CS839:
Probabilistic Graphical Models

Lecture 2: Directed Graphical Models

Theo Rekatsinas

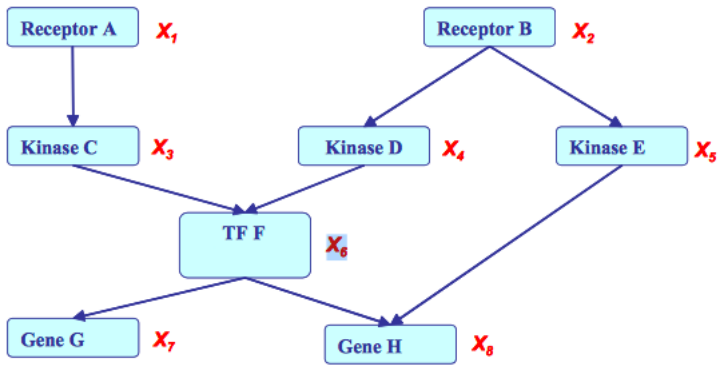
Questions?

- Waiting list
- Questions on other logistics

1. Intro to Bayes Nets

Representing Multivariate Distributions

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



$$
\begin{aligned}
&P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
=&P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2) \\
&P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)
\end{aligned}
$$

- If X_i's are **independent:** $P(X_i|\cdot) = P(X_i)$

$$
\begin{aligned}
&P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
=&P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)P(X_6)P(X_7)P(X_8)
\end{aligned}
$$

Notation

We will be using the notation from Koller and Friedman (at the end of the book)

- Random variable: X, Y, Z
- Random variable set (some time matrices): $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$
- Parameters: $\alpha, \beta, \gamma, \kappa, \theta$

Representation

-- example by Eric Xing

The Dishonest Casino

A casino has two dice:

- Fair die: P(1) = P(2) = … = P(6) = 1/6
- Loaded die: P(1) = P(2) = P(3) = P(5) = 1/10, P(6) = 1/2

The dealer switches back and forth between fair and loaded die once every 20 turns

Game:

1. You bet $X
2. You roll (always with a fair die)
3. Dealer rolls (maybe with a fair die, maybe with a loaded die)
4. Highest number wins $2X

Representation

-- example by Eric Xing

You have a sequence of rolls from the dealer

1245526462146146136136661664661…

Questions…

- How likely is this sequence, given our model of how the casino works? (Evaluation)
- What portion of the sequence was generated with a fair die, and what portion with a loaded one? (Decoding)
- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded and back? (Learning)

Knowledge Engineering

-- example by Eric Xing

- Modeling random variables
  - Latent (hidden)
  - Observed

Knowledge Engineering

-- example by Eric Xing

- Modeling random variables
  - Latent (hidden)
  - Observed

- Modeling the structure
  - Causal
  - Generative
  - Coupling

Knowledge Engineering

-- example by Eric Xing

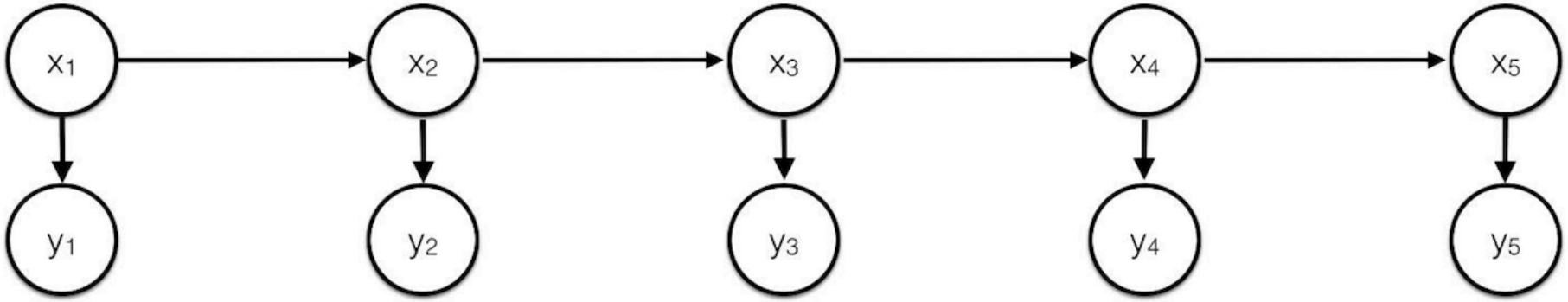
- Modeling random variables
  - Latent (hidden)
  - Observed
- Modeling the structure
  - Causal
  - Generative
  - Coupling
- Model probabilities
  - Conditional probabilities
  - Orders of magnitude

