Directed GMs: Bayesian Networks

Kayhan Batmanghelich

Announcements

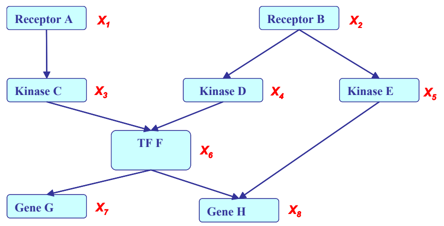
- HW0 is out
- Class recording on YouTube
- Readings will be posted today
- Piazza
- Office hours will be posted soon
- Who is going to scribe?

```
In [1]: import numpy as np

In [2]: row, col = np.random.randint(1,5,size=(1,)), np.random.randint(1,10,size=(1,))

In [3]: print row,col
[4] [6]
```

Two types of GMs

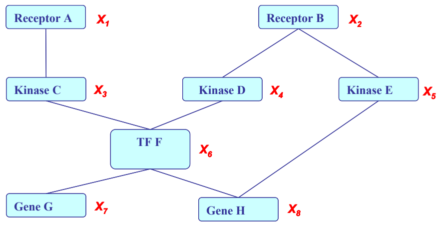
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

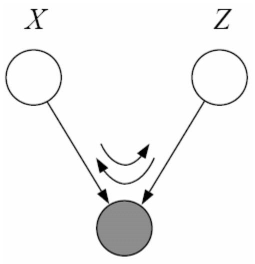


- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



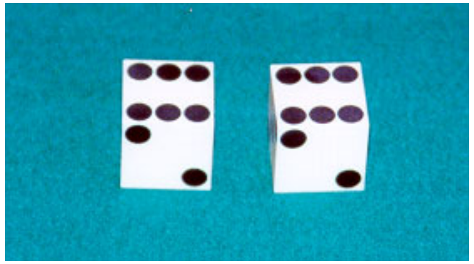
- **Representation of directed GM**



Notation

- Variable, value and index
- Random variable
- Random vector
- Random matrix
- Parameters

Example: The Dishonest Casino



A casino has two dice:

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

Casino player switches back-&-forth between fair and loaded die once every 20 turns

Game:

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

Puzzles regarding the dishonest casino

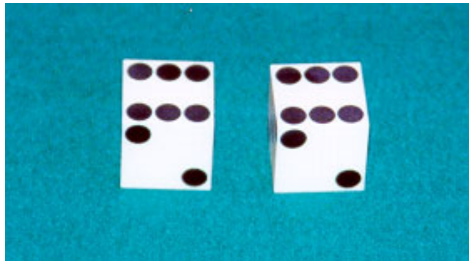
GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

- How likely is this sequence, given our model of how the casino works?
    - This is the **EVALUATION** problem
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
    - This is the **DECODING** question
- 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?
    - This is the **LEARNING** question

Knowledge Engineering

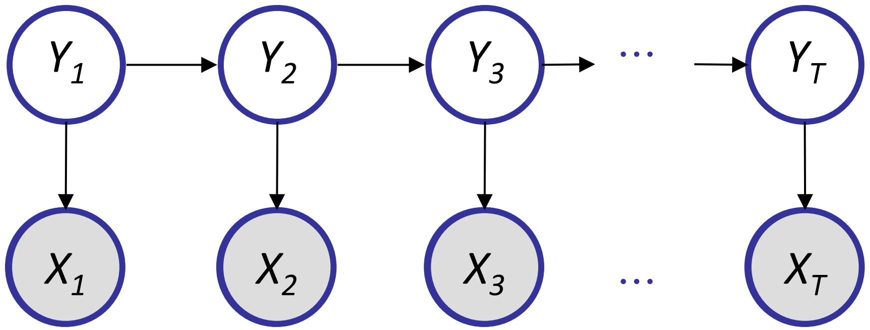


- **Picking variables**
  - **Observed**
  - **Hidden**
- **Picking structure**
  - **CAUSAL**
  - **Generative**
  - **Coupling**
- **Picking Probabilities**
  - **Zero probabilities**
  - **Orders of magnitudes**
  - **Relative values**

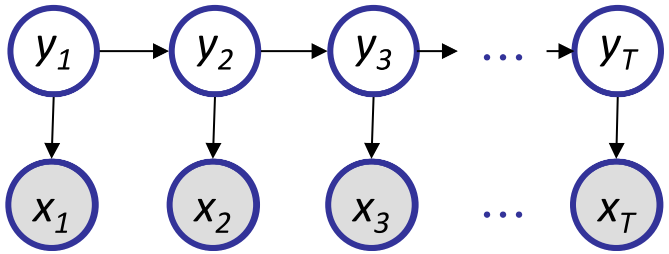
Hidden Markov Model

The underlying source:
Speech signal
genome function
dice

The sequence:
Phonemes
DNA sequence
sequence of rolls



Getting Insights from the Probability



- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$ and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T,$
- To find how likely is the parse: (given our HMM and the sequence)

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}) &= p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \qquad \text{(Joint probability)} \\
&= p(y_1)\, p(x_1 \mid y_1)\, p(y_2 \mid y_1)\, p(x_2 \mid y_2) \ldots p(y_T \mid y_{T-1})\, p(x_T \mid y_T) \\
&= p(y_1)\, P(y_2 \mid y_1) \ldots p(y_T \mid y_{T-1}) \times p(x_1 \mid y_1)\, p(x_2 \mid y_2) \ldots p(x_T \mid y_T) \\
&= p(y_1, \ldots\ldots, y_T)\, p(x_1 \ldots\ldots x_T \mid y_1, \ldots\ldots, y_T)
\end{aligned}
$$

- How far on the tail (Marginal probability): $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_N} \pi_{y_1} \prod_{t=2}^{T} a_{y_{t-1}, y_t} \prod_{t=1}^{T} p(x_t \mid y_t)$

