**10-708: Probabilistic Graphical Models, Spring 2020**

# 3: Directed Graphical Models (Bayesian Networks)

*Lecturer: Eric P. Xing Scribe: Suman Pokharel, Wendy Yang, Donghui Yan, Jingjing Tang, Wenhuan Sun*

## 1 Two Types of Graphical Models

### 1.1 Directed Graphical Models

In directed graphical models, nodes that represent random variables are connected by directed edges, which represent causality relationships between nodes. This type of directed GM is called Bayesian Network or Directed Graphical Model.

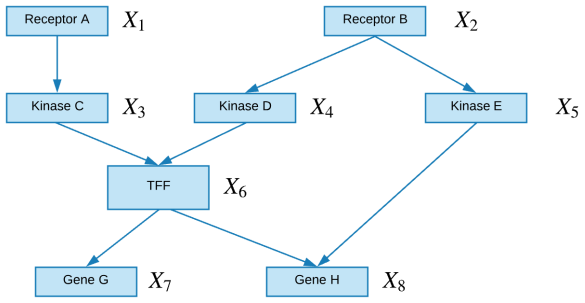



Figure 1: Directed Graph

For example, in the Directed GM above, the underlying joint probability can be written as:

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

### 1.2 Undirected Graphical Models

In undirected graphical models, nodes are connected by undirected edges, which represents correlations between nodes/variables. This type of GM is called Markov Random Field or Undirected Graphical model.

For example, in the undirected graphical models below, the underlying joint probability can be written as $P(X_1, X_2, ..., X_7, X_8) = 1/Zexp\{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) + E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\}$.

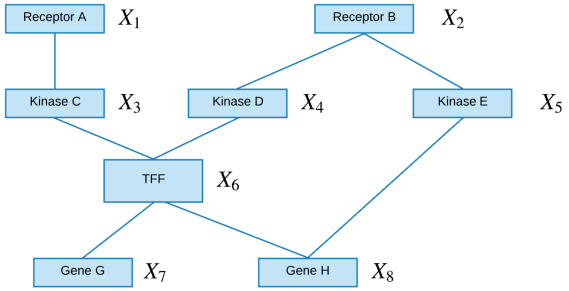



Figure 2: Undirected Graph

# 2 Examples

## 2.1 Expert Systems

Expert systems are good example of the application of directed graphical models, where the expert knowledge will be encoded as directed edges between nodes. For example, according to the ALARM network, a patient monitoring system, representing causal relationships presented by Beinlich et al. 1989, expert medical knowledge was encoded as directed edges between nodes that represent random variables (e.g. measurements (blood pressure, heart rate, respiratory rate, etc) and queries (presence of a disease)). In this case, inference is easier with such directed graphical model (e.g. P(kinked tube = true — measurements)).

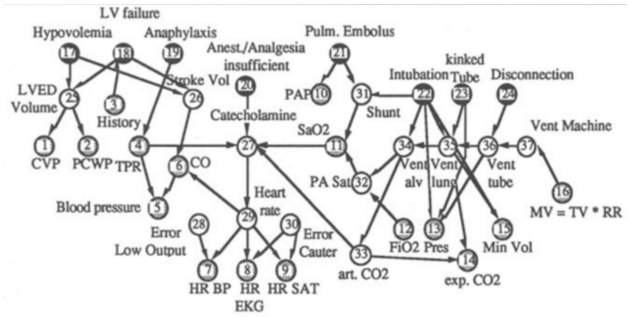



Figure 3: The ALARM network representing causal relationships, Beinlich et al. 1989

## 2.2 Dishonest Casino

Domain knowledge and knowledge engineering: Let discrete random variable $x$ be the outcome of dice-rolling taking values from $[1, 2, 3, 4, 5, 6]$ and categorical random variable $y$ be the choice of the dice taking values from [fair, loaded]. In this case, $x$ is an observed discrete variable, $y$ is a hidden categorical variable. There exists some causal relationships between $x_i$ and $y_i$, as well as between different $y_i$. For example, $P(x_i|y_i = fair)$ and $P(y_{i+1}|y_i)$

Two dices were used in a casino. The fair one has equal probability for each number ($P(x = i) = \frac{1}{6}, i \in [1, 6]$), and the loaded/unfair one has the following probability distribution: $P(x = i) = \frac{1}{10}$ if $i \neq 6$, else 0.5.

