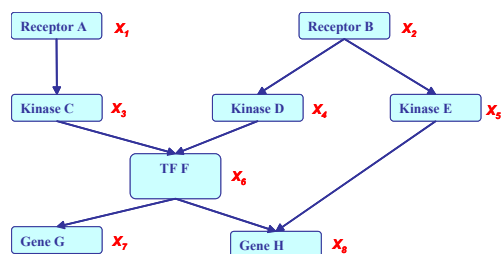
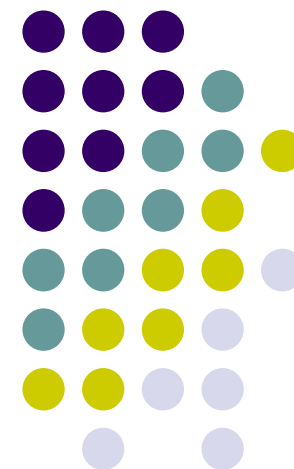




Probabilistic Graphical Models

Directed GMs: Bayesian Networks



Eric Xing

Lecture 2, January 23, 2017

Reading: see class homepage

Two types of GMs

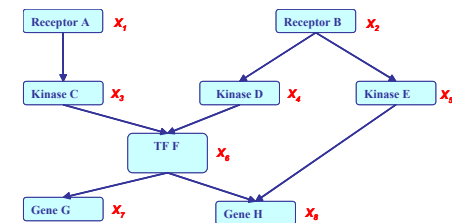


- **Directed edges** give **causality** relationships (Bayesian Network or Directed Graphical Model):

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

$$= P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2)$$

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

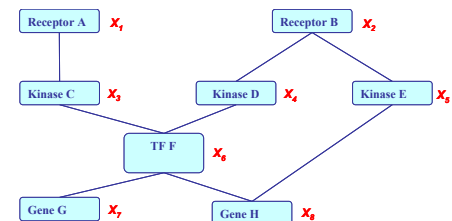


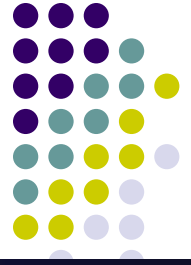
- **Undirected edges** simply give **correlations** between variables (Markov Random Field or Undirected Graphical model):

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

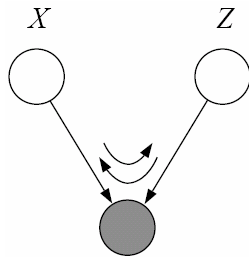
$$= \frac{1}{Z} \exp\{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2)$$

$$+ E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\}$$





- 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

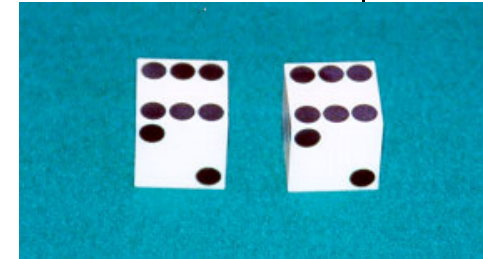
$$P(1) = P(2) = P(3) = P(5) = 1/10$$

$$P(6) = 1/2$$

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



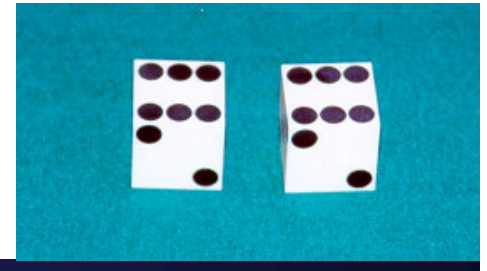
GIVEN: A sequence of rolls by the casino player

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
 - Discrete
 - Continuous
- **Picking structure**
 - CAUSAL
 - Generative
 - Coupling
- **Picking Probabilities**
 - “Natural”
 - Zero probabilities
 - Orders of magnitudes
 - Relative values

Hidden Markov Model

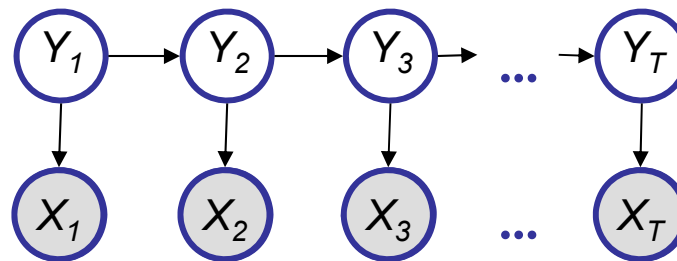


The underlying source:

Speech signal
genome function
dice

The sequence:

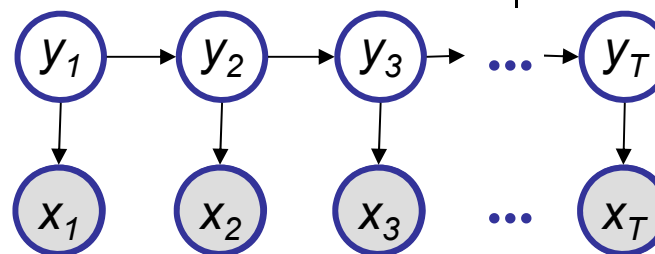
Phonemes
DNA sequence
sequence of rolls



Probability of a parse



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



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

- Marginal probability: $p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) = \sum_{y_1} \sum_{y_2} \dots \sum_{y_T} \pi_{y_1} \prod_{t=2}^T a_{y_{t-1}, y_t} \prod_{t=1}^T p(x_t | y_t)$
- Posterior probability: $p(\mathbf{y} | \mathbf{x}) = p(\mathbf{x}, \mathbf{y}) / p(\mathbf{x})$

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



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. In other words, each variable is a stochastic function of its parents.