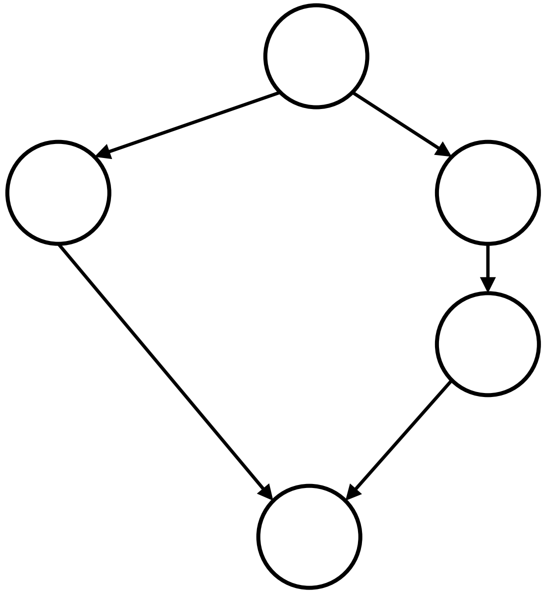
- When did he use unfair dice (Posterior probability): $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$

- We will learn how to do this explicitly (polynomial time)

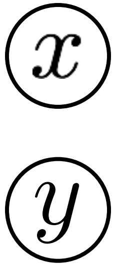
Directed Graphical Model (Bayesian Network)

- Nodes represent observed and unobserved random variables. Edges denote influence/dependence.

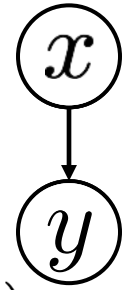
- The graph denotes the data generating procedure.



- It is a data structure/language to represent factorization of joint distribution.

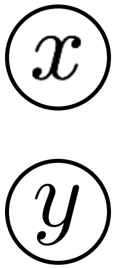


$$p(x, y) = p(x)p(y)$$

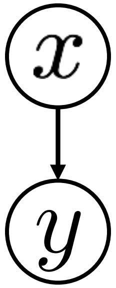


$$p(x, y) = p(x)p(y|x)$$

- One can read the set of conditional independence from the graph. .



$$x \perp\!\!\!\perp y$$



$$x \not\!\perp\!\!\!\perp y$$

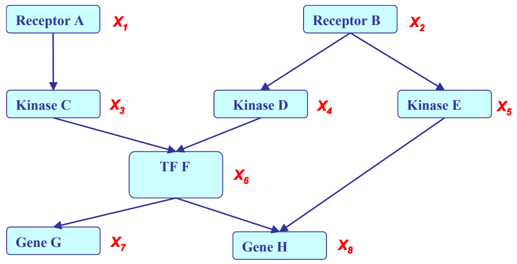
Bayesian Network: Factorization Theorem

- **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":

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

where $\mathbf{X}_{\pi_i}$ is the set of parents of X_i, d is the number of nodes (variables) in the graph.

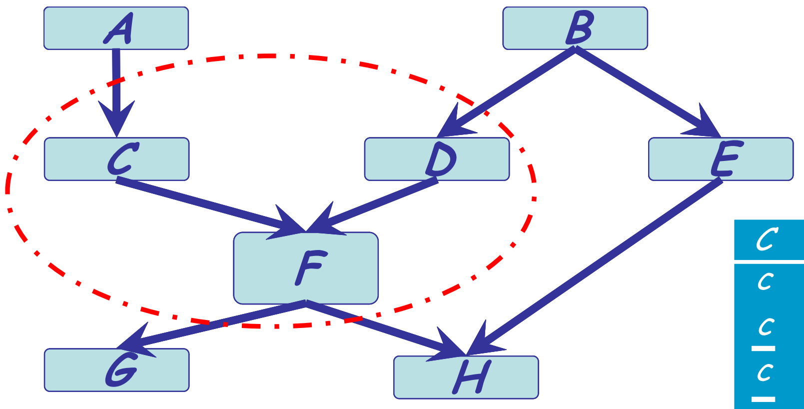


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= P(X_1) P(X_2) P(X_3| X_1) P(X_4| X_2) P(X_5| X_2) P(X_6| X_3, X_4) P(X_7| X_6) P(X_8| X_5, X_6)$$

Specification of a directed GM

- There are two components to any GM:
  - the *qualitative* specification specifies a family of distributions
  - the *quantitative* specification specifies a distribution from the family



C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

Where does the Qualitative Specification come from?

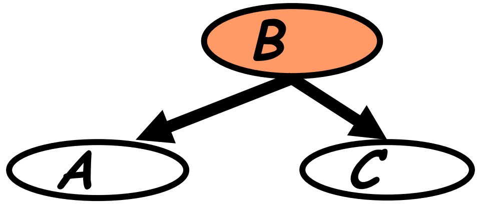
- Prior knowledge of causal relationships
- Prior knowledge of modular relationships
- Assessment from experts
- Learning from data
- We simply link a certain architecture (e.g. a layered graph)
- …

DAG and Independences

Local Structures & Independencies

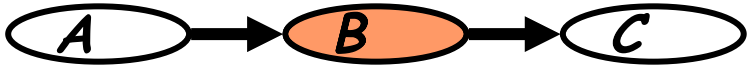
- Common parent
  - Fixing B decouples A and C

    "given the level of gene B, the levels of A and C are independent"



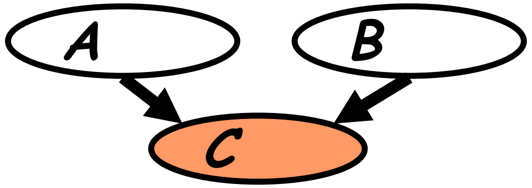
- Cascade
  - Knowing B decouples A and C

    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



- V-structure
  - Knowing C couples A and B
    because A can "explain away" B w.r.t. C

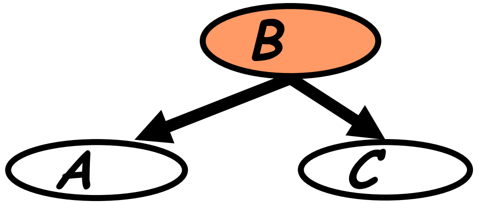
    "If A correlates to C, then chance for B to also correlate to B will decrease"



- The language is compact, the concepts are rich!

