Speech and Language Processing

Lecture 2
Graphical model

Information and Communications Engineering Course

Takahiro Shinozaki
Manabu Okumura

2024/9/30

Graphical Models

Integrations of probability and graph theories

Role of Graphical Model

- Visualize the structure of probabilistic models and algorithms
- Describe algorithms in terms of graphical manipulations

Conditional Probability

- Conditional probabilities are obtained from marginal/joint probabilities and the product rule

$$P(A, B \mid C) = \frac{P(A,B,C)}{\sum_{A,B} P(A,B,C)}$$

$$P(A \mid C) = \frac{\sum_{B} P(A,B,C)}{\sum_{A,B} P(A,B,C)}$$

Conditional Independence

Let A, B, and C be disjoint sets of random variables. When the following equation holds, we say that A is independent of B given C, and denote it as $A \perp\!\!\!\perp B | C$

$$P(A | B, C) = P(A | C)$$

$$A \perp\!\!\!\perp B | C$$

$A \perp\!\!\!\perp B | C \Rightarrow P(A, B | C) = P(A | C) P(B | C)$

$\because P(A, B | C) = P(A | B, C) P(B | C)$

$$A \perp\!\!\!\perp B | C \Leftrightarrow B \perp\!\!\!\perp A | C$$

Decomposition of Joint Probability

- By the product rule, arbitrary joint probability is decomposed to a product of conditional probabilities

$$P(A,B,C,D) = P(A)P(B \mid A)P(C \mid A,B)P(D \mid A,B,C)$$

- The contexts might be truncated if there exist conditional independence

e.g. $P(A,B,C,D) = P(A)P(B|A)P(C|A)P(D|C)$

Composition of Joint Probability

- A joint probability is defined by a product of conditional probabilities where the contexts consist of variables whose probabilities are appearing in the left-hand side of them

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

- Proof

The product has the following form

$$\prod_{i=1}^{N} P(X_i|C_i)\,, \quad C_i \subseteq \{X_1, X_2, \cdots, X_{i-1}\}$$

X_i does not appear to the conditional part of X_1 ,…, $X_{i\text{-}1}$. Therefore, by thinking the summation of the following order, we have:

$$\sum_{X_1}\sum_{X_2}\cdots\sum_{X_N}\prod_{i=1}^{N} P(X_i|C_i) = \sum_{X_1}\cdots\sum_{X_{N-1}} P(X_{N-1}|C_{N-1})\sum_{X_N} P(X_N|C_N) = 1$$

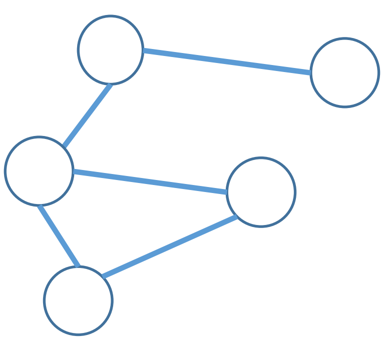
(It is non-negative and satisfies the sum-to-one constraint)

Basic Graph Theories

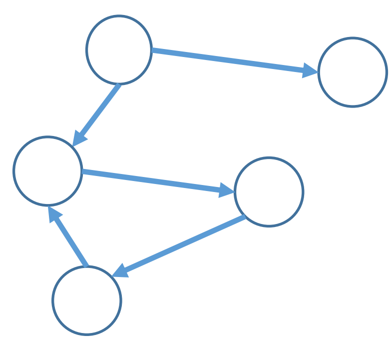
Graphs

- Undirected graph
  - A graph defined by nodes and undirected arcs
- Directed graph
  - A graph defined by nodes and directed arcs
- Directed Acyclic Graph: DAG
  - Directed graph that does not contain a directed cycle

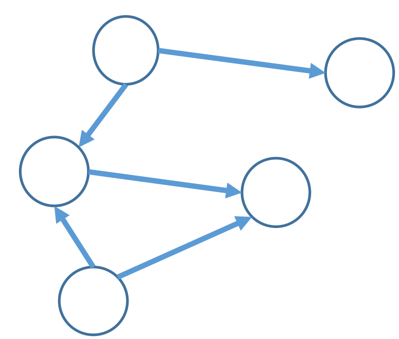
Examples:



Undirected graph



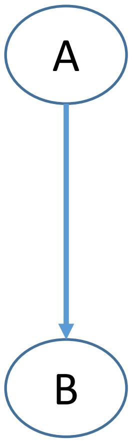
Directed graph
(Have a directed cycle)