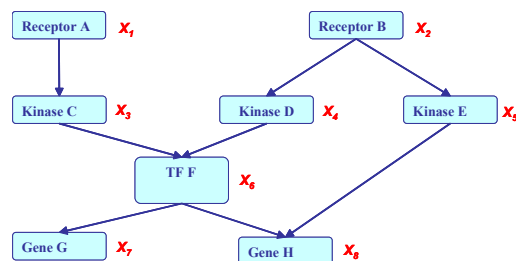
Bayesian Network: Factorization Theorem

- **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”:

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

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

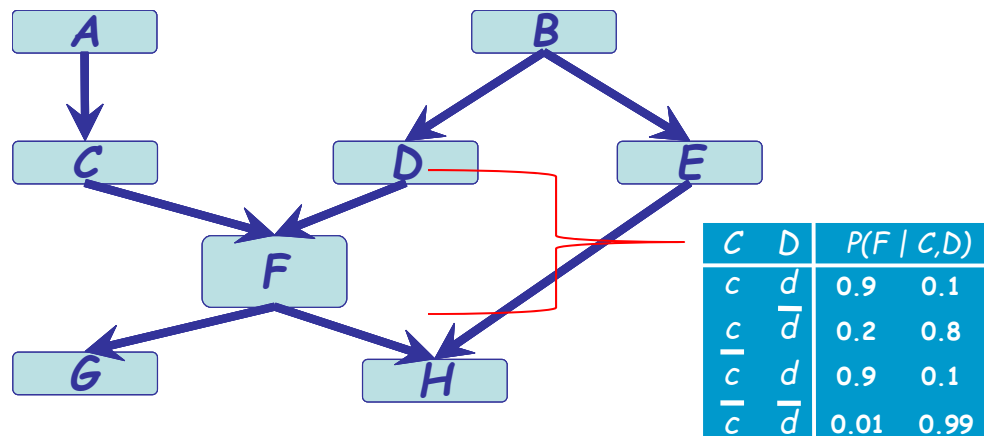


$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &\Rightarrow = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) \\ &\quad P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \end{aligned}$$



Specification of a directed GM

- There are two components to any GM:
 - the *qualitative* specification
 - the *quantitative* specification

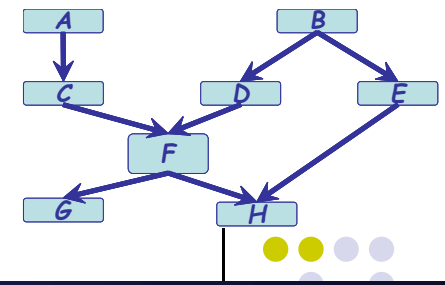




Qualitative Specification

- 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)
 - ...

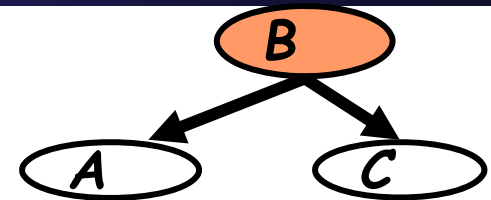
Local Structures & Independencies



- 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

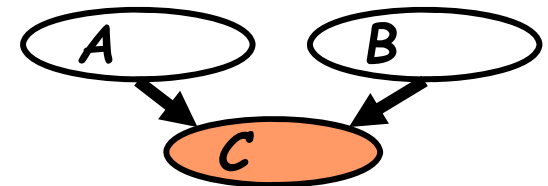
"given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



- V-structure

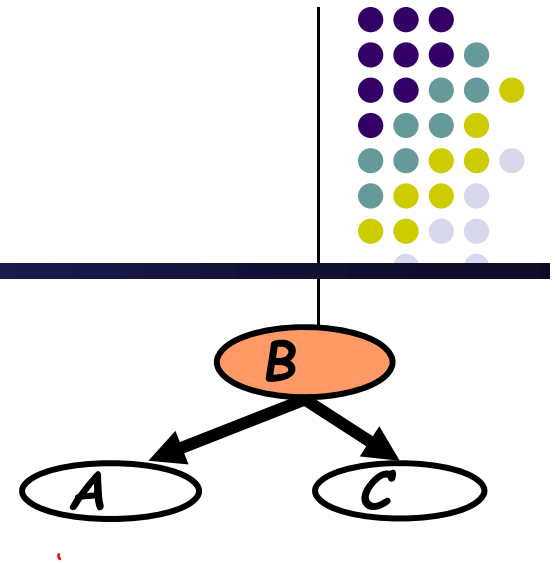
- Knowing C couples A and B
because A can "explain away" B w.r.t. C

"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 justification



I-maps



- **Defn** : Let P be a distribution over 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 , if $I(K) \subseteq I$.
- 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.