A simple proof:

Factorization by the graph $\equiv$ Independent Set



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

$$\mathcal{I}(\mathcal{G}) = \{A \perp\!\!\!\perp B|C\}$$

I-maps

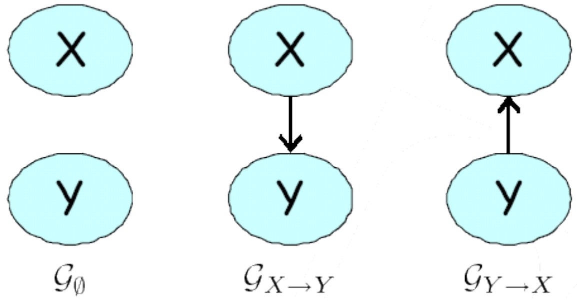
- **Defn :** Let P be a distribution over $\boldsymbol{X}$. We define I(P) to be the set of independence assertions of the form $(X \perp Y \mid Z)$ that hold in P (however how we set the parameter-values).

- **Defn :** 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, I(K) $\subseteq$ I.

- We now say that 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-map is a conservative specification of P

Ex: Which of the following graphs allows for both probability distributions?



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

Any independence that G asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in G.

The intuition behind I(G) local Markov assumptions of BN

Remember the *Bayesian network structure:*

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

- **Defn :**

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*** $I_\ell(G)$:

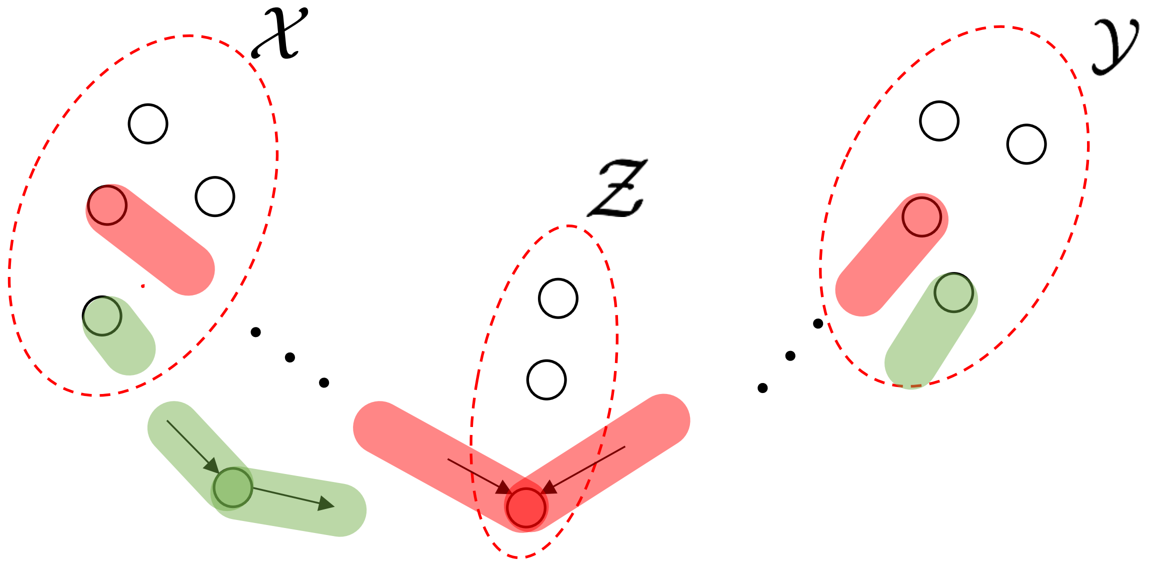
$$\mathcal{I}_\ell(\mathcal{G}) = \{X_i \perp\!\!\!\perp NonDescendants(X_i) \quad | pa(X_i) : \forall i\}$$

In other words, each node X_i is independent of its nondescendants given its parents.

d-connection and d-separation

Defn: If G is a directed graph in which $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are disjoint sets of vertices, then $\mathcal{X}$ and $\mathcal{Y}$ are d-connected by $\mathcal{Z}$ in $\mathcal{G}$ if and only if there exists an undirected path U between some vertex in $\mathcal{X}$ and some vertex in $\mathcal{Y}$ such that for every collider C on U, either C or a descendent of C is in $\mathcal{Z}$, and no non-collider on U is in $\mathcal{Z}$.

$\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$ in $\mathcal{G}$ if and only if they are not d-connected by $\mathcal{Z}$ in $\mathcal{G}$.

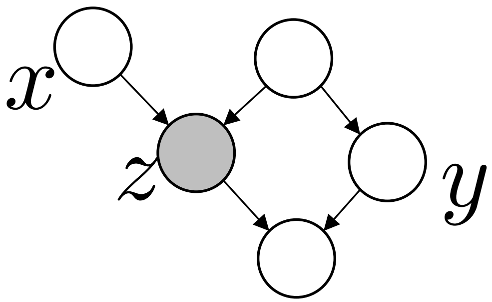


$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

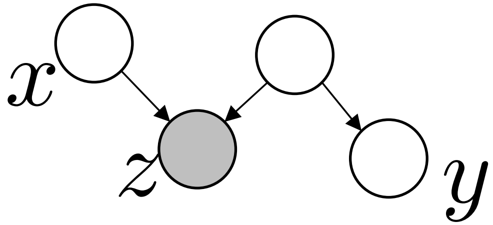
Alternative Definition

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

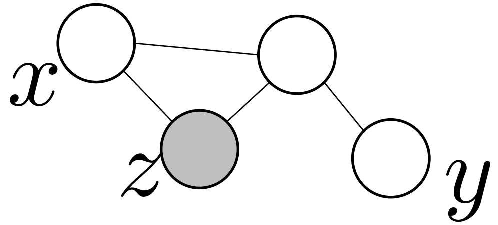
- Example:



Original graph



ancestral



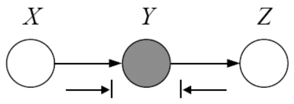
Moral ancestral

Bayes Ball Algorithm: Testing $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$

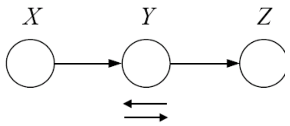
- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

Causal Trail:

blocked

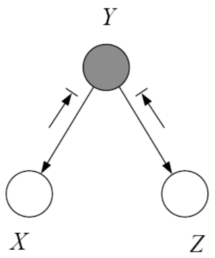


Active

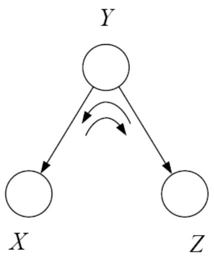


Common Cause:

blocked

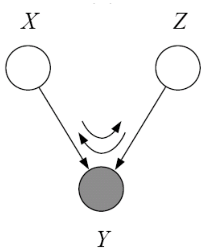


Active

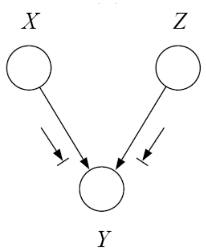


Common Effect:

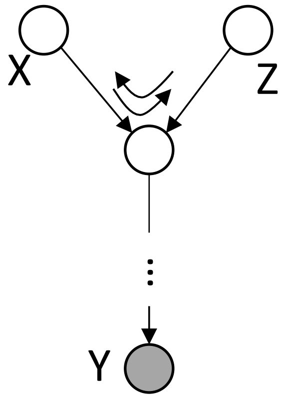
Active



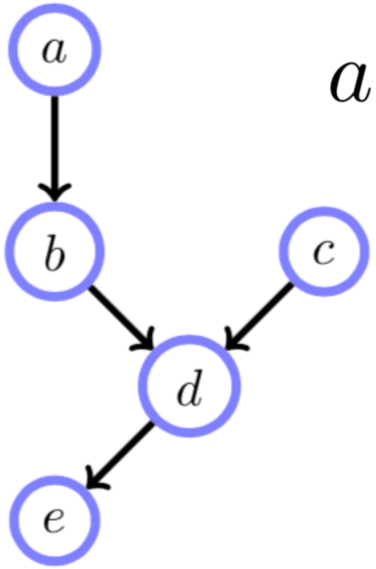
blocked



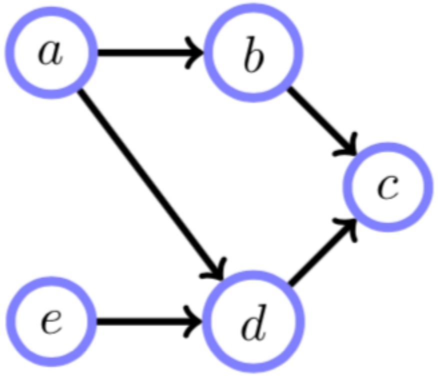
Active



Example:

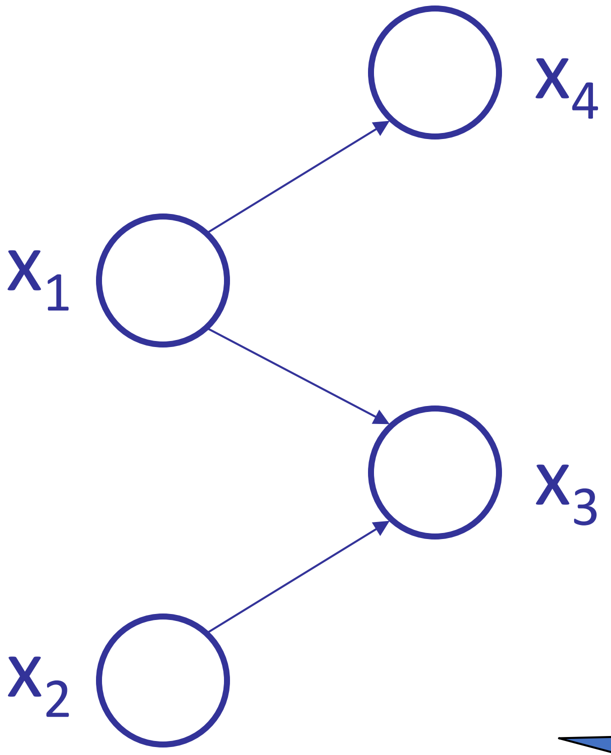


$a \perp\!\!\!\perp e \mid b?$



$a \perp\!\!\!\perp e \mid c?$

Example:



- Complete the I(G) of this graph:

Scriber please fill in
the rest of this slide !

A bit of Theories

Toward quantitative specification of probability distribution

- Separation properties in the graph imply independence properties about the associated variables

- **The 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,

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | pa(X_i))$$

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

Soundness and completeness

D-separation is sound and "complete" w.r.t. BN factorization law

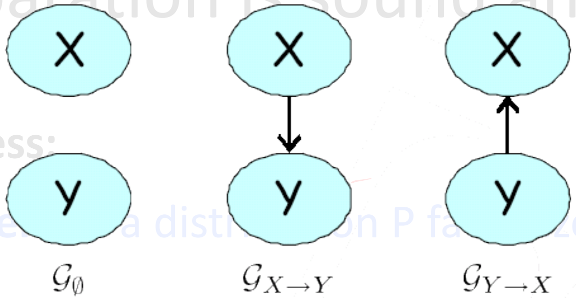
Soundness:

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 $(X \perp Y \mid Z) \in I(P)$ then $d\text{-}sep_G(X; Y \mid Z)$?

Soundness and completeness



P_1

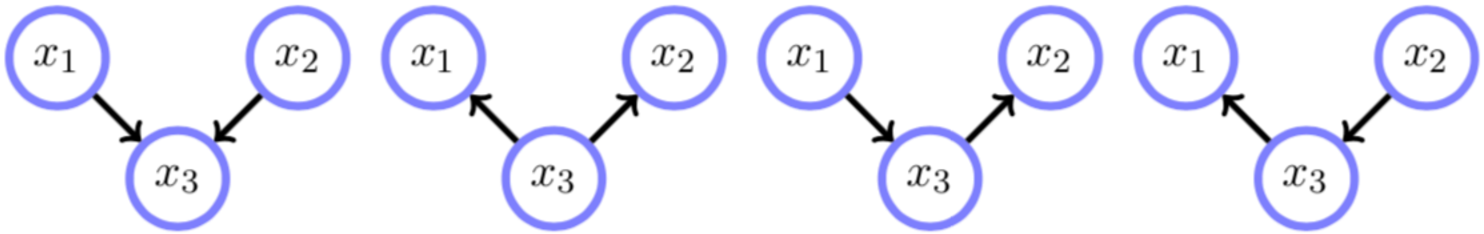
X	Y	$P(X,Y)$
x^0	y^0	0.08
x^0	y^1	0.32
x^1	y^0	0.12
x^1	y^1	0.48

- **Theorem :** For **almost all** distributions P that factorize over G, i.e., for all distributions except for a set of "measure zero" in the space of CPD parameterizations, we have that I(P) = I(G)

- **Thm**: 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 BN

- Which graphs satisfy $\mathcal{I}(\mathcal{G}) = \{x_1 \perp\!\!\!\perp x_2 | x_3\}$?