Given a sequence of 50 observed rolls (e.g. 1245.....666...2344), 3 types of question could be asked:

1. Evaluation: how likely is this sequence, given our knowledge/model of how the casino works? (e.g. P(observed sequence | domain knowledge))

2. Decoding: what portion of the sequence was generated with the fair die, and what portion with the loaded die? (e.g. P(choice of dice | observed sequence))

3. Learning: 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? (e.g. P(choice of dice))

A Hidden Markov Model can be used to model this casino problem, where the sequence is $\mathbf{x} = x_1, x_2, ..., x_T$ and the parse $\mathbf{y} = y_1, y_2, ..., y_T$.

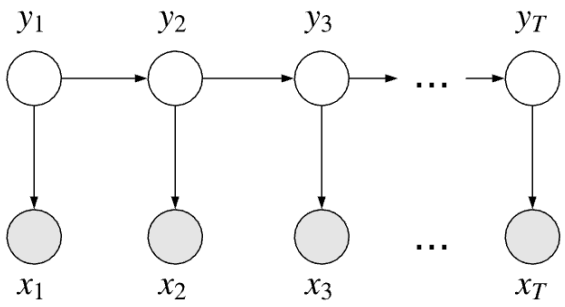



Figure 4: Hidden Markov Model for Casino

The probability of this parse can be answered as:

**Joint probability** $p(\mathbf{x},\mathbf{y}) = p(x_1, x_2, ..., x_T, y_1, y_2, ..., y_T) = p(y_1) \prod_{t=2}^{T} p(y_t|y_{t-1}) \prod_{t=1}^{T} p(x_t|y_t)$

$= p(y_1, y_2, ..., y_T)p(x_1, x_2, ..., x_T|y_1, y_2, ..., y_T)$

**Marginal probability** $p(\mathbf{x}) = \sum_y p(\mathbf{x},\mathbf{y}) = \sum_{y_1} \sum_{y_2} ... \sum_{y_N} \pi_{y_1} \prod_{t=2}^{T} a_{y_{t-1},y_t} \prod_{t=1}^{T} p(x_t|y_t))$

**Posterior probability** $p(\mathbf{y}|\mathbf{x}) = p(\mathbf{x},\mathbf{y})/p(\mathbf{x})$

In the case of calculating marginal probability, $k^N$ summation operations are needed, where $k$ represents the number of possible value of the categorical variable $y$. This could be improved to polynomial time.

# 3 Bayesian Networks

- A BN is a directed graph model whose nodes represent the random variables and whose edges represent directed influence among or between random variables.

- It is a data structure that provides the skeleton for representing **a joint distribution** compactly in a **systematic 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 the nature using a distribution that depends only on its parents.

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

## 3.1 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 givens 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.

Example:

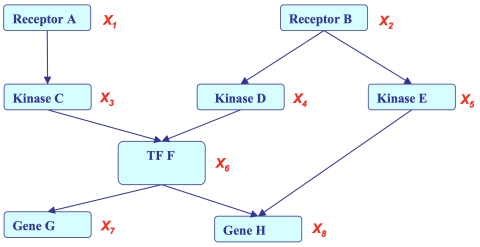



Figure 5: Directed Graph

The joint probability of the above directed graph can be written as:

$$P(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6, \mathbf{X}_7, \mathbf{X}_8) =$$
$$P(\mathbf{X}_1)P(\mathbf{X}_2)P(\mathbf{X}_3|\mathbf{X}_1)P(\mathbf{X}_4|\mathbf{X}_2)P(\mathbf{X}_5|\mathbf{X}_2)P(\mathbf{X}_6|\mathbf{X}_3, \mathbf{X}_4)P(\mathbf{X}_7|\mathbf{X}_6)P(\mathbf{X}_8|\mathbf{X}_5, \mathbf{X}_6)$$

## 3.2 Specification of a directed GM

There are two components to any GM:

- the **qualitative** specification
- the **quantitative** specification

### 3.2.1 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 like a certain architecture (e.g. a layered graph)
- ...

Is there a concise, unambiguous, and rigorous meaning/interpretation?

- Conditional independence between variables!

## 3.3 Local Structures and Independence

- Common parents
  - Fixing B **decouples** A and C
  - "Given the level of gene B, the levels of A and C are independent"
  - Expression: $P(A, C|B) = P(A|B)P(C|B)$

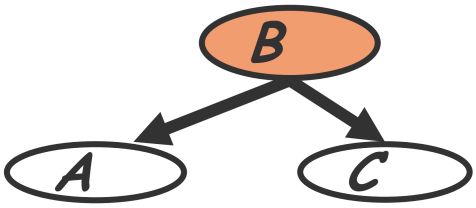



Figure 6: Common Parent Example

- Cascade
  - Knowing B **decouples** A and C
  - "Given the level of gene B, the levels of A provides no extra prediction value for the level of gene C"
  - Expression: $P(A, B, C) = P(A)P(B|A)P(C|B)$


Figure 7: Cascade Example

- V-structure
  - Knowing C **decouples** 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"
  - Expression: $P(A, B) = P(A)P(B)P(A, B|C)$
- The language is compact, the concept are rich!