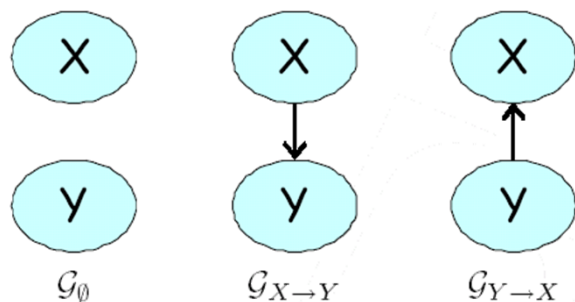


Facts about I-map

- For G 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

What is in $I(G)$ --- local Markov assumptions of BN



A *Bayesian network structure* G is a directed acyclic graph whose nodes represent random variables $X_1, \dots, X_n$.

local Markov assumptions

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

$$I_\ell(G): \{X_i \perp NonDescendants_{X_i} \mid Pa_{X_i} : \forall i\},$$

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

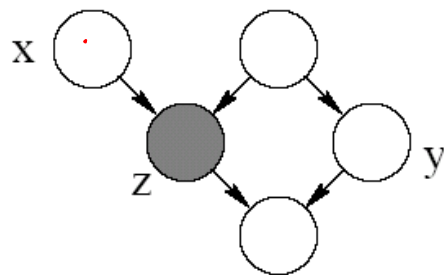


Graph separation criterion

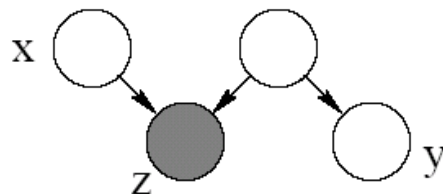
- D-separation criterion for Bayesian networks (D for Directed edges):

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

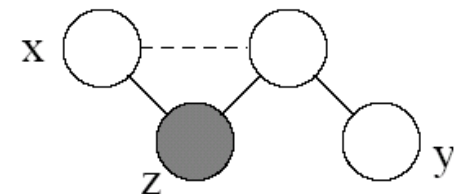
- Example:



original graph



ancestral



moral ancestral



Active trail

- **Causal trail** $X \rightarrow Z \rightarrow Y$: active if and only if Z is not observed.
- **Evidential trail** $X \leftarrow Z \leftarrow Y$: active if and only if Z is not observed.
- **Common cause** $X \leftarrow Z \rightarrow Y$: active if and only if Z is not observed.
- **Common effect** $X \rightarrow Z \leftarrow Y$: active if and only if either Z or one of Z 's descendants is observed

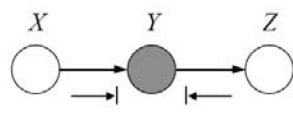
Definition : Let X, Y, Z be three **sets** of nodes in G . We say that X and Y are *d-separated given Z* , denoted $d\text{-sep}_{\mathcal{G}}(X; Y \mid Z)$, if there is **no** active trail between any node $X \in X$ and $Y \in Y$ given Z .

What is in $I(G)$ ---

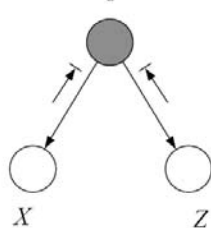
Global Markov properties of BN



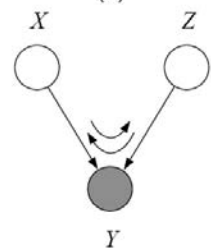
- X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "**Bayes-ball**" algorithm illustrated below (and plus some boundary conditions):



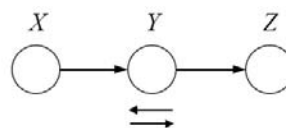
(a)



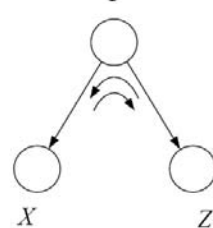
(a)



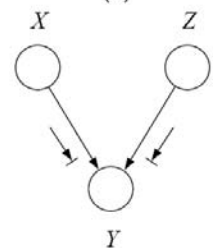
(a)



(b)



(b)