Hidden Markov Model

$p(y_t|x_t)$ observation probability

$p(x_t|x_{t-1})$ transition probability

$$p(X, Y) = p(x_1) \prod_{t=1}^{T-1} p(x_{t+1}|x_t) \prod_{t'=1}^{T} p(y_{t'}|x_{t'})$$



Bayesian Network

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

Bayesian Network

- **Factorization Theorem:**

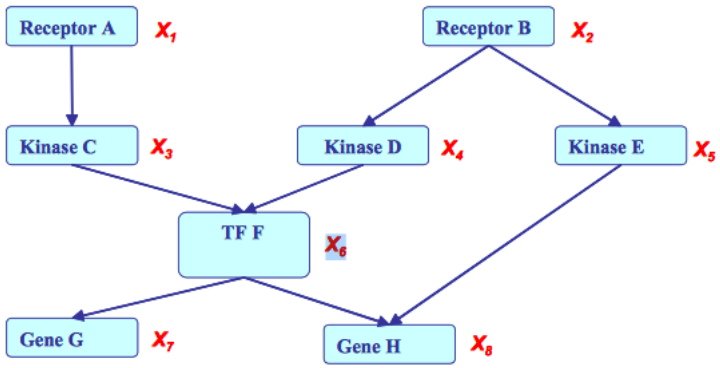
Given a DAG, the most general form of the probability distribution that is **consistent with** the graph factors according to "node given its parents":

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

where X_{π_i} is the set of parents of X_i, *d* is the number of nodes (variables) in the graph

Bayesian Network

- **Factorization Theorem Example**



$$
\begin{aligned}
&P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
=&P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_2)P(X_5|X_2) \\
&P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_5, X_6)
\end{aligned}
$$

Specification of a Bayesian Network

- Qualitative specification: structure and variables (roughly knowledge engineering)
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data

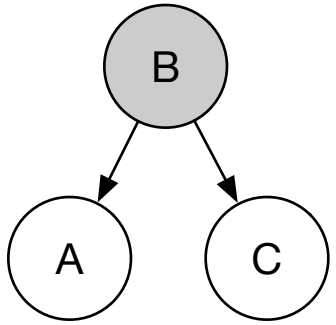
- Quantitative specification: instantiation of conditional probability distributions
  - Conditional probability tables
  - Continuous distributions

2. Local Structure and Independence

Local structure

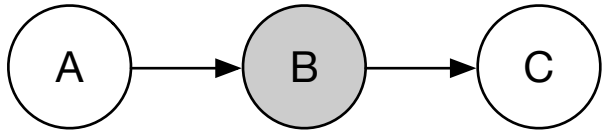
Conditional independences

- Common parent
  - Fixing B **decouples** A and C



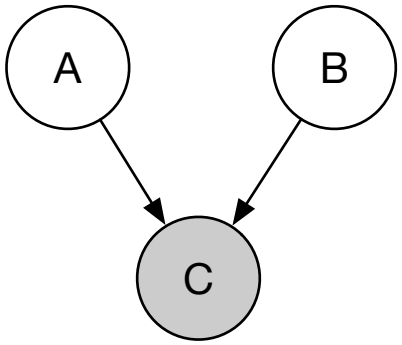
$$I(A \perp C | B)$$

- Cascade
  - Knowing B **decouples** A and C



$$I(C \perp A | B)$$

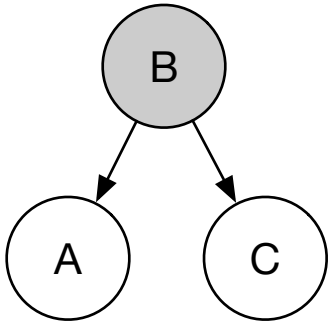
- V-structure
  - Knowing C **couples** A and B (knowing one variable explains the contribution of the other to a common child event)



Three foundational building blocks (compact language) for creating complex BNs

Why a graphical specification?

- Consider a factoring of P(A,B,C) = P(B)P(A|C)P(B|C)
- Consider $I(A \perp C | B)$
- Can we show that $I(P_\theta(A, B, C)) \equiv I(A \perp C | B)$?