(a) (b) (c) (d)

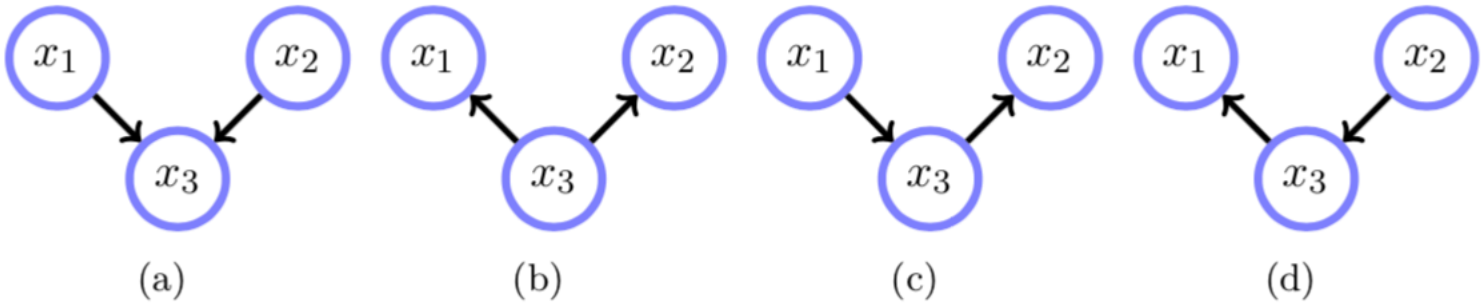
- You can see that in the factorization:

$$\underbrace{p(x_2|x_3)p(x_3|x_1)p(x_1)}_{graph(c)} = p(x_2, x_3)p(x_3, x_1)/p(x_3) = p(x_1|x_3)p(x_2, x_3)$$

$$= \underbrace{p(x_1|x_3)p(x_3|x_2)p(x_2)}_{graph(d)} = \underbrace{p(x_1|x_3)p(x_2|x_3)p(x_3)}_{graph(b)}$$

I-equivalence

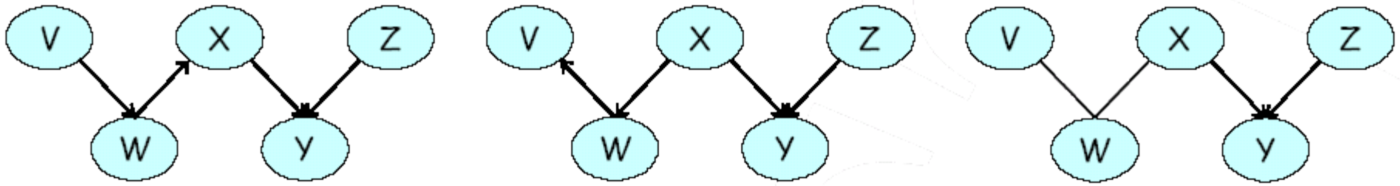
- Which graphs satisfy $\mathcal{I}(\mathcal{G}) = \{x_1 \perp\!\!\!\perp x_2 | x_3\}$?



- **Defn :** Two BN graphs G1 and G2 over $\boldsymbol{X}$ are *I-equivalent* if I(G1) = I(G2).
  - Any distribution P that can be factorized over one of these graphs can be factorized over the other.
  - Furthermore, there is no intrinsic property of P that would allow us associate it with one graph rather than an equivalent one.
  - This observation has important implications with respect to our ability to determine the directionality of influence.

Detecting I-equivalence

- **Defn :** The *skeleton* of a Bayesian network graph G over V is an undirected graph over V that contains an edge $\{X, Y\}$ for every edge (X, Y) in G.



(a) (b) (c)

- **Thm :** Let G_1 and G_2 be two graphs over V. If G_1 and G_2 have the same skeleton and the same set of v-structures then they are I-equivalent.

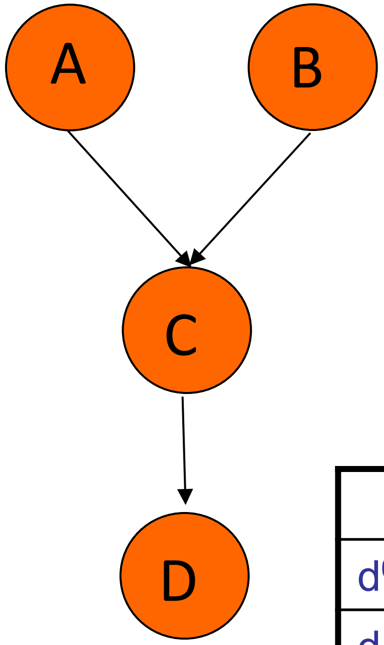
Practical Examples

Example of CPD for Discrete BN

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

P(a,b,c.d) =
P(a)P(b)P(c|a,b)P(d|c)



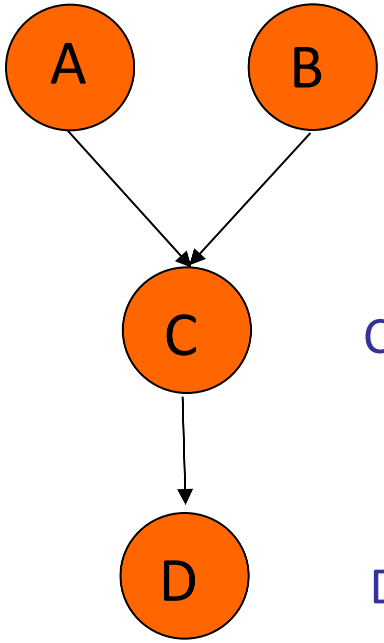
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