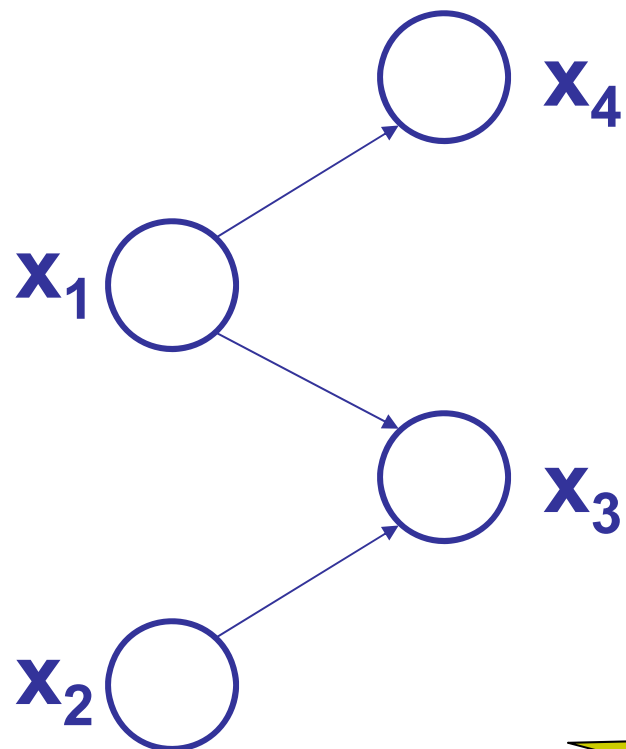
(b)

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

$$I(G) = \{X \perp Z | Y : \text{dsep}_G(X; Z | Y)\}$$

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

Example:



- Complete the $I(G)$ of this graph:

**Scriber please fill in
the rest of this slide !**

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 ,

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

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

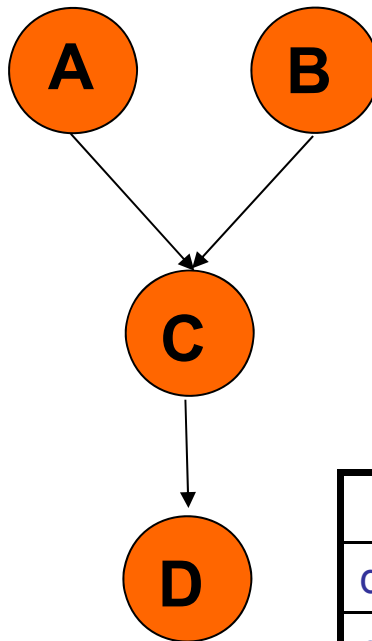
Conditional probability tables (CPTs)



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

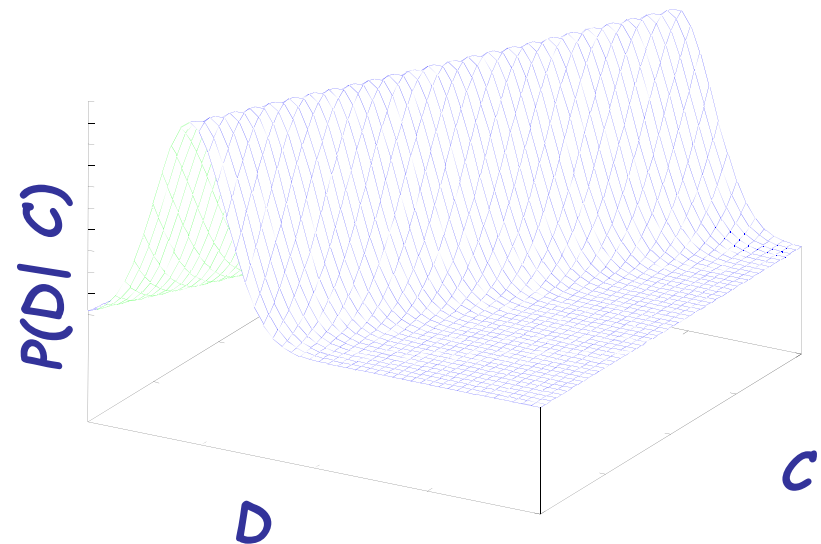
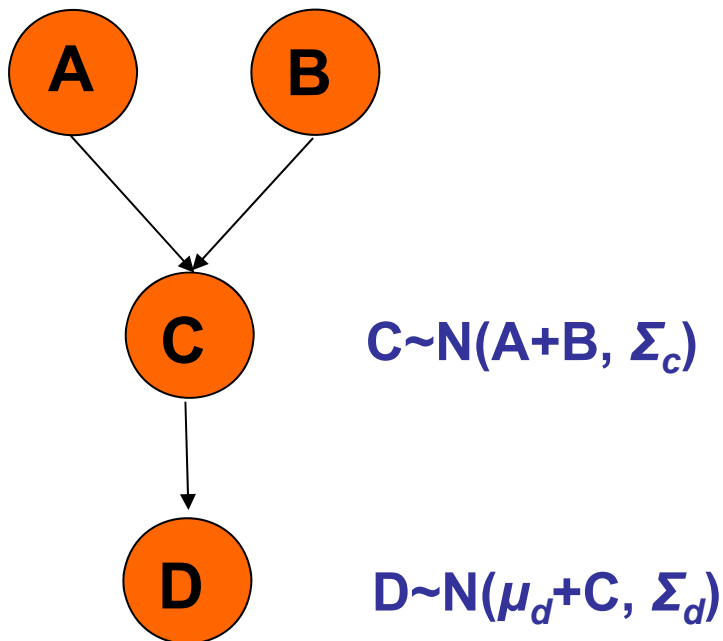
	c^0	c^1
d^0	0.3	0.5
d^1	0.7	0.5

Conditional probability density func. (CPDs)



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

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



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 independencies** that hold for every distribution P that factorizes over G ?