I-maps

- **Def:** Let P be a distribution over ***X***. We define I(P) to be the set of independence assertions of the form $(X \perp Y | Z)$ that hold in P (despite how we set the parameter-values).

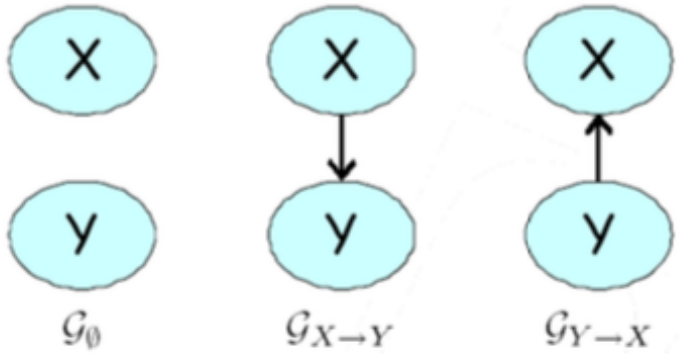
- **Def:** Let K be any graph object associated with a set of independencies I(K). We say that K is an I-map for a set of independencies I if $I(K) \subseteq I$

- We say that graph G is an I-map for P if G is an I-map for I(P), where we use I(G) as the set of independencies associated.

I-maps

- For F to be an I-map of P, it is necessary that G does not mislead us regarding independencies in P:
  - Any independence that G asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in G

- Example



P_1

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

P_2

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

From I(G) to local Markov assumptions of BNs

- A BN with structure G is a directed acyclic graph whose nodes represent random variables $X_1, \ldots, X_n$.

- **Local Markov assumptions** (one way to specify independencies:)
- **Def:**

Let Pa_{X_i} denote the parents of X_i in G and $NonDescendants_{X_i}$ denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of **local conditional independence assumptions** $I_l(G)$:

$$I_l(G) : \{X_i \perp NonDescendants_{X_i} | Pa_{X_i} : \forall I\}$$

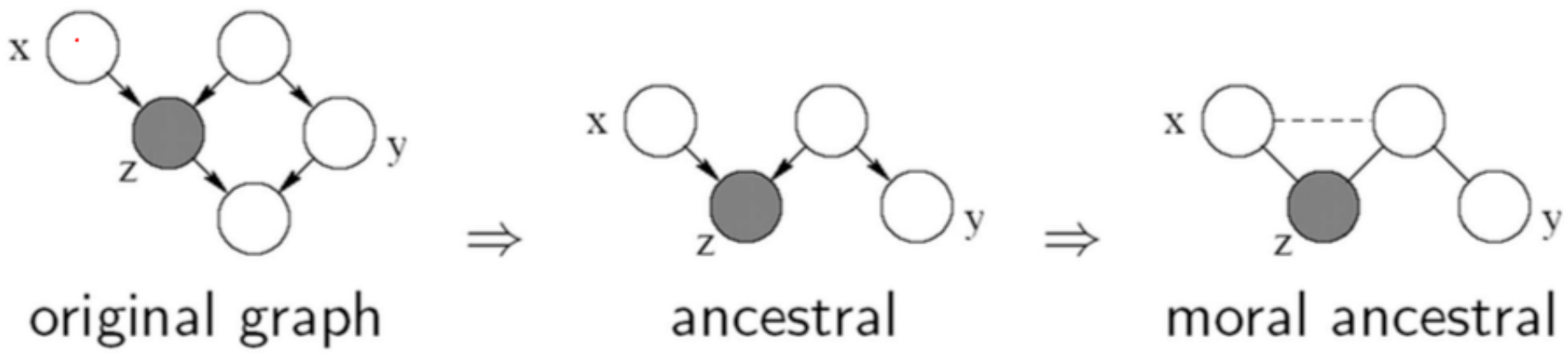
In other words, each node X_i is independent of its non-descendants given its parents.