Directed acyclic graph

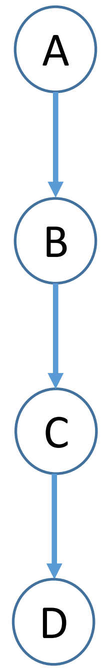
Parent, Child, Ancestor, Descendant

Node A is a
parent of node B



Node B is a
child of node A

Node A, B, and C are
ancestors of node D



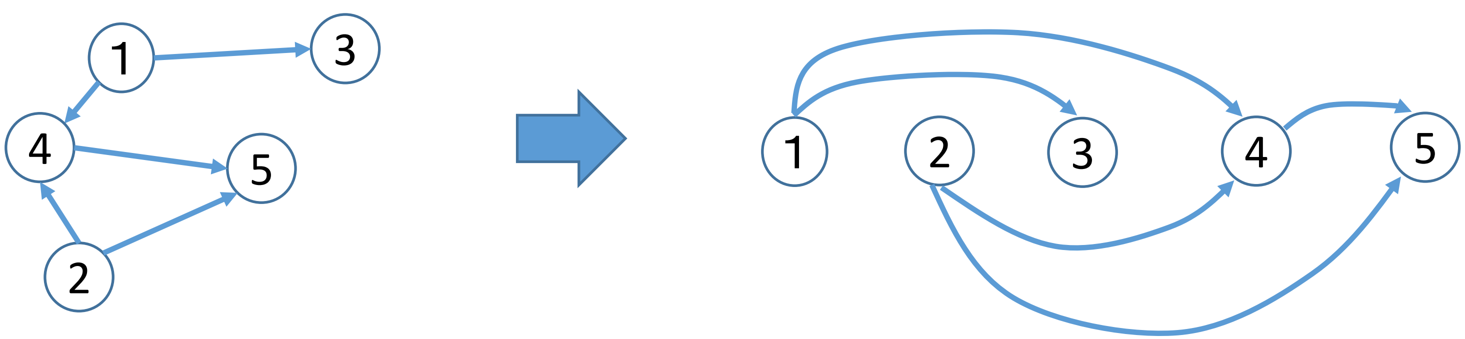
Node B, C, and D are
descendant of node A

Directed Graph and Node Ordering

A directed graph is a DAG

⇕ Eq

There is an ordering of nodes where all arcs face the same direction (=There is a numbering of nodes where all arcs go from a lower numbered to higher numbered nodes)



Outline of the Proof

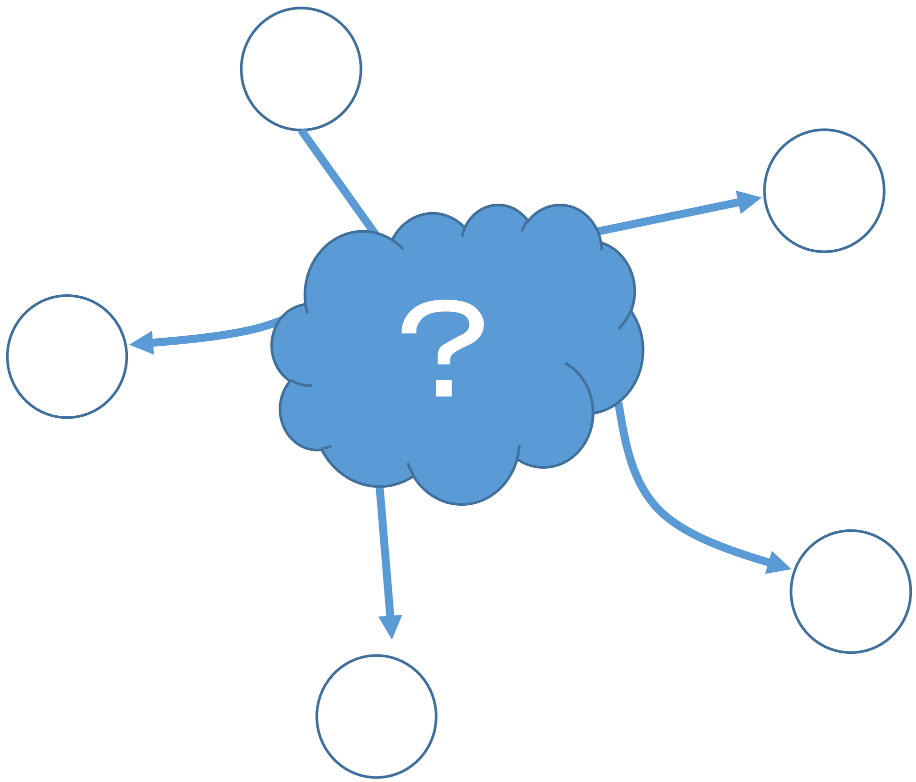
- Statement A:
  - There is an ordering of nodes where all the arcs face the same direction
- Statement B:
  - A graph does not contain a directed cycle