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 $d\text{-sep}_G(X; Y \mid Z)$.

Contrapositive of the completeness statement

- "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?

Distributional equivalence and I-equivalence

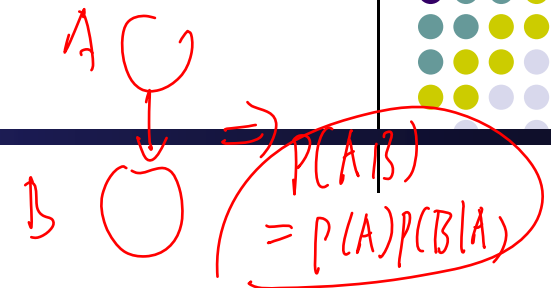
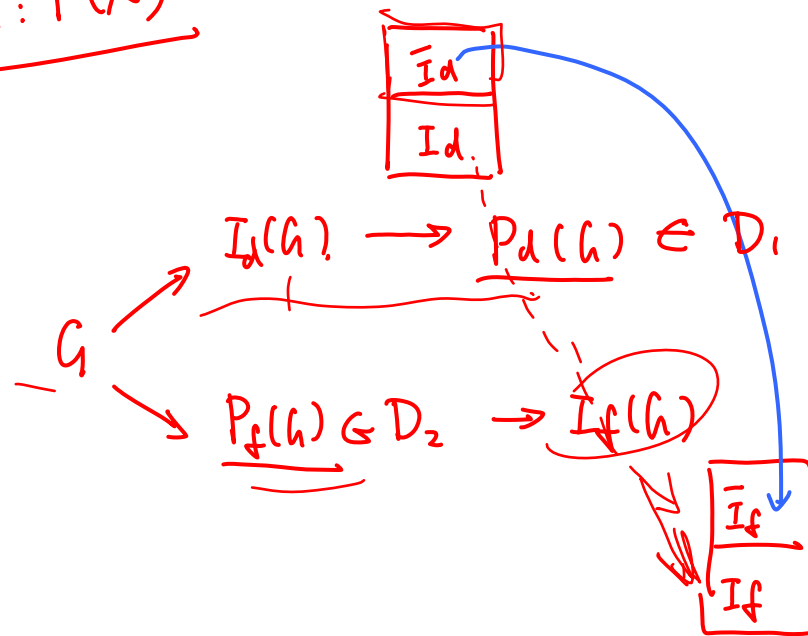


- All independence in $I_d(G)$ will be captured in $I_f(G)$, is the reverse true?
- Are "not-independence" from G all honored in P_f ?

Distributional equivalence and I-equivalence



$X: P(X)$



$$X \perp Y | Z \in I_d$$



$$X \perp Y | Z \in I_f.$$

- All independence in $I_d(G)$ will be captured in $I_f(G)$, is the reverse true?
- Are "not-independence" from G all honored in P_f ?



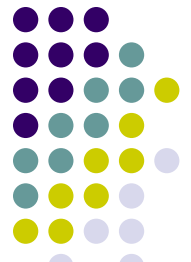
Soundness and completeness

- Contrapositive of the completeness statement
 - "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?
- No. Even if a distribution factorizes over G , it can still contain **additional independencies** that are not reflected in the structure
 - Example: graph $A \rightarrow B$, for actually independent A and B (the independence can be captured by some subtle way of parameterization)
- **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 .

A	b^0	b^1
a^0	0.4	0.6
a^1	0.4	0.6

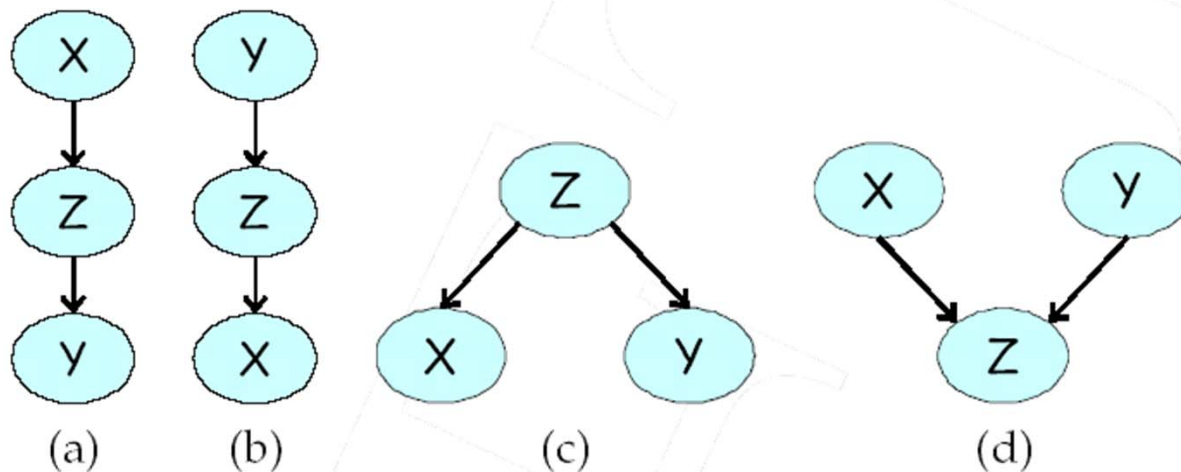


- **Theorem** : For **almost all** distributions P that factorize over G , i.e., for all distributions except for a set of "measure zero" in the space of CPD parameterizations, we have that $I(P) = I(G)$



