Probabilistic Graphical Models

The Big Objective(s)

In a wide variety of application fields two main problems need to be addressed over and over:

1. **How can (expert) knowledge of complex domains be efficiently represented?**
2. **How can inferences be carried out within these representations?**
3. **How can such representations be (automatically) extracted from collected data?**

We will deal with all three questions during the lecture.

Example 1: Planning in car manufacturing

Available information

"Engine type e_1 can only be combined with transmission t_2 or t_5."

"Transmission t_5 requires crankshaft c_2."

"Convertibles have the same set of radio options as SUVs."

Possible questions/inferences:

"Can a station wagon with engine e_4 be equipped with tire set y_6?"

"Supplier S_8 failed to deliver on time. What production line has to be modified and how?"

"Are there any peculiarities within the set of cars that suffered an aircondition failure?"

Example 2: Medical reasoning

Available information:

"Malaria is much less likely than flu."

"Flu causes cough and fever."

"Nausea can indicate malaria as well as flu."

"Nausea never indicated pneunomia before."

Possible questions/inferences

"The patient has fever. How likely is he to have malaria?"

"How much more likely does flu become if we can exclude malaria?"

Common Problems

Both scenarios share some severe problems:

Large Data Space
It is intractable to store all value combinations, i. e. all car part combinations or inter-disease dependencies.

(Example: VW Bora has 10^{200} theoretical value combinations*)

Sparse Data Space
Even if we could handle such a space, it would be extremely sparse, i. e. it would be impossible to find good estimates for all the combinations.

(Example: with 100 diseases and 200 symptoms, there would be about 10^{62} different scenarios for which we had to estimate the probability.*)

* The number of particles in the observable universe is estimated to be between 10^{78} and 10^{85}.

Idea to Solve the Problems

Given: A large (high-dimensional) distribution δ representing the domain knowledge.

Desired: A set of smaller (lower-dimensional) distributions $\{\delta_1, \ldots, \delta_s\}$ (maybe overlapping) from which the original δ *could* be reconstructed with no (or as few as possible) errors.

With such a decomposition we can draw any conclusions from $\{\delta_1, \ldots, \delta_s\}$ that could be inferred from δ — without, however, actually reconstructing it.

Example: Car Manufacturing

Let us consider a car configuration is described by three attributes:

- Engine E, $\mathrm{dom}(E) = \{e_1, e_2, e_3\}$
- Breaks B, $\mathrm{dom}(B) = \{b_1, b_2, b_3\}$
- Tires T, $\mathrm{dom}(T) = \{t_1, t_2, t_3, t_4\}$

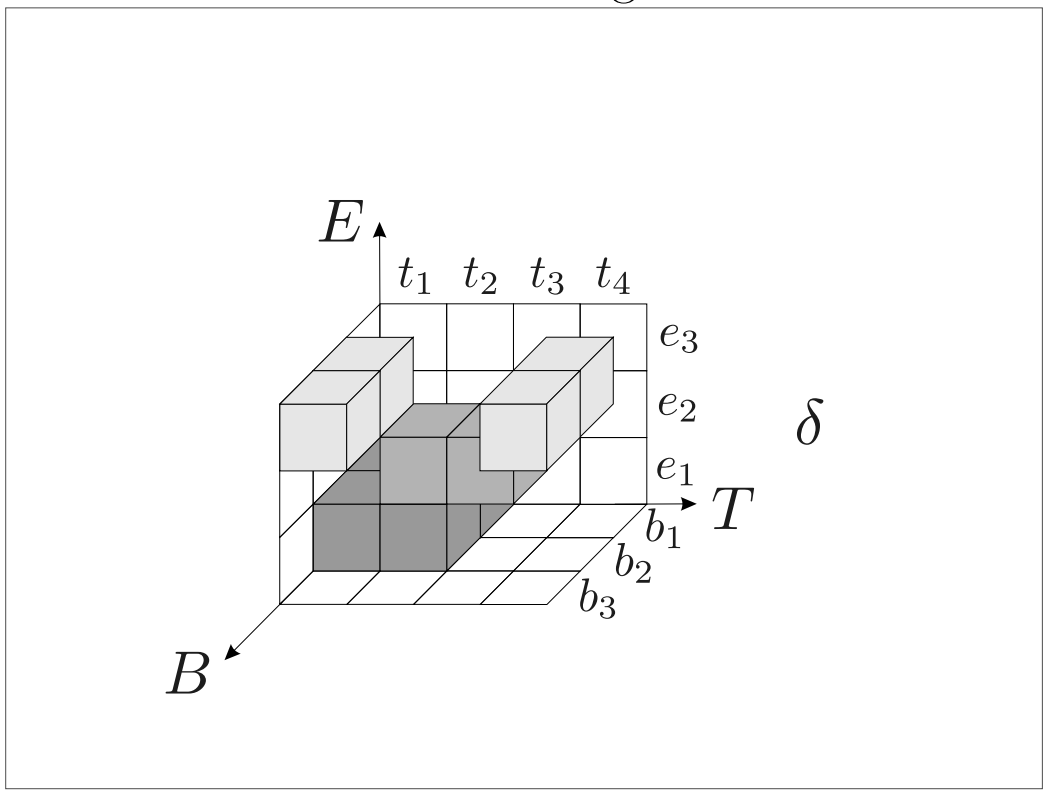
Therefore the set of all (theoretically) possible car configurations is:

$$\Omega = \mathrm{dom}(E) \times \mathrm{dom}(B) \times \mathrm{dom}(T)$$

Since not all combinations are technically possible (or wanted by marketing) a set of rules is used to cancel out invalid combinations.