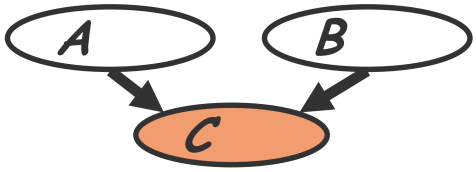
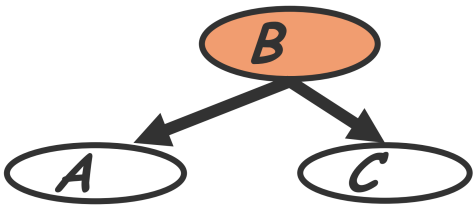



Figure 8: V-structure Example

### 3.4 A simple justification




Equation:
$P(A, C|B) = \frac{P(A,B,C)}{P(B)} = \frac{P(B)P(A|B)P(C|B)}{P(B)} = P(A|B)P(C|B)$

## 4 I-Maps

We use I-maps to establish the relationship between graph and distribution. A distribution $\mathcal{P}$ satisfies the local independencies associated with a graph $\mathcal{G}$, if and only if $\mathcal{P}$ is representable as a set of Conditional Probability Distributions (CPDs) associated with the graph $\mathcal{G}$.

Independencies associated with a distribution $\mathcal{P}$

**Definition:** Let $\mathcal{P}$ be a distribution over $\mathcal{X}$. We define $\mathcal{I}(\mathcal{P})$ to be the set of independence assertions of the form $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$ that hold in $\mathcal{P}$.

I-Map

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

**Corollary:** $\mathcal{G}$ is an I-map for $\mathcal{P}$ if $\mathcal{G}$ is an I-map for $\mathcal{I}(\mathcal{P})$, where we use $\mathcal{I}(\mathcal{G})$ as set of independencies associated

### 4.1 Facts about I-maps

For $\mathcal{G}$ to be I-map of $\mathcal{P}$, it is necessary that $\mathcal{G}$ does not mislead us regarding independencies in $\mathcal{P}$:

any independence that $\mathcal{G}$ asserts must also hold in $\mathcal{P}$. Conversely, $\mathcal{P}$ may have additional independencies that are not reflected in $\mathcal{G}$

**Example**

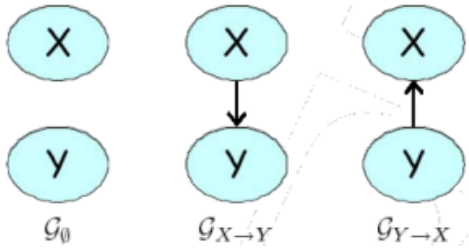



Figure 10: I-map example

X and Y are independent in graph $\mathcal{G}_\emptyset$ only.

**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 |

In $P_1 : P(X,Y) = P(X)P(Y)$, i.e. $0.6 \times 0.8 = 0.48$. Hence, $I(P_1) = X \perp Y$ and is shown by graph $\mathcal{G}_\emptyset$

In $P_2 : P(X,Y) \neq P(X)P(Y)$. Hence, $I(P_1) = \emptyset$ and is shown by both graphs $\mathcal{G}_{\mathcal{X}\to\mathcal{Y}}$ and $\mathcal{G}_{\mathcal{Y}\to\mathcal{X}}$

## 4.2 What is $\mathcal{I}(\mathcal{G})$

### 4.2.1 Local Markovian Assumptions of Bayesian Network

A Bayesian Network structure $\mathcal{G}$ is a directed acyclic graph (DAG) whose nodes represent random variables $X_1, X_2, ..., X_N$.

**Definition:**

Let $Pa_{X_i}$ denote the parents of $X_i$ in $\mathcal{G}$, and $NonDescendants_{X_i}$ denote the variable in the graph that are non-descendants of $X_i$. Then $\mathcal{G}$ encodes the following set of local conditional independence assumptions $I_\ell(\mathcal{G})$ :

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

Each node $X_i$ is independent of its non-descendants given its parents.