Uniqueness of BN

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

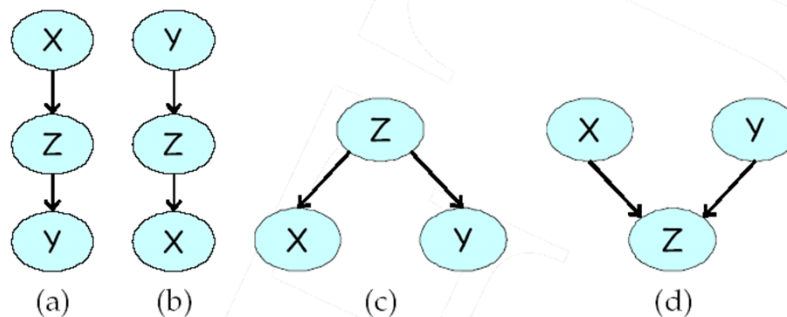


$(X \perp Y \mid Z)$.

I-equivalence



- **Defn** : Two BN graphs $G1$ and $G2$ over X are *I-equivalent* if $I(G1) = I(G2)$.
- The set of all graphs over X is partitioned into a set of mutually exclusive and exhaustive *I-equivalence classes*, which are the set of equivalence classes induced by the I-equivalence relation.

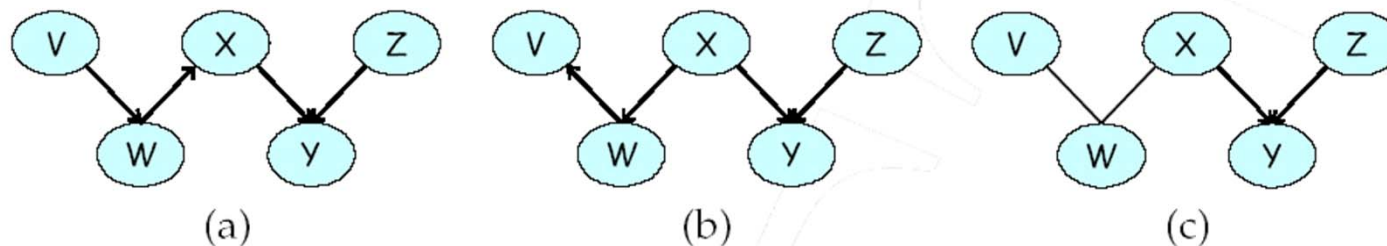


- Any distribution P that can be factorized over one of these graphs can be factorized over the other.
- Furthermore, there is no intrinsic property of P that would allow us associate it with one graph rather than an equivalent one.
- This observation has important implications with respect to our ability 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 .



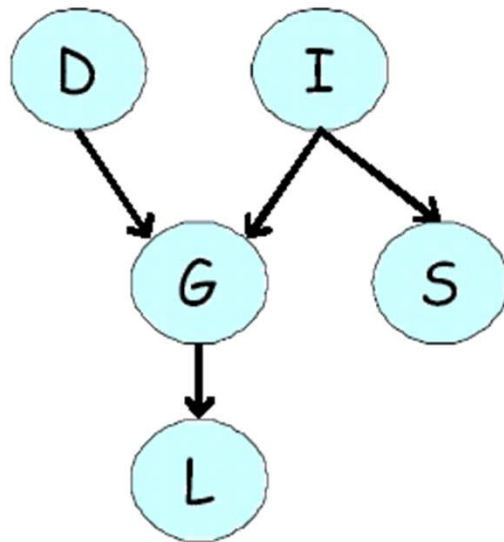
- **Thm** : Let G_1 and G_2 be two graphs over V . If G_1 and G_2 have the same skeleton and the same set of v-structures then they are I-equivalent.
- graph equivalence
- Same trail
- But not necessarily active



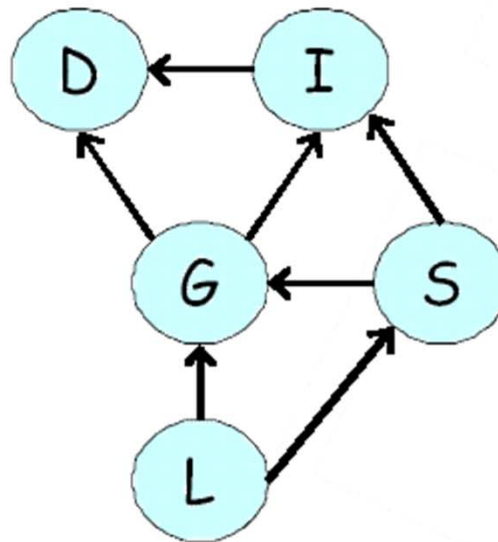
Minimum I-MAP

- 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.