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

Example of CPD for Continuous BN

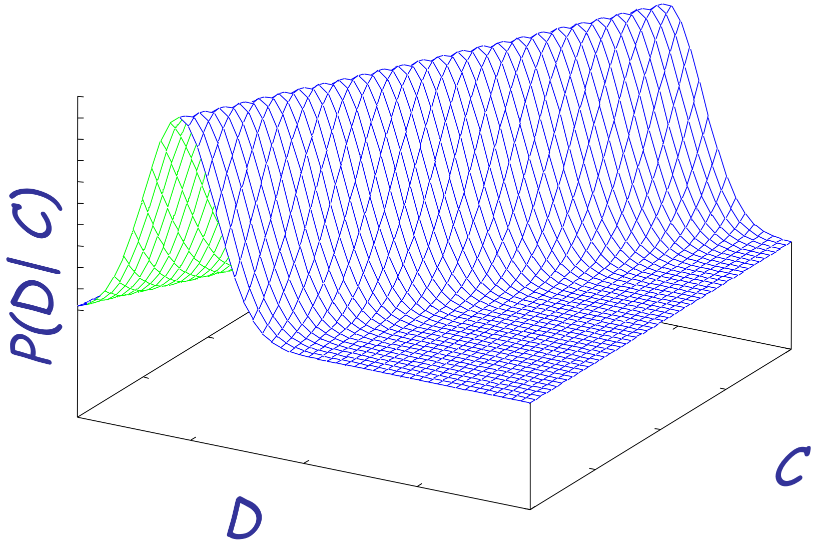
$A \sim N(\mu_a, \Sigma_a)$ $\quad$ $B \sim N(\mu_b, \Sigma_b)$