Graph separation criterion

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

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

Example:



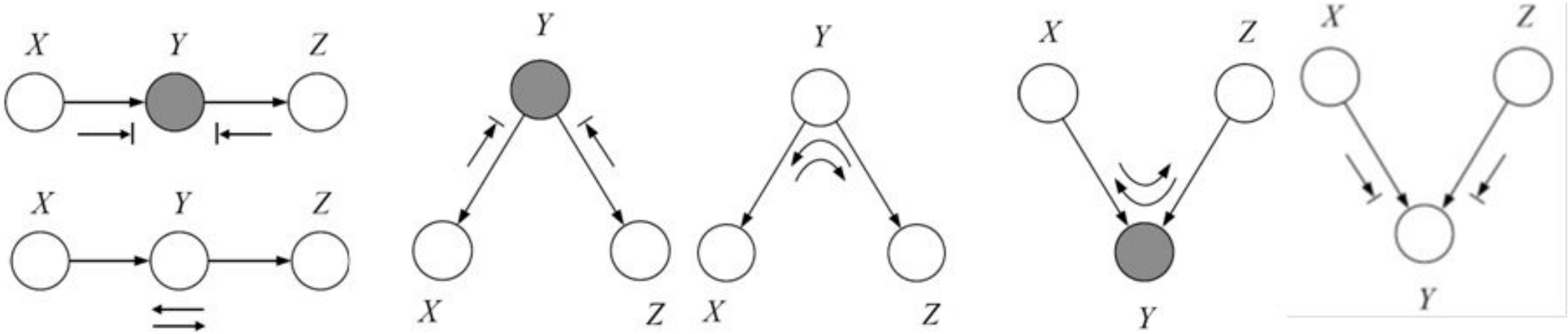
Active trail condition

- **Causal trail** X → Z → Y: active if and only if Z is not observed.
- **Evidential trail** X ← Z ← Y: active if and only if Z is not observed.
- **Common cause** X ← Z → Y: active if and only if Z is not observed.
- **Common effect (V-structure)** X → Z ← Y: active if and only if either Z or one of Z's descendants is observed.

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

Global Markov properties of BN

- X is d-separated (directed-separated) from Z given Y if we cannot send a ball from any node in X to any node in Z using the “Bayes-ball” algorithm.



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

Quantitative specification of P

- Separation properties in the graph imply independence properties about the associated variables
- **Equivalence Theorem**

  For a graph G,
  Let $\mathcal{D}_1$ denote the family of **all distributions** that satisfy $I(G)$.
  Let $\mathcal{D}_2$ denote the family of **all distributions** that factor according to G

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

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

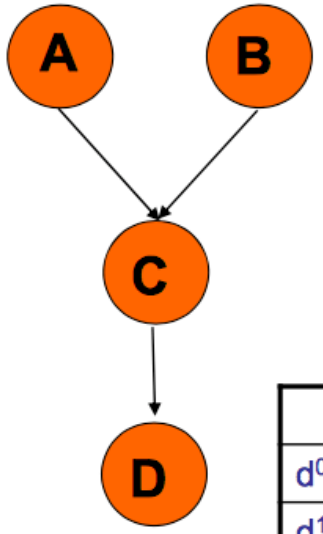
- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

Specification of BNs

- Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67



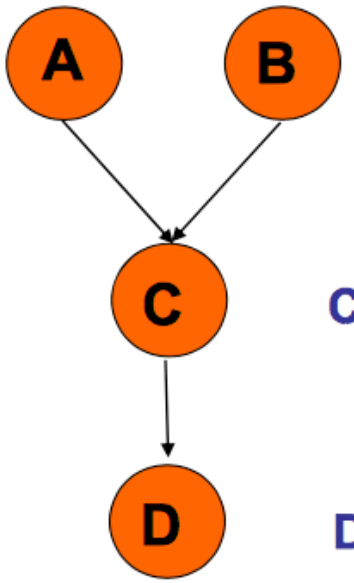
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

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

Specification of BNs

- Conditional probability density functions (CPDs)

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



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

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

Summary of BN semantics

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

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

Summary of BN semantics

- **Def:** A Bayesian network is a pair (G,P) where P factorizes over G, and P is specified as a set of CPDs associated with G's nodes.
  - Conditional independencies imply factorization
  - Factorization according to G implies the associated conditional independencies.
- Are there **other independences** that hold for every distribution P that factorizes over G?

Soundness and completeness

- Soundness:
  - **Theorem:** if a distribution P factorizes according to G, then I(G)$\subseteq$ I(P).