Example: Car Manufacturing

Possible car configurations

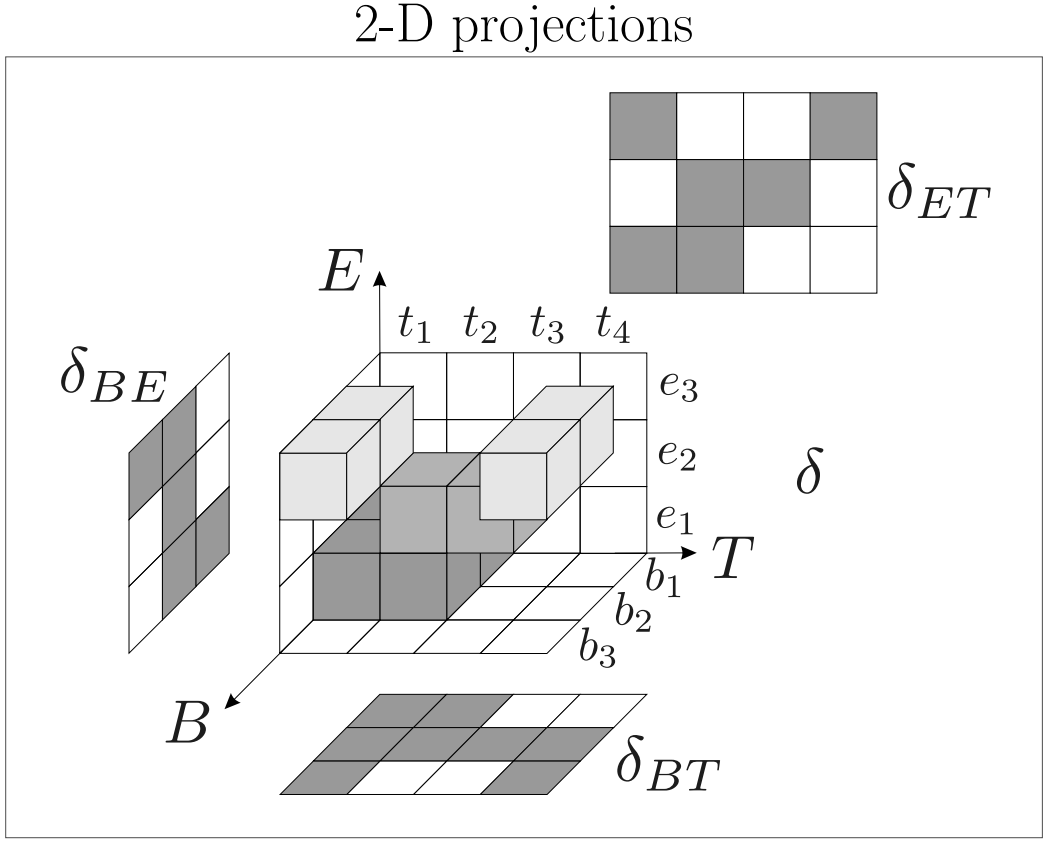


Every cube designates a valid value combination.

10 car configurations in our model.