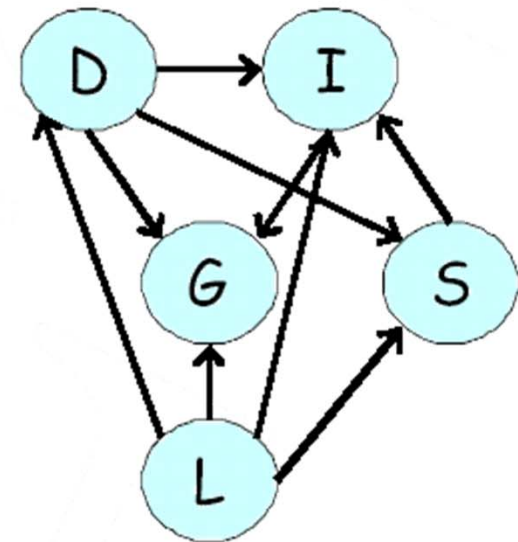
Minimum I-MAP is not unique



(a)

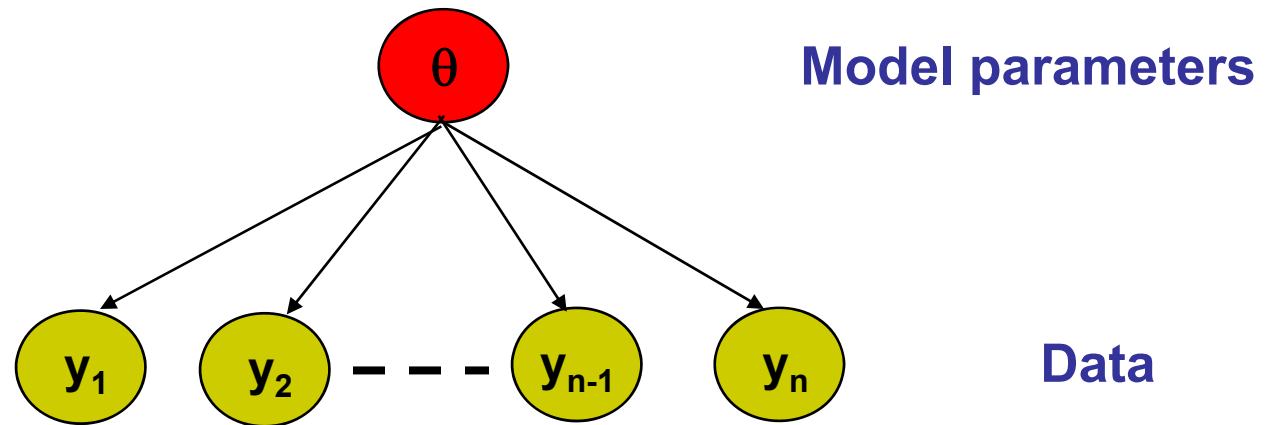


(b)

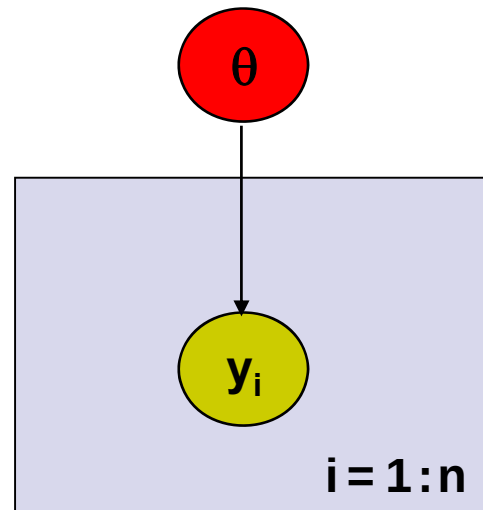


(c)

Simple BNs: Conditionally Independent Observations



The “Plate” Micro



Model parameters

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

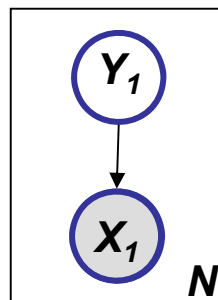
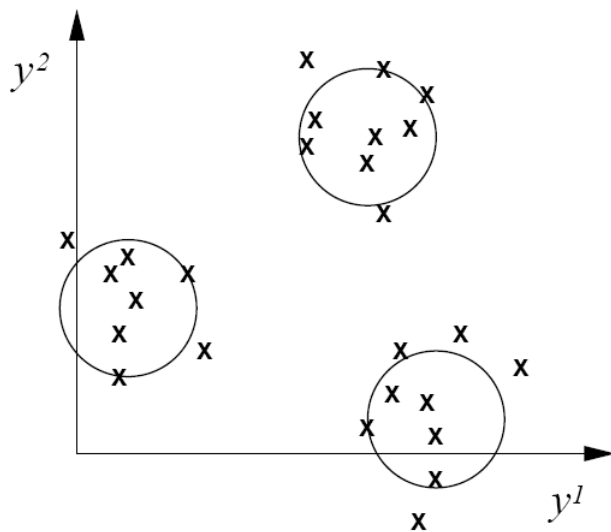
Plate = rectangle in graphical model