- Completeness:
  - **Claim:** for any distribution P that factorizes over G, if (X$\perp$Y | Z)$\in$I(P) then $d - sep_{\mathcal{G}}(X;Y|Z)$

Soundness and completeness

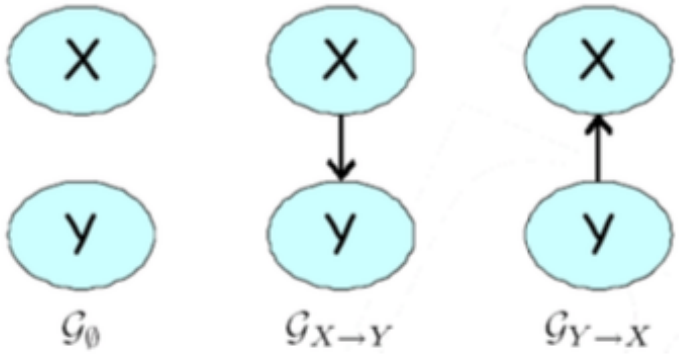
- Soundness:
  - **Theorem:** if a distribution P factorizes according to G, then I(G)$\subseteq$ I(P).

- Completeness:
  - **Claim:** for any distribution P that factorizes over G, if (X$\perp$Y | Z)$\in$I(P) then $d - sep_{\mathcal{G}}(X;Y|Z)$

- **Think of this:**
  - If X and Y are not d-separated given Z in G, then X and Y are dependent in all distributions P that factorize over G
  - Is this true?

Soundness and completeness

- Completeness:
  - **Claim:** for any distribution P that factorizes over G, if (X⊥Y | Z)∈I(P) then $d-sep_{\mathcal{G}}(X;Y|Z)$



P₁

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

P₂

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

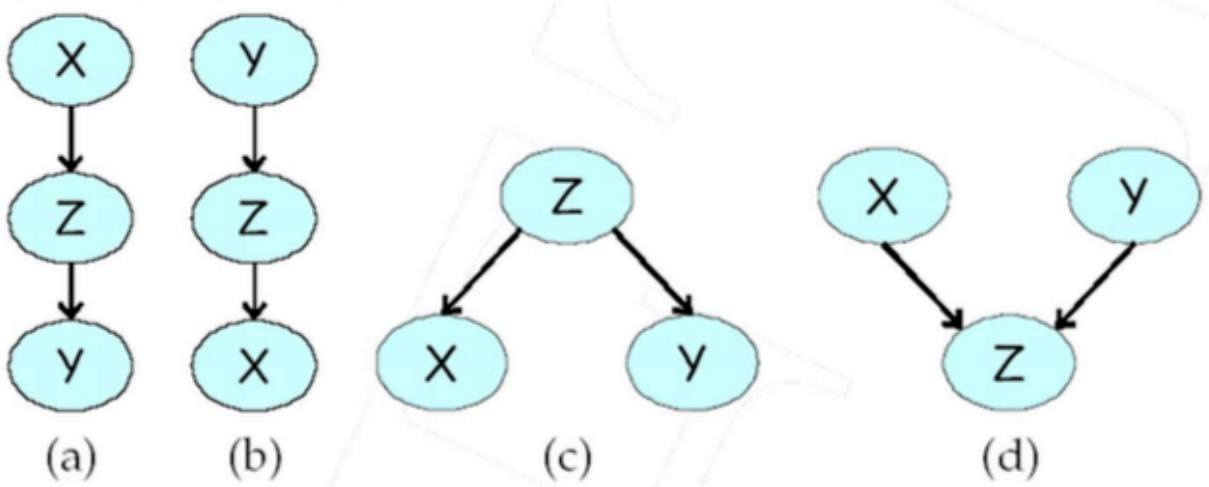
Soundness and completeness

- Completeness:
  - **Claim:** for any distribution P that factorizes over G, if $(X \perp Y \mid Z) \in I(P)$ then $d - sep_{\mathcal{G}}(X; Y|Z)$

- **Theorem:** Let G be a BN graph. If X and Y are not d-separated given Z in G, then X and Y are dependent in some distribution P that factorizes over G.