**Graph Separation Criterion**

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

**Definition:**

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

**Example**

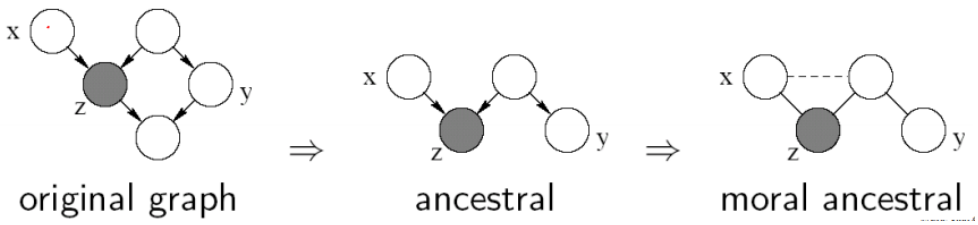



Figure 11: D-separation criterion example

Construct the ancestral graph by removing all nodes except the random variables of interest and their ancestors. Then perform moralization on ancestral graph by removing all directions on edges and connecting nodes that are originally unconnected and have a common child node.

The example shows conditional independence. If there is a way to travel from one node to another node using any path (not through the given), then those two nodes are not conditionally independent.

If $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$, then we say $\mathbf{Z}$ D-separates $\mathbf{X}$ and $\mathbf{Y}$.

### 4.2.2 Global Markovian Assumptions of Bayesian Network

Practical definition of $\mathcal{I}(\mathcal{G})$

$\mathbf{X}$ is D-separated from $\mathbf{Z}$ given $\mathbf{Y}$ if we can't send a ball from any node in $\mathbf{X}$ to any node in $\mathbf{Z}$ using the *"Bayes-ball"* algorithm illustrated by the following examples (plus some boundary conditions):

**Example**

Causal Trail: $\mathbf{X} \rightarrow \mathbf{Z} \rightarrow \mathbf{Y}$, active $\iff$ $\mathbf{Z}$ is not observed.

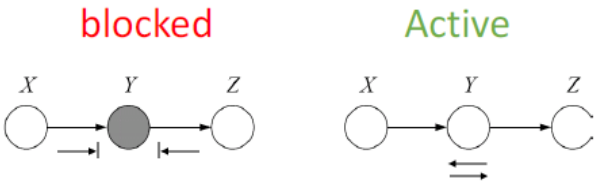



Figure 12: Causal Trail

Common cause: $\mathbf{X} \leftarrow \mathbf{Z} \rightarrow \mathbf{Y}$, active $\iff$ $\mathbf{Z}$ is not observed.

Common effect: $\mathbf{X} \rightarrow \mathbf{Z} \leftarrow \mathbf{Y}$, active $\iff$ either $\mathbf{Z}$ or one of $\mathbf{Z}$'s descendents is observed.

**Definition:**

All independence properties that correspond to d-separation:

$$\mathcal{I}(\mathcal{G}) = \{X \perp Z|Y : dsep_{\mathcal{G}}(X; Z|Y)\}$$

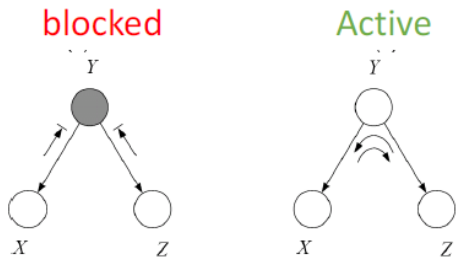



Figure 13: Common Cause

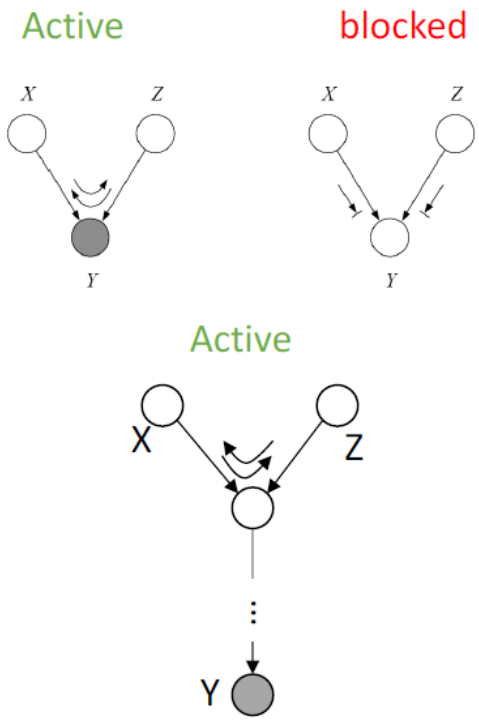



Figure 14: Common Effect

**Example**

To find I(G) for the graph below, we try all possible trails and see what conditional independence we can draw from the trail.

For example, for trail $X_4 \leftarrow X_1 \rightarrow X_3$, the trail is not active only if $X_1$ is observed based on the common cause structure. So we get $X_4 \perp X_3|X_1$, $X_4 \perp X_3|\{X_1, X_2\}$.