A ⇒ B ...Trivial

B ⇒ A ...Lemma 2.1

Lemma 2.1

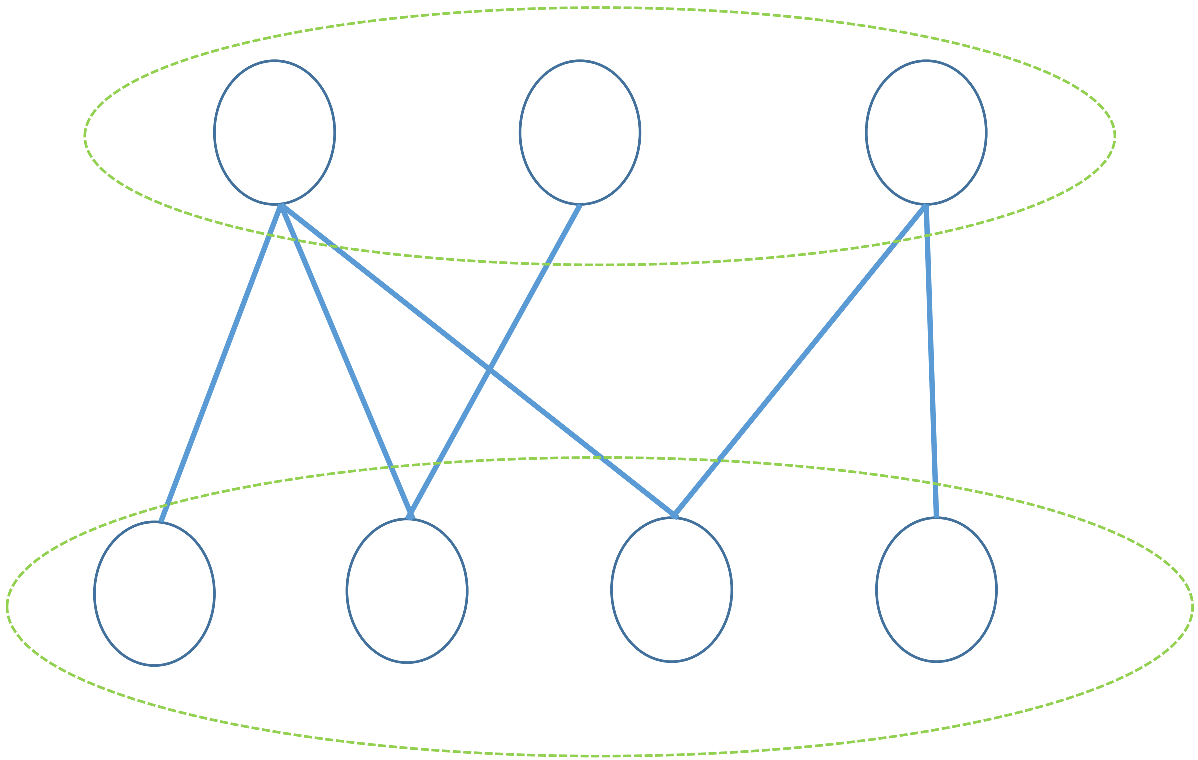
- If a graph does not contain a directed cycle, then there exist at least one node that has no incoming arc



Bipartite

When nodes of a graph are separated to two groups and there is no arc inside the groups, it is called a bipartite

Example of Bipartite:



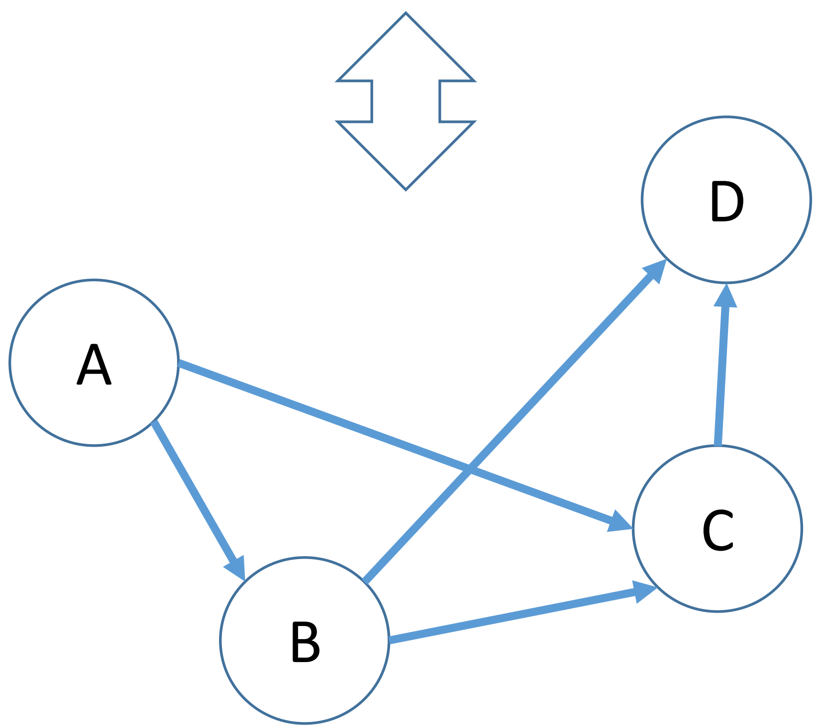
Bayesian Network (BN)