Uniqueness of BNs

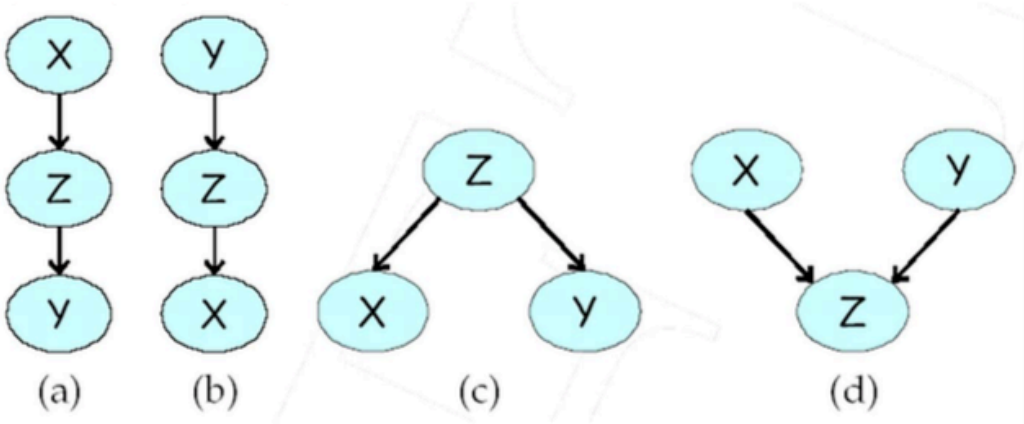
- Very different BN graphs can actually be equivalent, in that they encode precisely the same set of conditional independence assertions.



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

I-equivalence

- Def: Two BN graphs G1 and G2 over **X** are I-equivalent if I(G1) = I(G2).

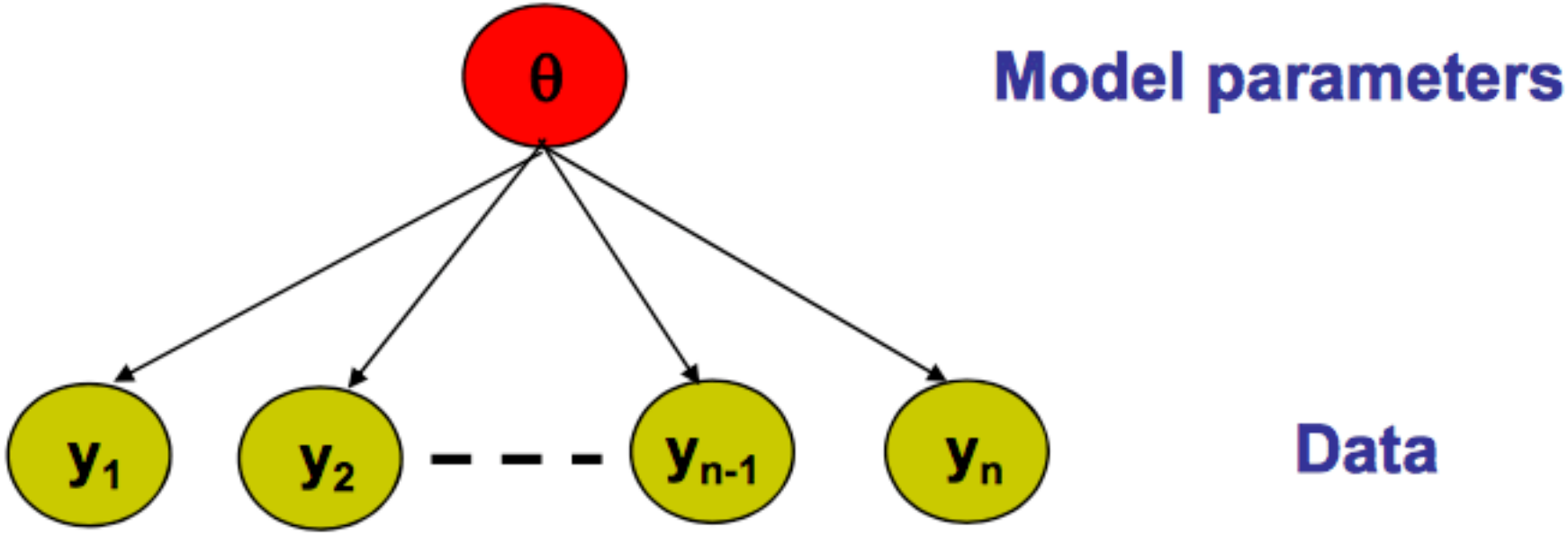


$(X \perp Y \mid Z)$.

- Any distribution P that can be factorized over one of these graphs can be factorized over the other. Implications when trying to determine directionality of influence.