Similarly, we can find $\{X_1 \perp X_2, X_1 \perp X_2|X_4\}$ from trail $X_1 \rightarrow X_3 \leftarrow X_2$ and $\{X_2 \perp X_4, X_2 \perp \{X_1, X_4\}$, $X_2 \perp X_4|\{X_1, X_2\}$, $X_2 \perp X_4|\{X_1, X_3\}\}$ from trail $X_4 \leftarrow X_1 \rightarrow X_3 \leftarrow X_2$

Thus,

$$I(G) = \{X_4 \perp X_3|X_1, X_4 \perp X_3|\{X_1, X_2\}, X_1 \perp X_2, X_1 \perp X_2|X_4,$$
$$X_2 \perp X_4, X_2 \perp \{X_1, X_4\}, X_2 \perp X_4|X_1, X_2 \perp X_4|\{X_1, X_3\}\}$$

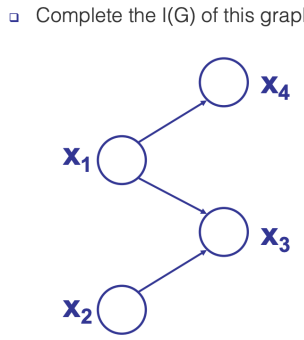



Figure 15: I(G) example

## 4.3 Equivalence Theorem

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

Formally, for any graph G

let $D_1$ denote the family of all distributions that satisfy I(G), $D_2$ denote the family of all distributions that factor according to G.

Then we have $D_1 \equiv D_2$. The two families are the same.

In other words, when building the distribution, we can directly use the factorization law to assemble a distribution mechanically by $P(X) = \prod_{i=1:d} P(X_i|X_{\pi_i})$.

## 4.4 Conditional probability tables (CPTs)

To build the joint distribution for the graph with discrete random variables below, we can use conditional probability tables.

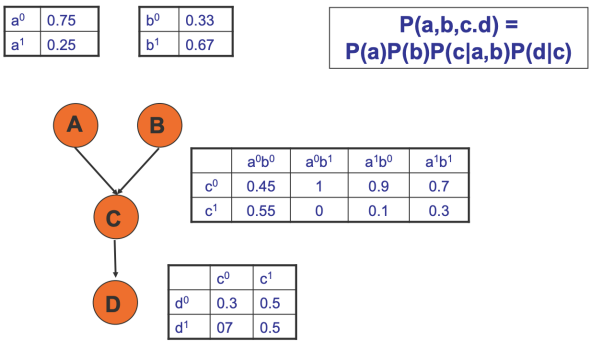



Figure 16: CPT example

### 4.5 Conditional probability density (CPDs)

To build the joint distribution for the graph with continuous random variables below, we can use conditional probability density functions. Here is an example of defining a continuous random variable dependent on other continuous random variables.

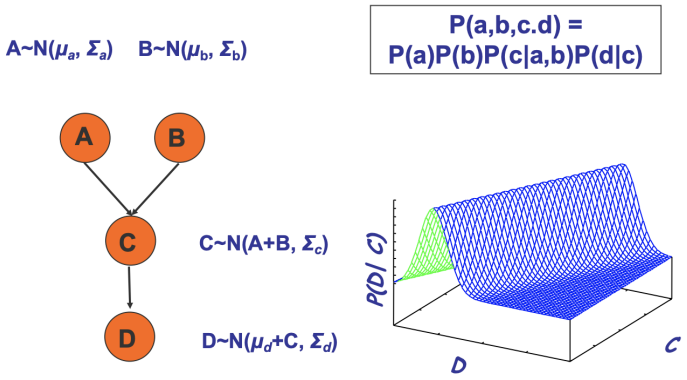



Figure 17: CPD example

### 4.6 Summary of BN semantics

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

## 5 Soundness and completeness

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

- **Completeness**
  Claim: For any distribution P that factorizes over G, if $(X \perp Y|Z) \in I(P)$ then $d - sep_G(X;Y|Z)$ (not guaranteed)

- Contrapositive of the completeness statement
  - If and Y are not d-separated given Z in G, are X and Y guaranteed to be dependent in all distributions P that factorize over G?
  - No. Even if a distribution factorizes over G, it can still contain additional independencies that are not reflected in the structure.
  - Example: Consider graph A → B, which indicates that A and B are dependent. However, the conditional distribution of B—A could be arbitrarily picked so that A and B are actually independent. (The independence can be captured by some subtle way of pasteurization)
  - **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.

| $A$ | $b^0$ | $b^1$ |
|---|---|---|
| $a^0$ | 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 $I(P) = I(G)$