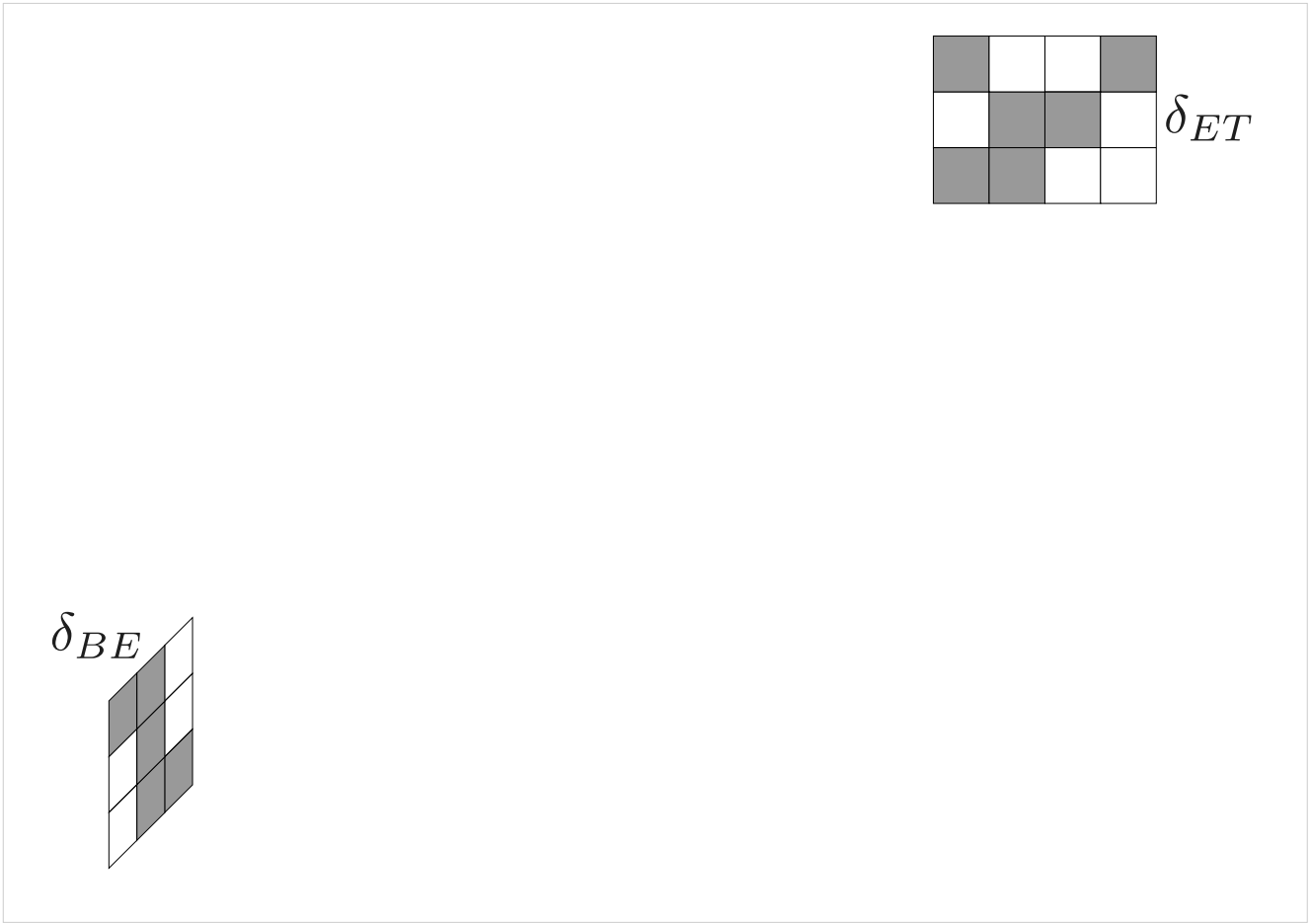
Different colors are intended to distinguish the cubes only.

Example

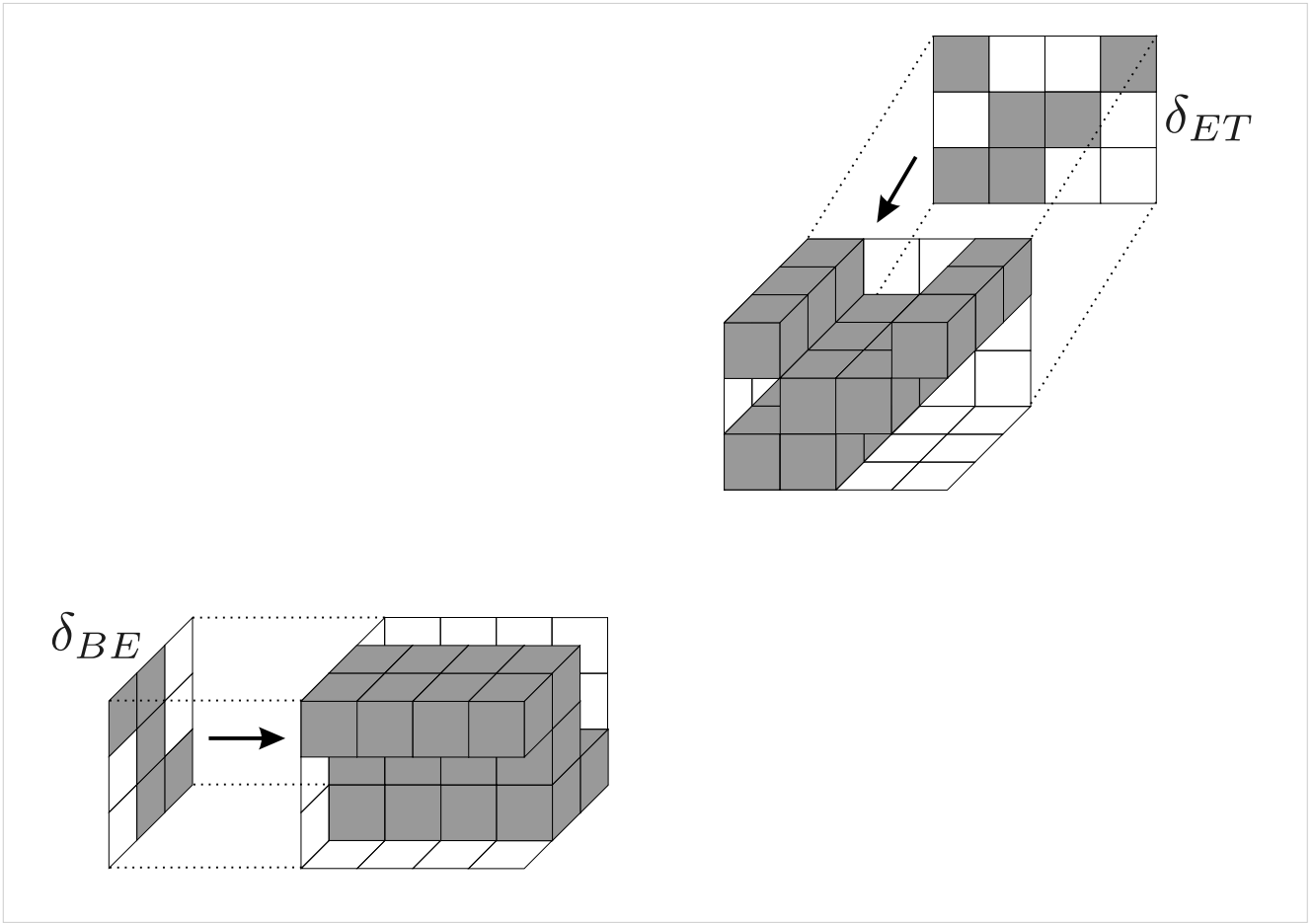


Is it possible to reconstruct δ from the δ_i?

