $x \perp z|y$
- $X \rightarrow Y \leftarrow Z$
  - $p(x,y,z) = p(x)p(z)p(y|x,z)$
  - $x \not\perp z|y \quad \text{but} \quad x \perp z$

Example



- $x_2 \perp x_6 | x_5$, since $2 \rightarrow 5 \rightarrow 6$ is blocked by x_5 (observed)
- $2 \rightarrow 4 \rightarrow 7 \rightarrow 6$ is blocked by x_7
- $2 \rightarrow 1 \rightarrow 3 \rightarrow 6$ is blocked by x_1
- Is $x_2 \perp x_6 | x_5, x_7$? No, if x_7 is observed

More Markov Property

- Directed Local Markov Property (L)

$$t \perp nd(t)\backslash \mathrm{pa}(t) \mid pa(t)$$

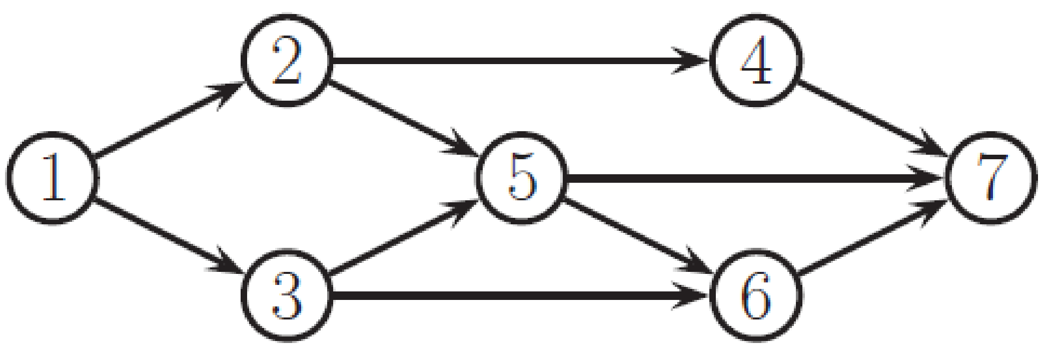
  - nd(t) means non-descendants of t
  - pa(t) means parents of t
- Ordered Markov Property (O)

$$t \perp pred(t)\backslash \mathrm{pa}(t) \mid pa(t)$$

  - pred(t) means predcessors of t
- Three Markov properties for DAGs
  - $G \Leftrightarrow L \Leftrightarrow O$

Markov blanket and full conditionals

- A Markov blanket of node t is
  - $mb(t) \triangleq ch(t) \cup pa(t) \cup copa(t)$
    - copa(t) means co-parents, have the same child
  - Given Markov blanket, t will be CI with all other nodes in the Graph.



Influence (decision) diagrams*

decision nodes	rectangles
utility nodes	diamonds
chance nodes	ovals