A joint probability model based on DAG and conditional probabilities

Definition of BN

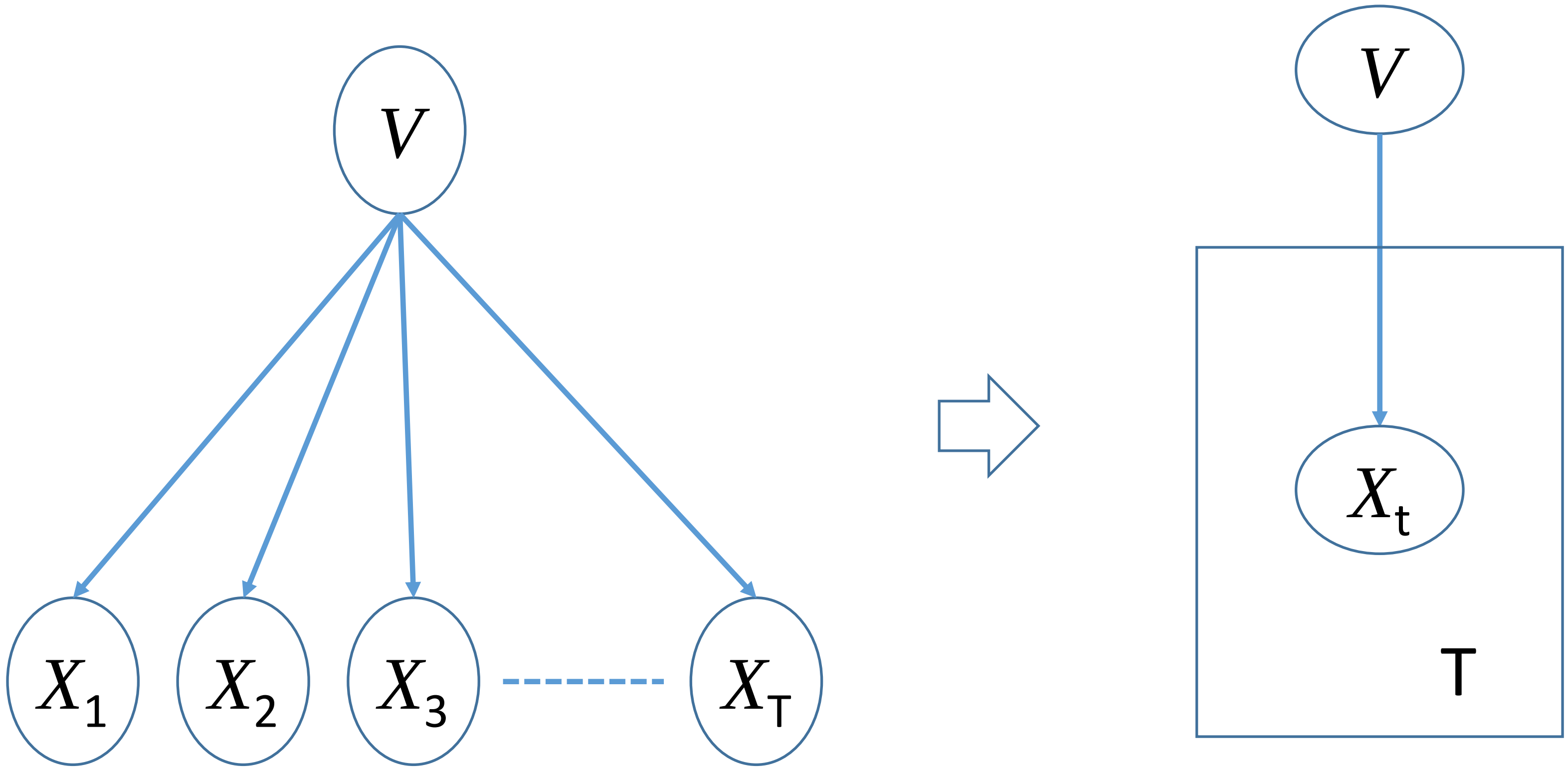
BN is a graphical model where a node represent a random variable and arcs represent dependency of the variables

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



Representation of Repeated Structure

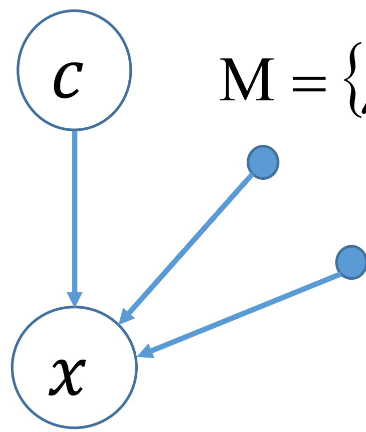
A shorthand notation for a repeated structure is to surrounding the repeating unit and associating the number of repetition



Explicit Representation of Parameters

Small circles represent parameters

Example:
BN representation of a GMM with an explicit description of its parameters



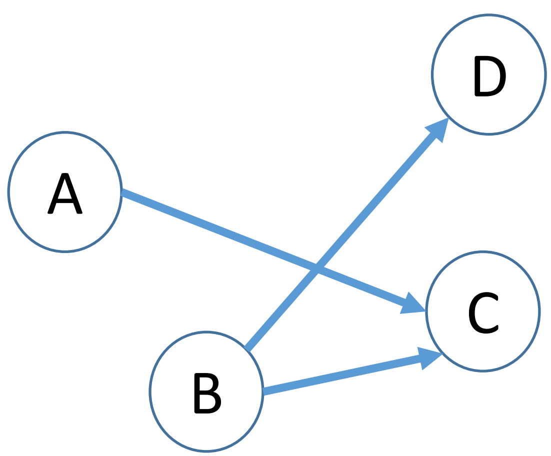
$$\mathrm{M} = \{\mu_1, \mu_2, \mu_3 \cdots \mu_N\}$$