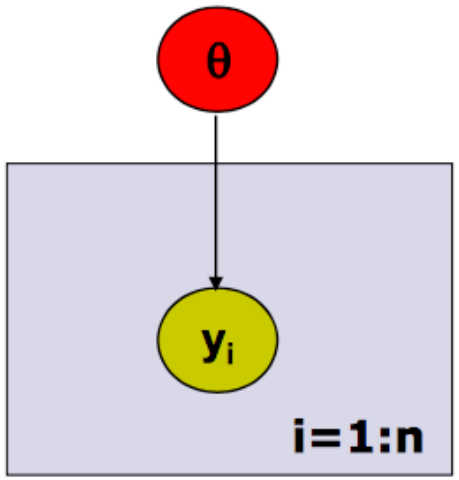
Simple BNs

- Conditionally Independent Observations



Simple BNs

- Plate model



Model parameters

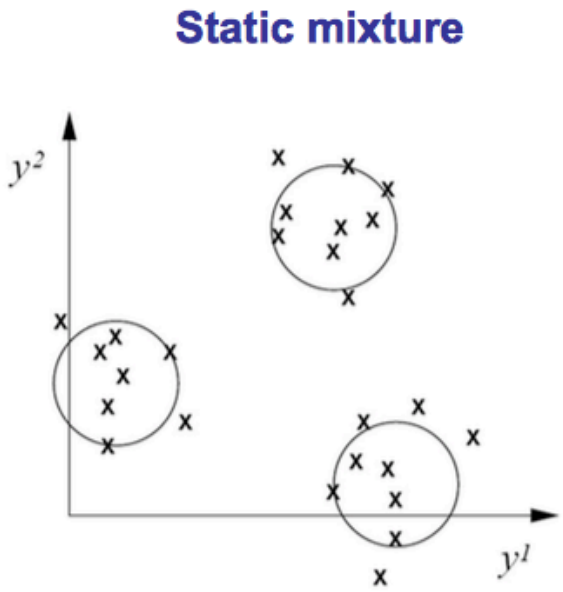
Data = $\{y_1,\ldots y_n\}$

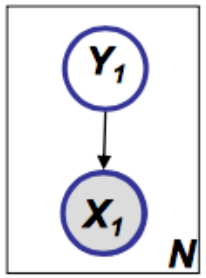
Plate = rectangle in graphical model

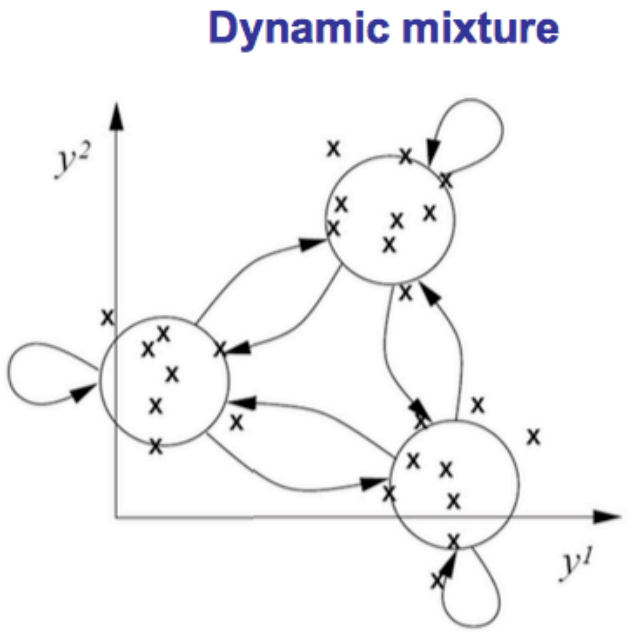
variables within a plate are replicated in a conditionally independent manner

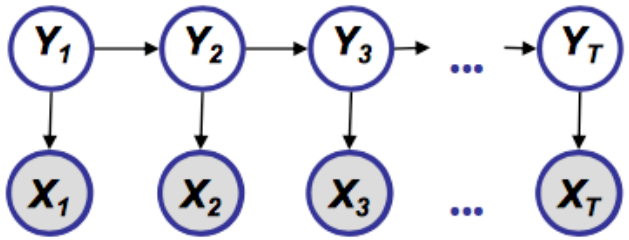
Simple BNs

- Hidden Markov Model









Summary

- A Bayesian network is a pair (G, P) where P factorizes over G, and where P is specified as a set of local conditional probability distributions.
- A BN captures “causality”, “generative schemes”, “asymetric influences” etc.
- Local and global independence properties are identifiable via d-separation criteria