$P(a,b,c.d) =$
$P(a)P(b)P(c|a,b)P(d|c)$

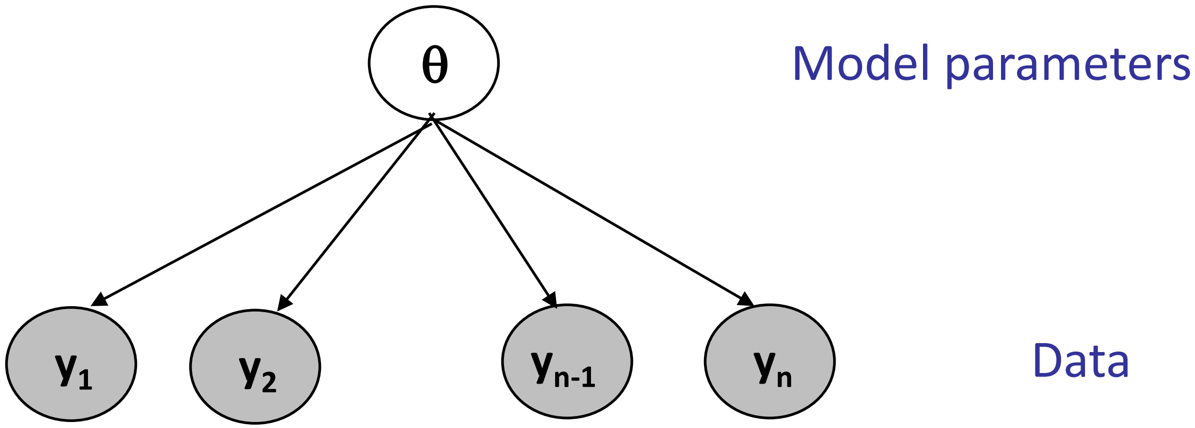


$C \sim N(A+B, \Sigma_c)$

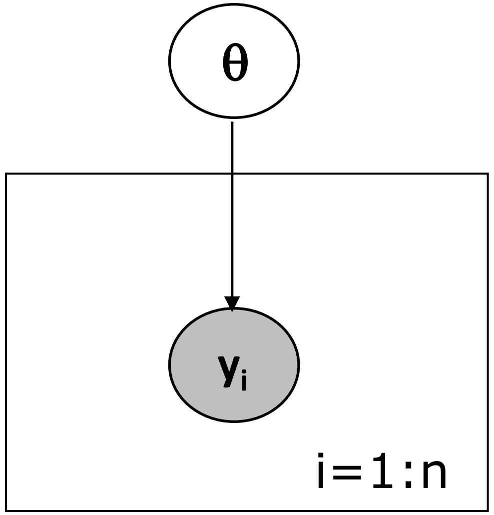
$D \sim N(\mu_d + C, \Sigma_d)$



Simple BNs:
Conditionally Independent Observations



The “Plate” Micro



Model parameters

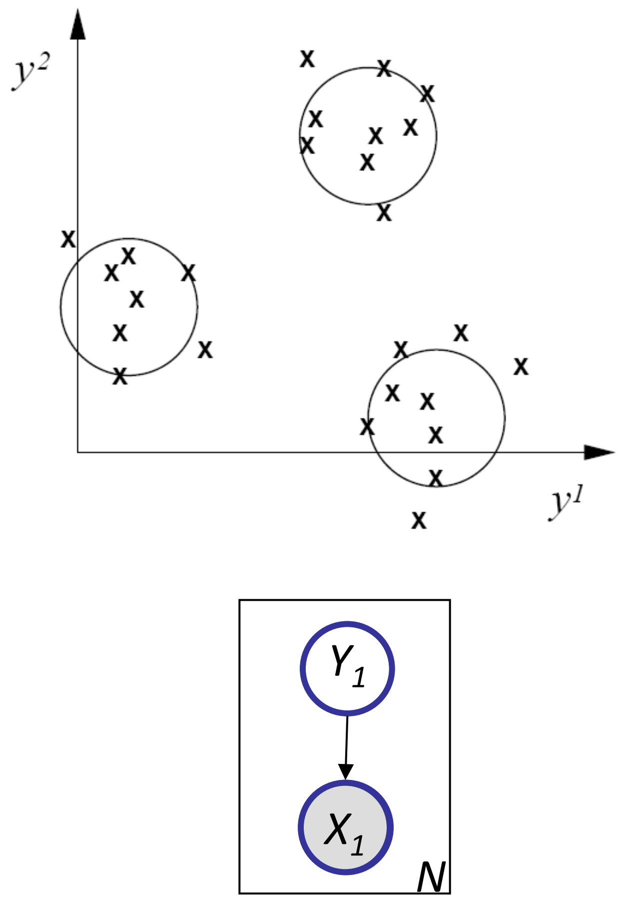
Data = $\{y_1, \ldots y_n\}$

Plate = rectangle in graphical model

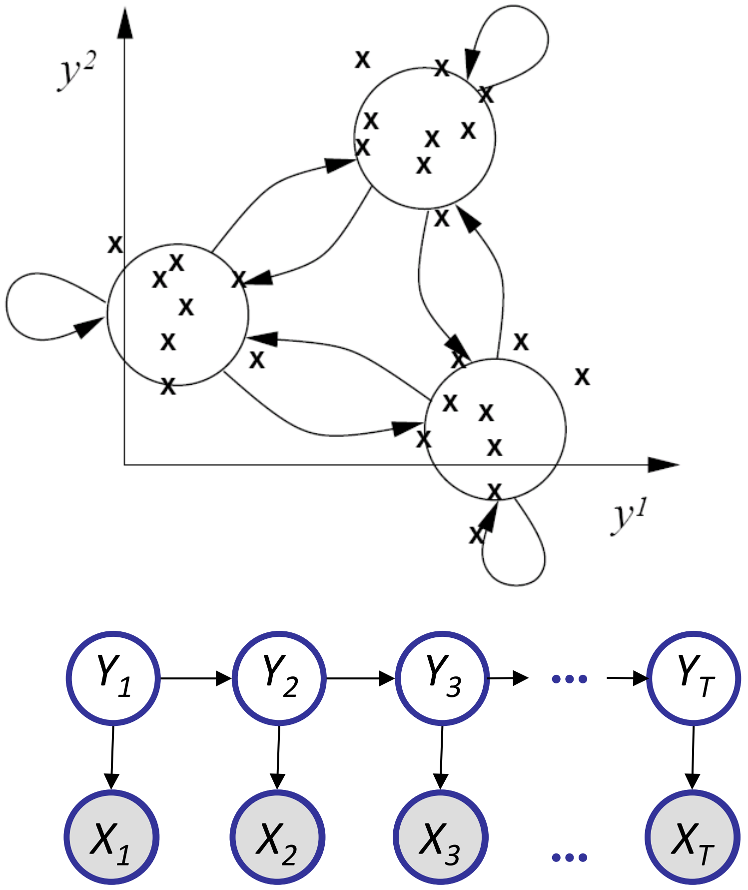
variables within a plate are replicated
in a conditionally independent manner

Hidden Markov Model:
from static to dynamic mixture models

Static mixture



Dynamic mixture



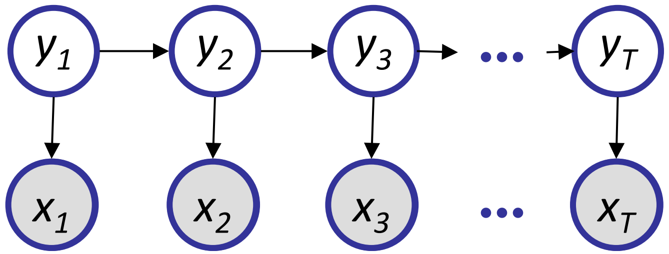
Definition (of HMM)

- Observation space

  Alphabetic set: $\mathrm{C} = \{c_1, c_2, \cdots, c_K\}$

  Euclidean space: R^d

- Index set of hidden states $\mathrm{I} = \{1, 2, \cdots, M\}$



- Transition probabilities between any two states

$$p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j},$$

or $$p(y_t \mid y_{t-1}^i = 1) \sim \mathrm{Multinomial}(a_{i,1}, a_{i,1}, \ldots, a_{i,M}), \forall i \in \mathrm{I}.$$

- Start probabilities

$$p(y_1) \sim \mathrm{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$$

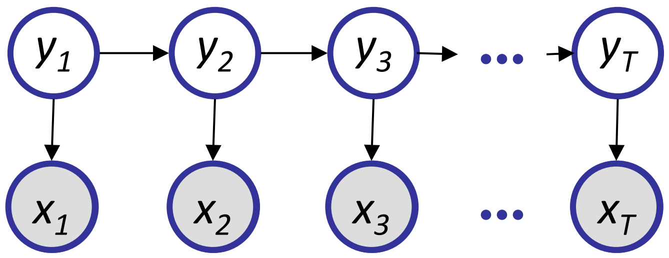
- Emission probabilities associated with each state

$$p(x_t \mid y_t^i = 1) \sim \mathrm{Multinomial}(b_{i,1}, b_{i,1}, \ldots, b_{i,K}), \forall i \in \mathrm{I}.$$

or in general: $$p(x_t \mid y_t^i = 1) \sim \mathrm{f}(\cdot \mid \theta_i), \forall i \in \mathrm{I}.$$

Probability of a parse

- Given a sequence $\mathbf{x} = x_1 \ldots\ldots x_T$ and a parse $\mathbf{y} = y_1, \ldots\ldots, y_T,$
- To find how likely is the parse:
  (given our HMM and the sequence)



$$
\begin{aligned}
p(\mathbf{x}, \mathbf{y}) &= p(x_1 \ldots\ldots x_T, y_1, \ldots\ldots, y_T) \qquad \text{(Joint probability)} \\
&= p(y_1)\, p(x_1 \mid y_1)\, p(y_2 \mid y_1)\, p(x_2 \mid y_2) \ldots p(y_T \mid y_{T-1})\, p(x_T \mid y_T) \\
&= p(y_1)\, \mathrm{P}(y_2 \mid y_1) \ldots p(y_T \mid y_{T-1}) \times p(x_1 \mid y_1)\, p(x_2 \mid y_2) \ldots p(x_T \mid y_T) \\
&= p(y_1, \ldots\ldots, y_T)\, p(x_1 \ldots\ldots x_T \mid y_1, \ldots\ldots, y_T)
\end{aligned}
$$

Summary: take home messages

- **Defn (3.2.5):** A *Bayesian network* is a pair (G, P) where P factorizes over G, and where P is specified as set of local conditional probability dist. CPDs associated with G's nodes.
- A BN capture "causality", "generative schemes", "asymmetric influences", etc., between entities
- Local and global independence properties identifiable via d- separation criteria (Bayes ball)
- Computing joint likelihood amounts multiplying CPDs
  - But computing marginal can be difficult
  - Thus inference is in general hard
- Important special cases:
  - Hidden Markov models
  - Tree models

A few myths about graphical models

- They require a localist semantics for the nodes
- They require a causal semantics for the edges
- They are necessarily Bayesian
- They are intractable