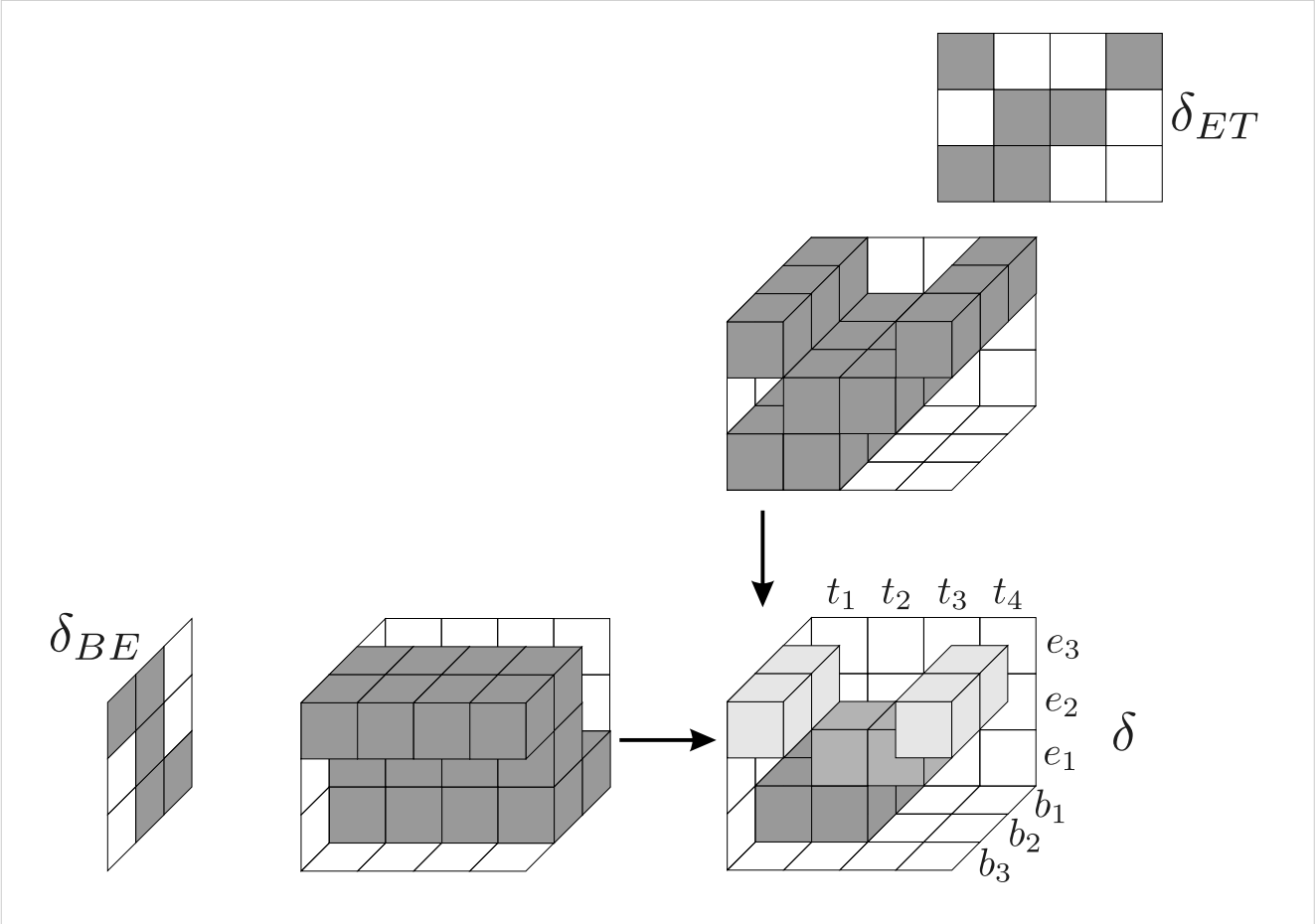
Example: Reconstruction of δ with δ_{BE} and δ_{ET}



Example: Reconstruction of δ with δ_{BE} and δ_{ET}



Example: Reconstruction of δ with δ_{BE} and δ_{ET}



Objective

Is it possible to exploit local constraints (wherever they may come from — both structural and expert knowledge-based) in a way that allows for a decomposition of the large (intractable) distribution $P(X_1, \ldots, X_n)$ into several sub-structures $\{C_1, \ldots, C_m\}$ such that:

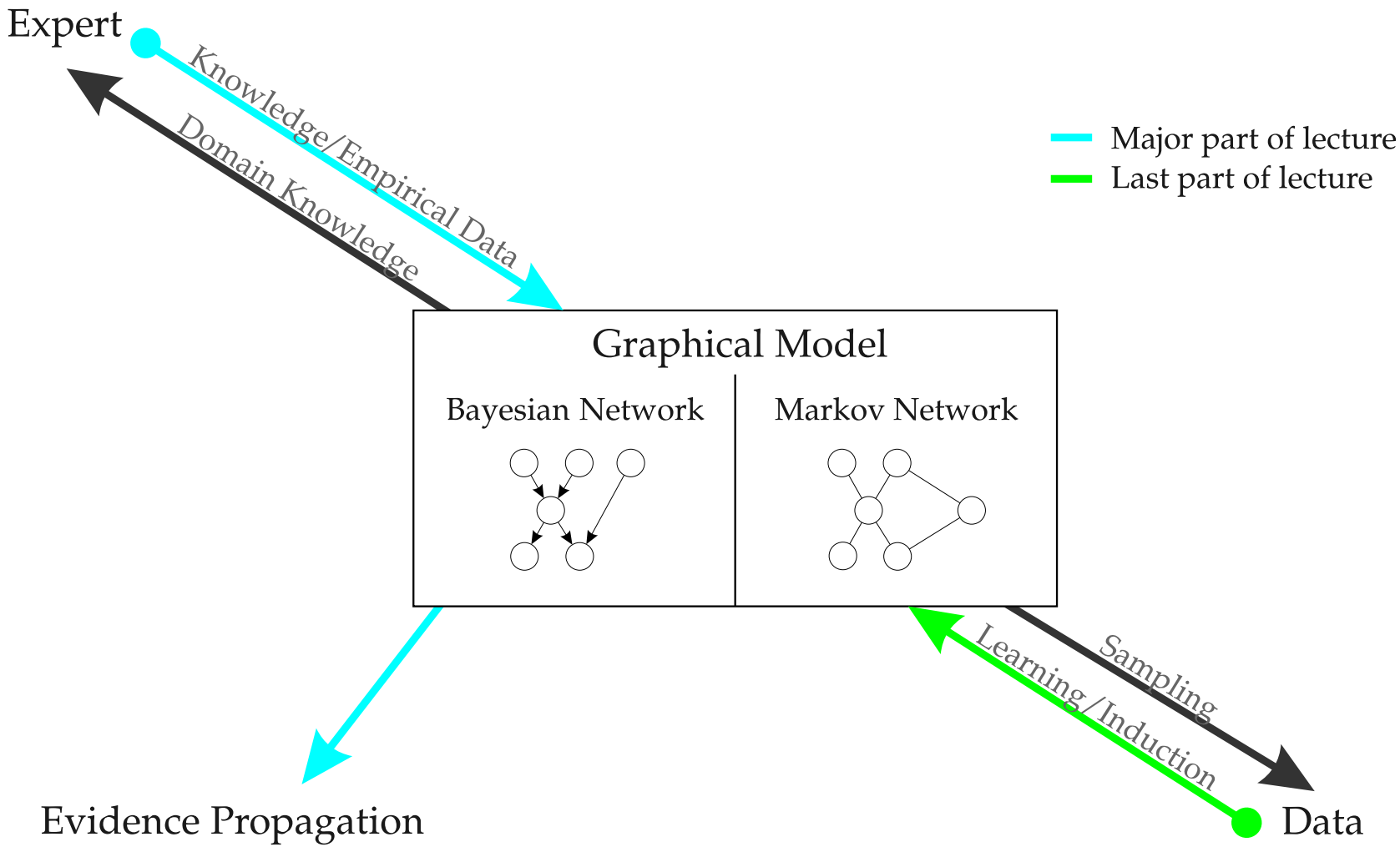
The collective size of those sub-structures is much smaller than that of the original distribution P.

The original distribution P is recomposable (with no or at least as few as possible errors) from these sub-structures in the following way:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{m} \Psi_i(c_i)$$

where c_i is an instantiation of C_i and $\Psi_i(c_i) \in \mathbb{R}^+$ a *factor potential*.

The Big Picture / Lecture Roadmap



Bayes Networks

Bayes Network

A *Bayes Network* (V, E, P) consists of a set $V = \{X_1, \ldots, X_n\}$ of random variables and a set E of directed edges between the variables.

Each variable has a finite set of mutual exclusive and collectively exhaustive states.

The variables in combination with the edges form a directed, acyclic graph.

Each variable with parent nodes $B_1, \ldots, B_m$ is assigned a table $P(A \mid B_1, \ldots, B_m)$.

Note, that the connections between the nodes not necessarily express a causal relationship.

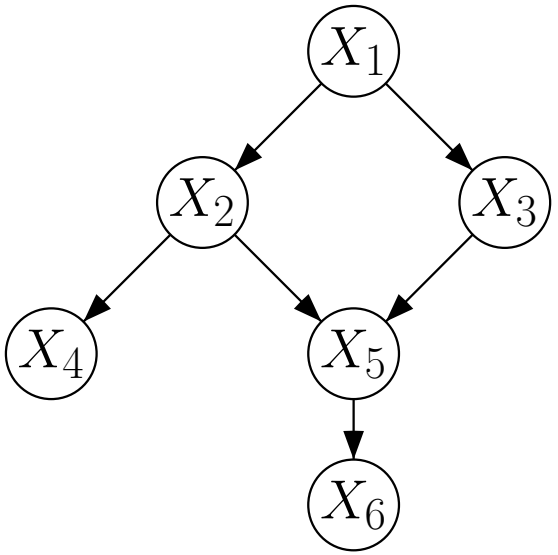
For every belief network, the following equation holds:

$$P(V) \quad = \prod_{v \in V : P(c(v)) > 0} P(v \mid c(v))$$

with $c(v)$ being the parent nodes of v.