$$S = \{\sigma_1, \sigma_2, \sigma_3 \cdots \sigma_N\}$$

$$P(x) = GMM(x) = \sum_{c=1}^{N} \underline{P(c)N(x;\ \mu_c, \sigma_c^2)}$$

Graph Structure and Conditional Independence

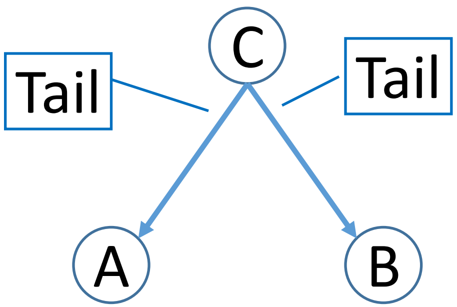
- By investigating the graph structure, we can read relationships between random variables



$A \perp B \mid C$?

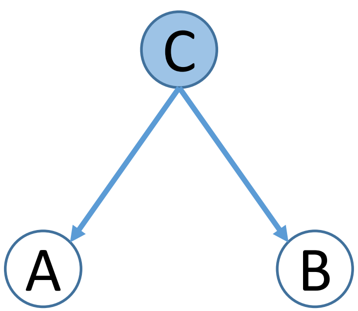
$A \perp D \mid C,B$?

Tail-To-Tail



In general, P(A,B) is not expressed as P(A)P(B). Therefor, $A \perp\!\!\!\perp B \mid \Phi$ does not hold. (Φ is an empty set)

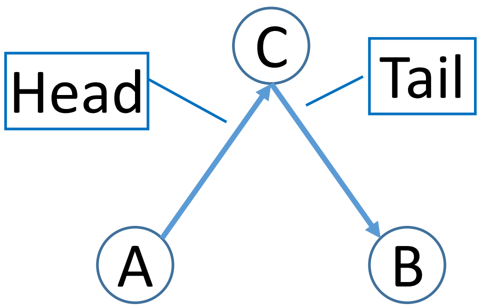
$$P(A,B)=\sum_{C}P(A,B,C)=\sum_{C}P(A\mid C)P(B\mid C)P(C)$$



P(A,B|C) is expressed as P(A|C)P(B|C). Therefore $A \perp\!\!\!\perp B \mid C$ holds.

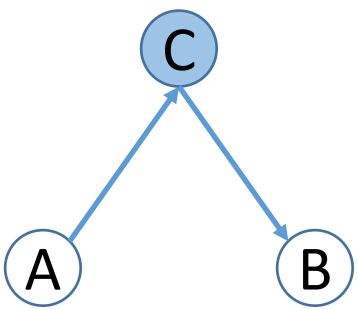
$$P(A,B\mid C)=\frac{P(A,B,C)}{P(C)}=\frac{P(A\mid C)P(B\mid C)P(C)}{P(C)}$$
$$=P(A\mid C)P(B\mid C)$$

Head-To-Tail



In general, P(A,B) is not expressed as P(A)P(B). Therefor, $A \perp\!\!\!\perp B \mid \Phi$ does not hold.

$$P(A,B)=\sum_{C} P(A,B,C)=\sum_{C} P(A)P(B \mid C)P(C \mid A)$$

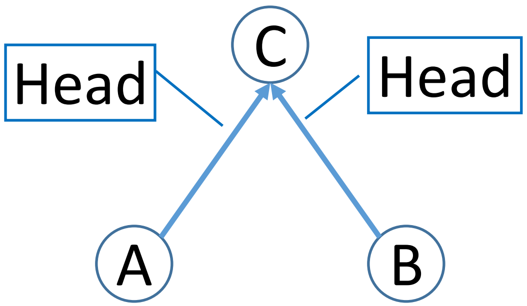


P(A,B|C) is expressed as P(A|C)P(B|C). Therefore $A \perp\!\!\!\perp B \mid C$ holds.

$$P(A,B \mid C)=\frac{P(A,B,C)}{P(C)}=\frac{(P(C \mid A)P(A))P(B \mid C)}{P(C)}$$

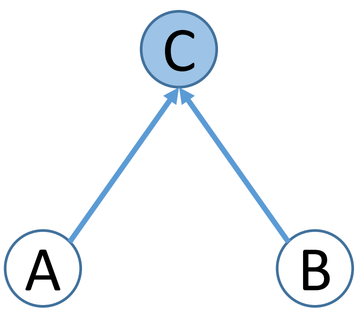
$$=P(A \mid C)P(B \mid C)$$

Head-To-Head



In general, P(A,B) is expressed as P(A)P(B).
Therefor, $A \perp B \mid \Phi$ holds.