Extra Slides

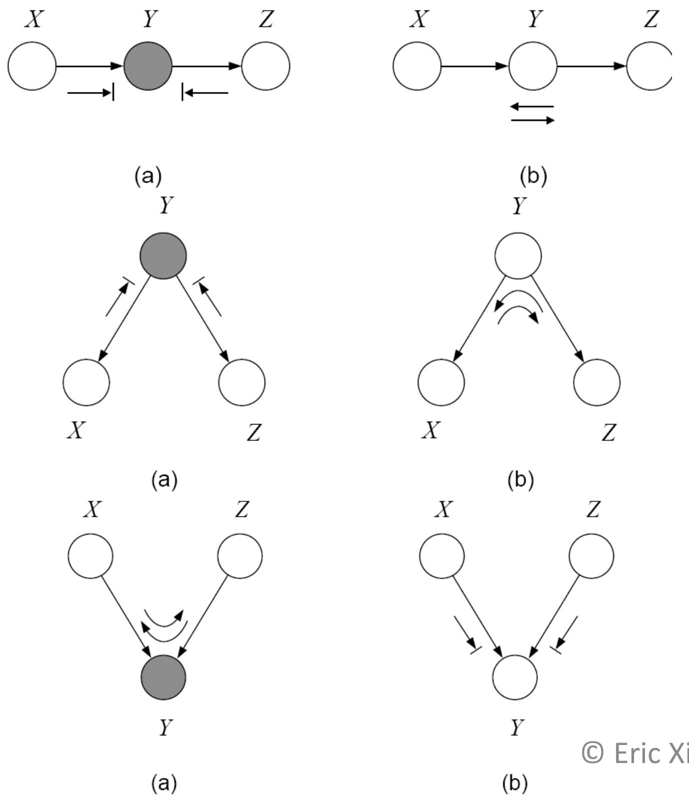
Active trail

- **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** X → Z ← Y **:** active if and only if either Z or one of Z's descendants is observed

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

What is in I(G) ---
Global Markov properties of BN

- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):



- Defn: $\mathcal{I}(\mathcal{G})$=all independence properties that correspond to d-separation:

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

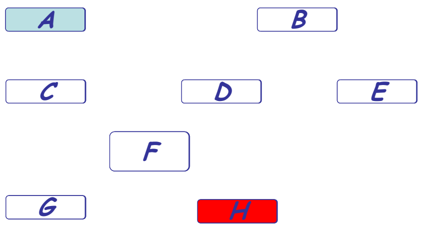
- D-separation is sound and complete (more details later)

Summary:
Representing Multivariate Distribution

- Representation: what is the joint probability dist. on multiple variables?

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8,)$$

  - How many state configurations in total? --- 2^8
  - Are they all needed to be represented?
  - **Do we get any scientific/medical insight?**



- Factored representation: the chain-rule

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

  - This factorization is true for any distribution and any variable ordering
  - Do we save any parameterization cost?

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

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

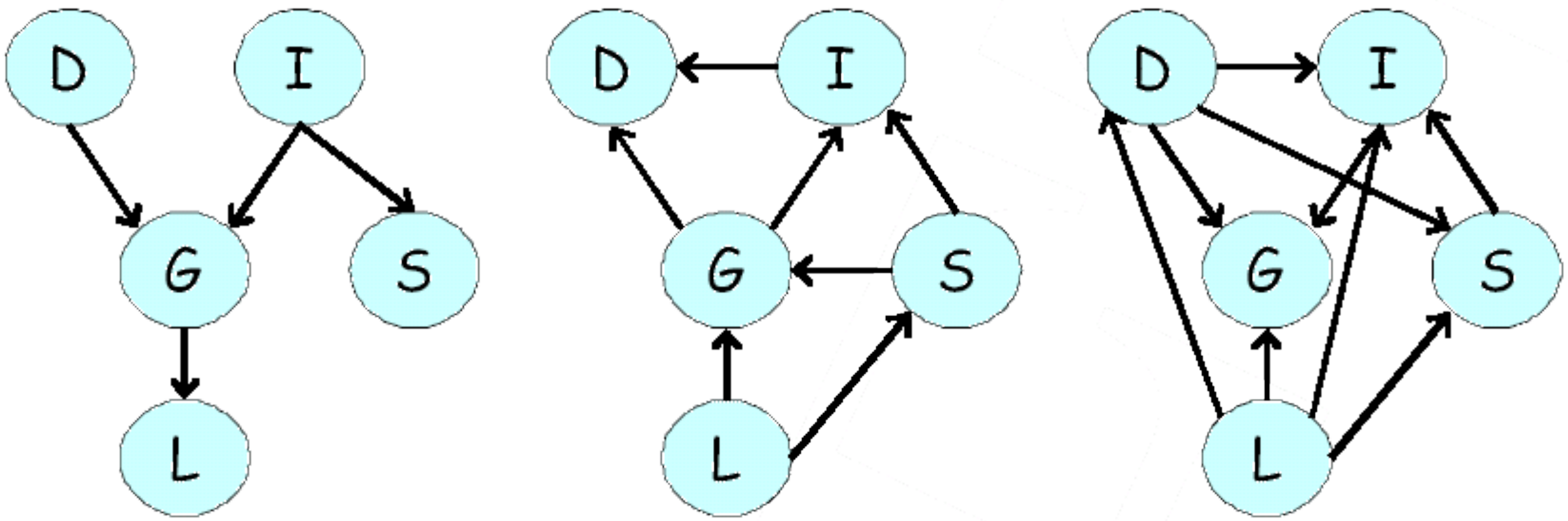
- **What do we gain?**
- **What do we lose?**

Minimum I-MAP

- Complete graph is a (trivial) I-map for any distribution, yet it does not reveal any of the independence structure in the distribution.
  - Meaning that the graph dependence is arbitrary, thus by careful parameterization an dependencies can be captured
  - We want a graph that has the maximum possible I(G), yet still $\subseteq$ I(P)

- **Defn :** A graph object G is a *minimal I-map* for a set of independencies I if it is an I-map for I, and if the removal of even a single edge from G renders it not an I-map.

Minimum I-MAP is not unique



(a) (b) (c)

Summary of BN semantics

- **Defn :** A *Bayesian network* is a pair (G, P) where P factorizes over G, and where P is specified as 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?