Probabilistic Dependency Networks

Probabilistic dependency networks are directed acyclic graphs (DAGs) where the nodes represent propositions or variables and the directed edges model a direct dependence between the connected nodes. The strength of dependence is defined by conditional probabilities.

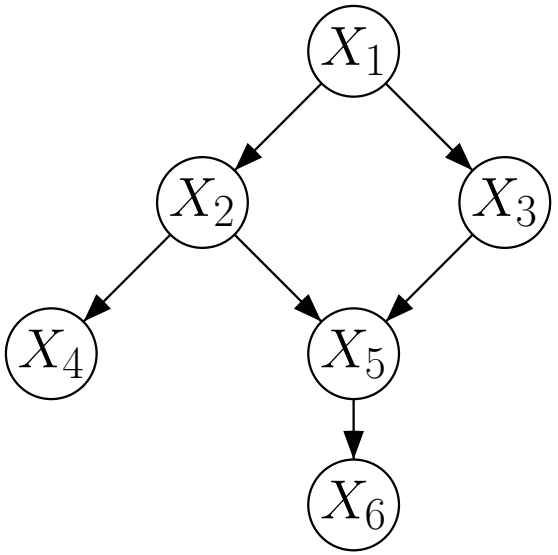


In general (according chain rule):

$$\begin{aligned} P(X_1, \ldots, X_6) \;= \; & P(X_6 \mid X_5, \ldots, X_1) \cdot \\ & P(X_5 \mid X_4, \ldots, X_1) \cdot \\ & P(X_4 \mid X_3, X_2, X_1) \cdot \\ & P(X_3 \mid X_2, X_1) \cdot \\ & P(X_2 \mid X_1) \cdot \\ & P(X_1) \end{aligned}$$

Probabilistic Dependency Networks

Probabilistic dependency networks are directed acyclic graphs (DAGs) where the nodes represent propositions or variables and the directed edges model a direct causal dependence between the connected nodes. The strength of dependence is defined by conditional probabilities.



According graph (independence structure):

$$\begin{aligned} P(X_1, \ldots, X_6) \;= \; & P(X_6 \mid X_5) \cdot \\ & P(X_5 \mid X_2, X_3) \cdot \\ & P(X_4 \mid X_2) \cdot \\ & P(X_3 \mid X_1) \cdot \\ & P(X_2 \mid X_1) \cdot \\ & P(X_1) \end{aligned}$$

Formal Framework

Nomenclature for the next slides:

$X_1, \ldots, X_n$	Variables (properties, attributes, random variables, propositions)
$\Omega_1, \ldots, \Omega_n$	respective finite domains (also designated with $\mathrm{dom}(X_i)$)
$\Omega = \bigtimes_{i=1}^{n} \Omega_i$	Universe of Discourse (tuples that characterize objects described by $X_1, \ldots, X_n$)
$\Omega_i = \{x_i^{(1)}, \ldots, x_i^{(n_i)}\}$	$n = 1, \ldots, n,\ n_i \in \mathbb{N}$

Formal Framework

The product space $(\Omega, 2^{\Omega}, P)$ is unique iff $P(\{(x_1, \ldots, x_n)\})$ is specified for all $x_i \in \{x_i^{(1)}, \ldots, x_i^{(n_i)}\}$, $i = 1, \ldots, n$.

When the distribution $P(X_1, \ldots, X_n)$ is given in tabular form, then $\prod_{i=1}^{n} |\Omega_i|$ entries are necessary.

For variables with $|\Omega_i| \geq 2$ at least 2^n entries.

The application of DAGs allows for the representation of existing (in)dependencies.

Constructing a DAG

input $P(X_1, \ldots, X_n)$
output a DAG G

1: Set the nodes of G to $\{X_1, \ldots, X_n\}$.

2: Choose a total ordering on the set of variables
(e. g. $X_1 \prec X_2 \prec \cdots \prec X_n$)

3: For X_i find the smallest (uniquely determinable) set $S_i \subseteq \{X_1, \ldots, X_n\}$ sucht that $P(X_i \mid S_i) = P(X_i \mid X_1 \ldots, X_{i-1})$.

4: Connect all nodes in S_i with X_i and store $P(X_i \mid S_i)$ as quantization of the dependencies for that node X_i (given its parents).

5: **return** G

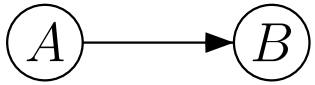
Example

Let a_1, a_2, a_3 be three blood groups and b_1, b_2, b_3 three indications of a blood group test.

Variables:	A (blood group)	B (indication)
Domains:	$\Omega_A = \{a_1, a_2, a_3\}$	$\Omega_B = \{b_1, b_2, b_3\}$

It is conjectured that there is a causal relationship between the variables.