$$P(A,B)=\sum_{C}P(A,B,C)=\sum_{C}P(A)P(B)P(C\mid A,B)=P(A)P(B)$$

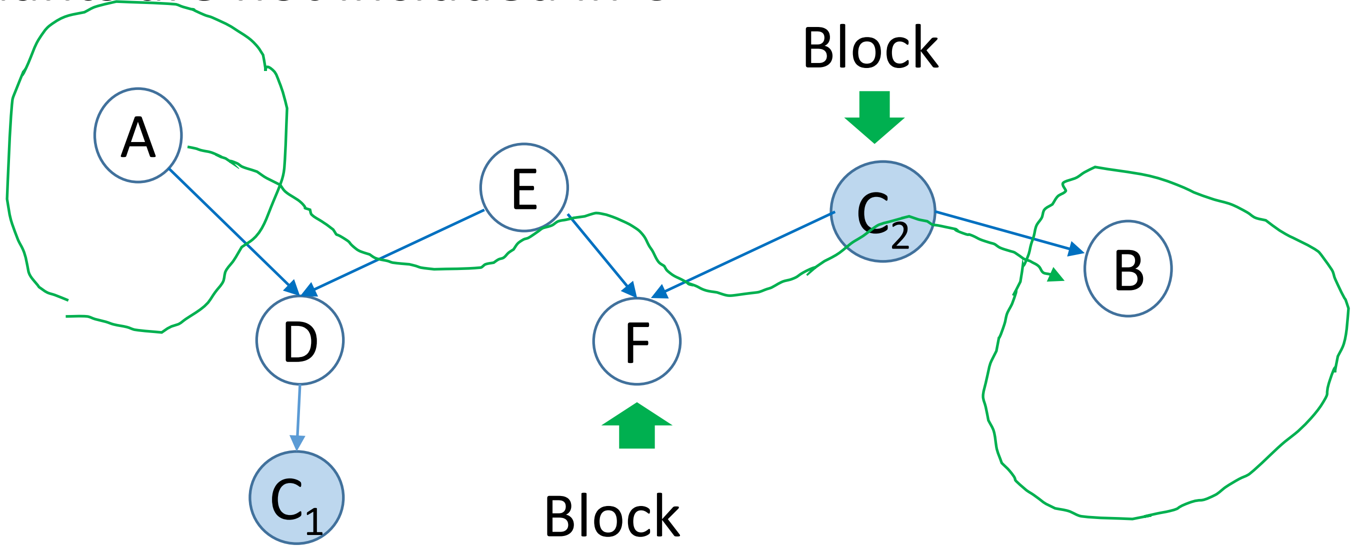


P(A,B|C) is not expressed as P(A|C)P(B|C).
Therefore $A \perp B \mid C$ does not hold.

$$P(A,B\mid C)=\frac{P(A,B,C)}{P(C)}=\frac{P(A)P(B)P(C\mid A,B)}{P(C)}$$

Blocking a Path

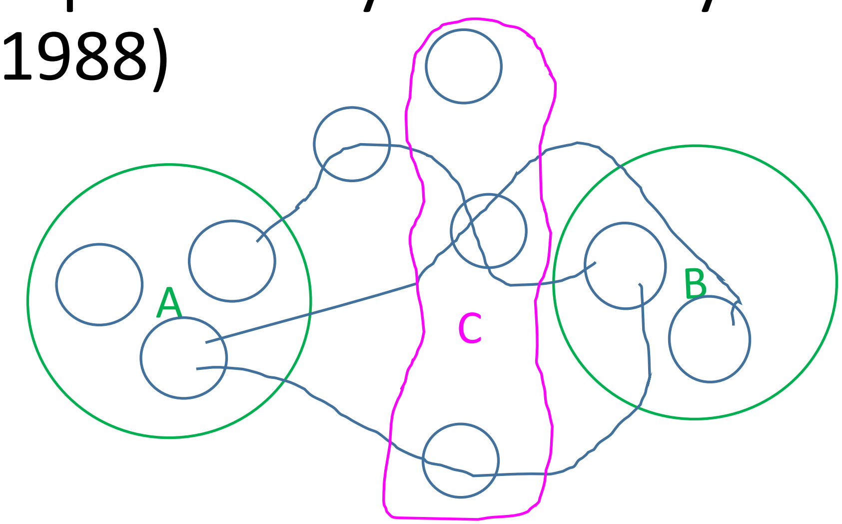
- For a Bayesian network, let A and B be a node, and C be a set of nodes that does not include A and B. We say a path from A to B is blocked when either of the followings holds
  - On the path from A to B, there is a node in C and the connection of the arcs is tail-to-tail or head-to-tail
  - At one of the nodes on the path from A to B, the connection of the arcs is head-to-head. In addition, the node and its all descendants are not included in C