**variables within a plate are replicated
in a conditionally independent manner**

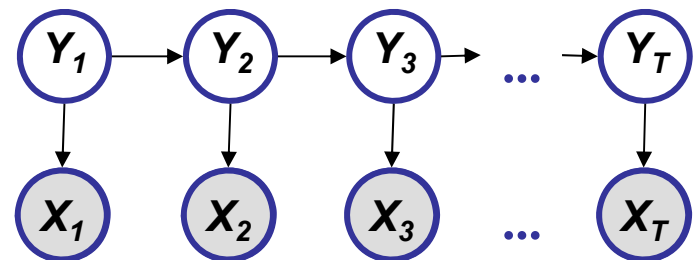
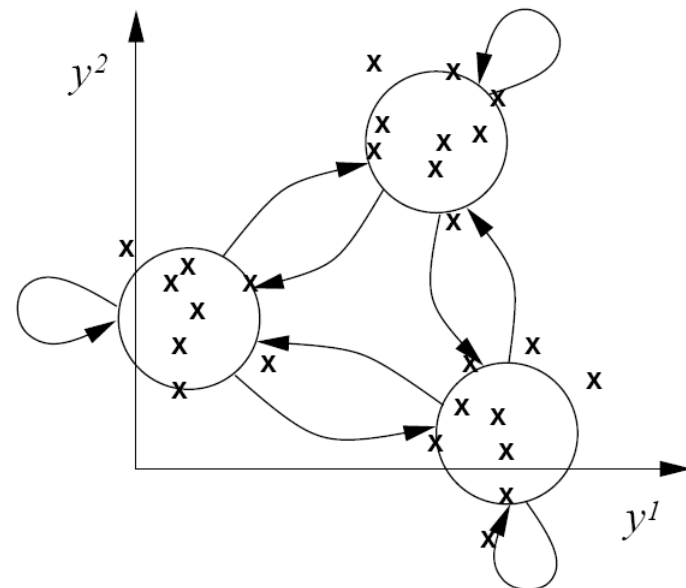
Hidden Markov Model: from static to dynamic mixture models



Static mixture



Dynamic mixture



Definition (of HMM)



- **Observation space**

Alphabetic set: $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$

Euclidean space: $\mathbb{R}^d$

- **Index set of hidden states**

$$\mathcal{I} = \{1, 2, \dots, M\}$$

- **Transition probabilities** between any two states

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

or $p(y_t | y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,1}, \dots, a_{i,M}), \forall i \in \mathcal{I}.$

- **Start probabilities**

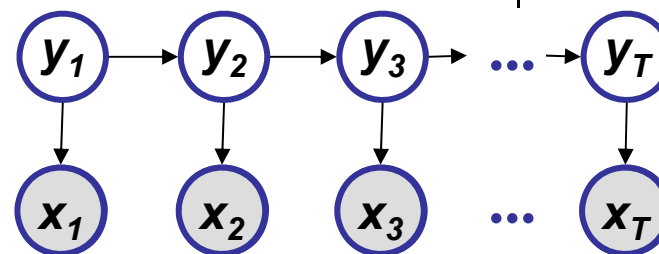
$$p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \dots, \pi_M).$$

- **Emission probabilities** associated with each state

$$p(x_t | y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,1}, \dots, b_{i,K}), \forall i \in \mathcal{I}.$$

or in general:

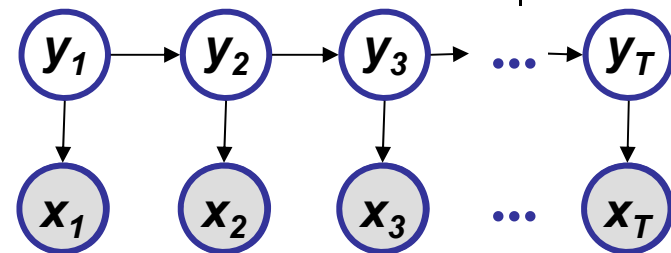
$$p(x_t | y_t^i = 1) \sim f(\cdot | \theta_i), \forall i \in \mathcal{I}.$$



Probability of a parse



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



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

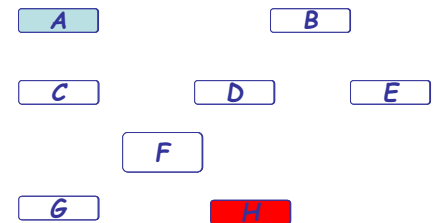
Summary: Representing Multivariate Distribution



- 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

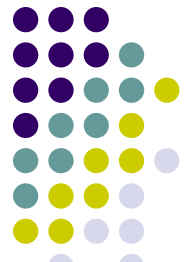
$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &= P(X_1)P(X_2 | X_1)P(X_3 | X_1, X_2)P(X_4 | X_1, X_2, X_3)P(X_5 | X_1, X_2, X_3, X_4)P(X_6 | X_1, X_2, X_3, X_4, X_5) \\ &\quad P(X_7 | X_1, X_2, X_3, X_4, X_5, X_6)P(X_8 | X_1, X_2, X_3, X_4, X_5, X_6, X_7) \end{aligned}$$

- 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?



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 ✗