$P(\{(a_i, b_j)\})$	b_1	b_2	b_3	$\sum$
a_1	0.64	0.08	0.08	0.8
a_2	0.01	0.08	0.01	0.1
a_3	0.01	0.01	0.08	0.1
$\sum$	0.66	0.17	0.17	1



$P(A, B) = P(B \mid A) \cdot P(A)$

We are dealing with a belief network.

Expert Knowledge

Metastatic cancer is a possible cause of brain cancer, and an explanation for elevated levels of calcium in the blood. Both phenomena together can explain that a patient falls into a coma. Severe headaches are possibly associated with a brain tumor.

Special Case

The patient has severe headaches.

Question

Will the patient is go into a coma?

Choice of universe of discourse

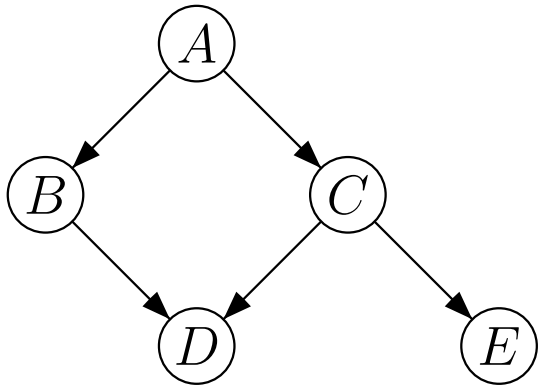
	Variable	Domain
A	metastatic cancer	$\{a_1, a_2\}$
B	increased serum calcium	$\{b_1, b_2\}$
C	brain tumor	$\{c_1, c_2\}$
D	coma	$\{d_1, d_2\}$
E	headache	$\{e_1, e_2\}$

$(\cdot_1$ — present,$\cdot_2$ — absent)

$$\Omega = \{a_1, a_2\} \times \cdots \times \{e_1, e_2\}$$

$$|\Omega| = 32$$

Analysis of dependencies



Example

Choice of probability parameters

$$P(a, b, c, d, e) \overset{\text{abbr.}}{=} P(A = a, B = b, C = c, D = d, E = e)$$
$$= P(e \mid c)P(d \mid b, c)P(c \mid a)P(b \mid a)P(a)$$

Shorthand notation

11 values to store instead of 31

Consult experts, textbooks, case studies, surveys, etc.

Calculation of conditional probabilities

Calculation of marginal probabilities

Crux of the Matter

Knowledge acquisition (Where do the numbers come from?)
$\rightarrow$ learning strategies

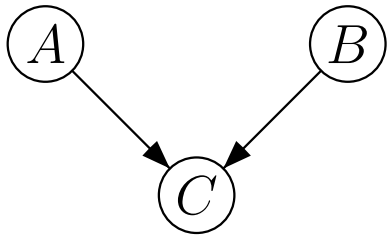
Computational complexities
$\rightarrow$ exploit independencies

Problem:

When does the independency of X and Y given Z hold in (V, E, P)?

How to determine a decomposition based of the graph structure?

Example



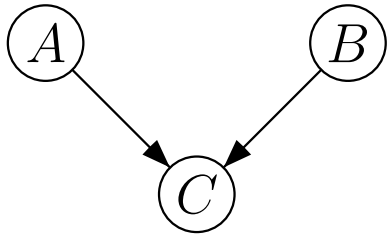
Meal quality	
A	quality of ingredients
B	cook's skill
C	meal quality

If C is not known, A and B are independent.

If C is known, then A and B become (conditionally) dependent given C.

$A \not\perp\!\!\!\perp B \mid C$

Converging Connection: Marginal Independence



Decomposition according to graph:

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

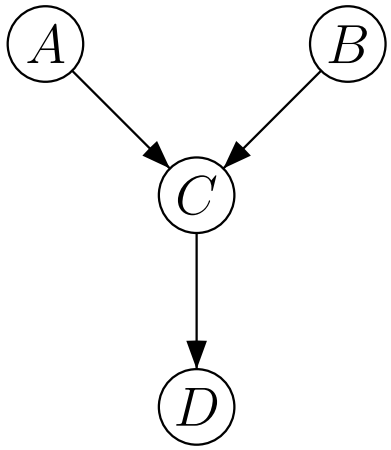
Embedded Independence:

$$P(A, B, C) = \frac{P(A, B, C)}{P(A, B)} \cdot P(A) \cdot P(B) \qquad \text{with} \quad P(A, B) \neq 0$$

$$P(A, B) = P(A) \cdot P(B)$$

$$\Rightarrow A \perp\!\!\!\perp B \mid \emptyset$$

Example (cont.)



Meal quality	
A	quality of ingredients
B	cook's skill
C	meal quality
D	restaurant success

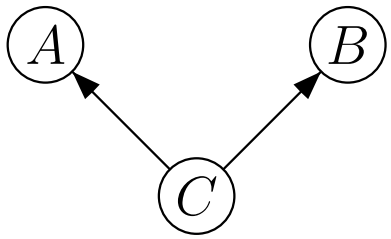
If nothing is known about the restaurant success or meal quality or both, the cook's skills and quality of the ingredients are unrelated, that is, *independent*.