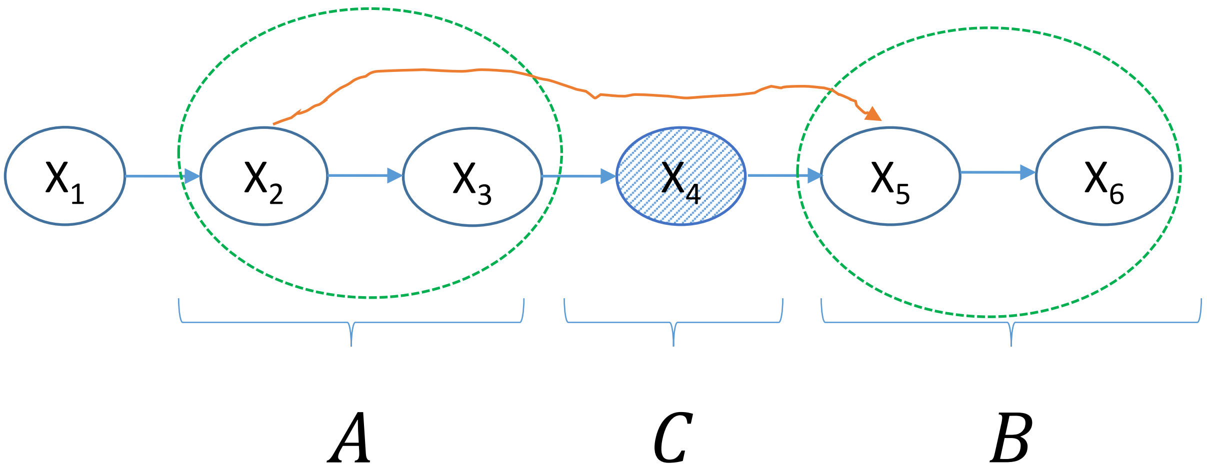
d-separation

For a Bayesian network, let A, B, and C be exclusive sets of nodes

- We say A is d-separated from B by C if all the paths starting from a node in A and ending at a node in B is blocked
- When A is d-separated from B by C, $A \perp\!\!\!\perp B \mid C$ holds for the joint probability defined by the Bayesian network (Pearl 1988)



Example



When X_4 is observed, all the paths from A to B is blocked at X_4

➡ $A \perp\!\!\!\perp B \mid C$

$$P(A, C, B) = P(A)P(C|A)P(B|A, C) = P(A)P(C|A)P(B|C)$$
$$= P(X_2, X_3)P(X_4|X_2, X_3)P(X_5, X_6|X_4)$$

Factor Graph

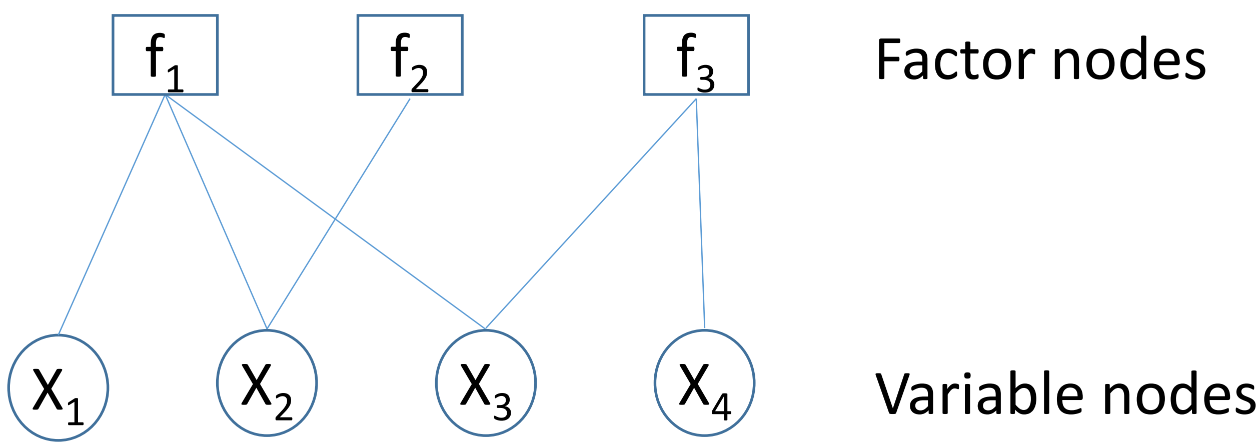
A joint probability model based on a bipartite and factorization of a function

Factor Graph

- A bipartite graph where one side of nodes represent random variables and the others represent functions
- The arcs represent dependencies of the functions to the variables
- A factor graph defines a joint probability

$$P(X_1, X_2 \cdots X_N) = \prod_{s \in subsets\ of\ variables} f_s(X_s)$$

Example:

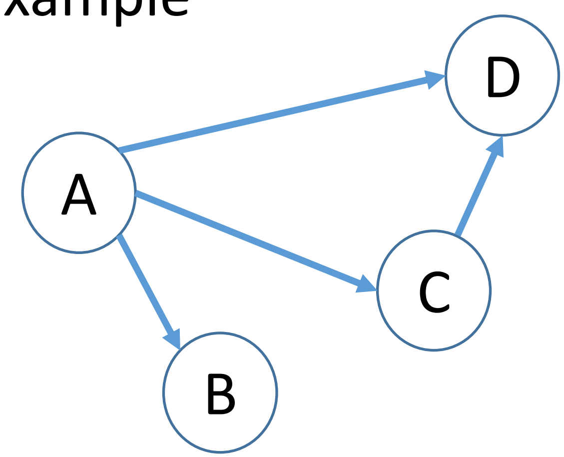


$$P(X_1, X_2, X_3, X_4) = f_1(X_1, X_2, X_3) f_2(X_2) f_3(X_3, X_4)$$

Factor Graph Representation of Bayesian Network

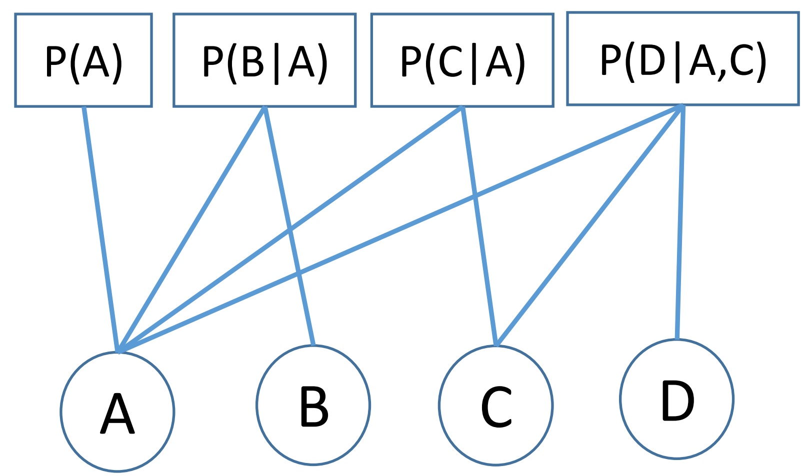
Each conditional probability can be regarded as a factor

Example



$$P(A)P(B \mid A)P(C \mid A)P(D \mid A, C)$$

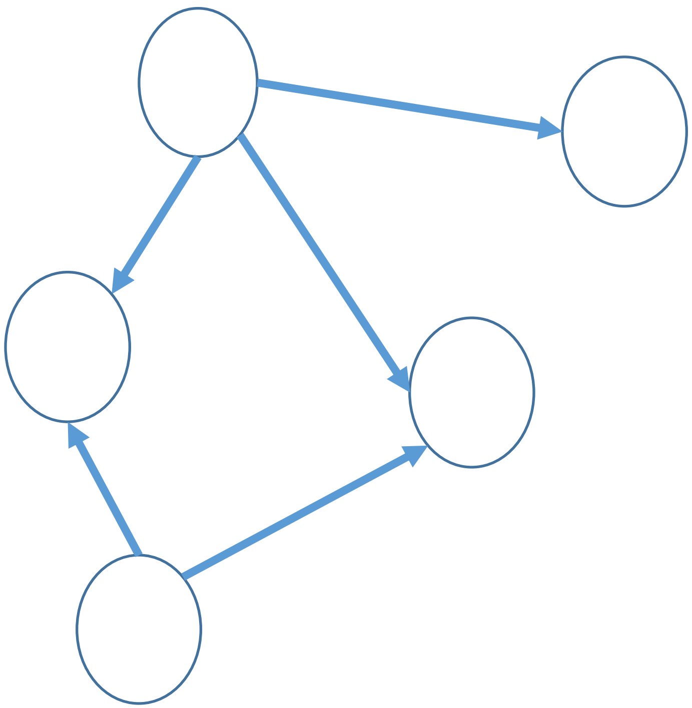
Bayesian network



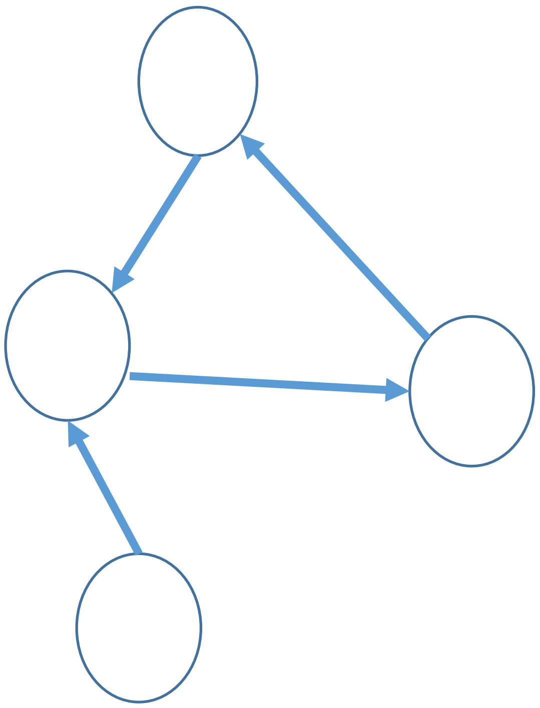
Factor graph

Exercise 2.1, 2.2

Q2.1) Is directed graph A a DAG?
Q2.2) Is directed graph B a DAG?



Graph A



Graph B

Exercise 2.3, 2.4, 2.5

Q2.3) Does $\{X_7, X_8\} \perp\!\!\!\perp \{X_4\} \mid \{X_1, X_3\}$ hold?

$$P(X_7, X_8 | X_4, X_1, X_3) = P(X_7, X_8 | X_1, X_3)$$

Q2.4) Does $\{X_4, X_5\} \perp\!\!\!\perp \{X_6\} \mid \{X_3\}$ hold?

$$P(X_4, X_5 | X_3, X_6) = P(X_4, X_5 | X_3)$$

Q2.5) Does $\{X_4, X_5\} \perp\!\!\!\perp \{X_6\} \mid \{X_3, X_8\}$ hold?

$$P(X_4, X_5 | X_3, X_6, X_8) = P(X_4, X_5 | X_3, X_8)$$