However, if we observe that the restaurant has no success, we can infer that the meal quality might be bad.

If we further learn that the ingredients quality is high, we will conclude that the cook's skills must be low, thus rendering both variables *dependent*.

$A \not\perp\!\!\!\perp B \mid D$

Diverging Connection



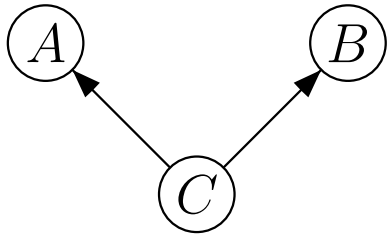
Diagnosis	
A	body temperature
B	cough
C	disease

If C is unknown, knowledge about A ist relevant for B and vice versa, i. e. A and B are marginally dependent.

However, if C is observed, A and B become conditionally independent given C.

A influences B via C. If C is known it in a way blocks the information from flowing from A to B, thus rendering A and B (conditionally) independent.

Diverging Connection: Conditional Independence



Decomposition according to graph:

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

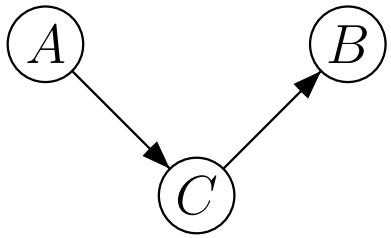
Embedded Independence:

$$\begin{aligned} P(A, B \mid C) &= P(A \mid C) \cdot P(B \mid C) \\ &\Rightarrow A \perp\!\!\!\perp B \mid C \end{aligned}$$

Alternative derivation:

$$\begin{aligned} P(A, B, C) &= P(A \mid C) \cdot P(B, C) \\ P(A \mid B, C) &= P(A \mid C) \\ &\Rightarrow A \perp\!\!\!\perp B \mid C \end{aligned}$$

Serial Connection

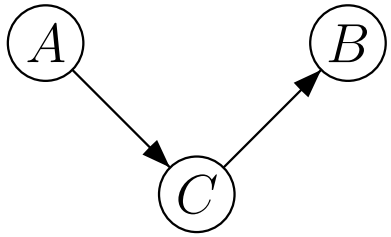


Accidents	
A	rain
B	accident risk
C	road conditions

Analog scenario to case 2

A influences C and C influences B. Thus, A influences B.
If C is known, it blocks the path between A and B.

Serial Connection: Conditional Independence



Decomposition according to graph:

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

Embedded Independence:

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

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

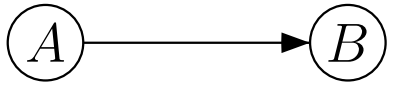
$$\Rightarrow A \perp\!\!\!\perp B \mid C$$

Trivial Cases:

Marginal Independence:

$$P(A, B) = P(A) \cdot P(B)$$

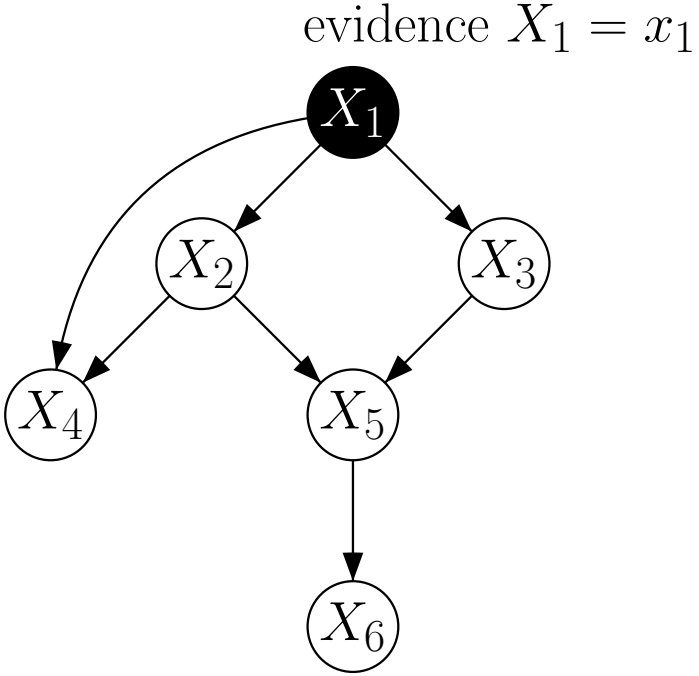
Marginal Dependence:



$$P(A, B) = P(B \mid A) \cdot P(A)$$

Question

Question: Are X_2 and X_3 independent given X_1?



Repetition: d-Separation

Let $G = (V, E)$ a DAG and $X, Y, Z \in V$ three nodes.

a) A set $S \subseteq V \backslash \{X, Y\}$ *d-separates* X and Y, if S blocks all paths between X and Y. (paths may also route in opposite edge direction)

b) A path π is d-separated by S if at least one pair of consecutive edges along π is blocked. There are the following blocking conditions:

1.	$X \leftarrow Y \rightarrow Z$	tail-to-tail
2.	$X \leftarrow Y \leftarrow Z$ $X \rightarrow Y \rightarrow Z$	head-to-tail
3.	$X \rightarrow Y \leftarrow Z$	head-to-head

c) Two edges that meet tail-to-tail or head-to-tail in node Y are blocked if $Y \in S$.

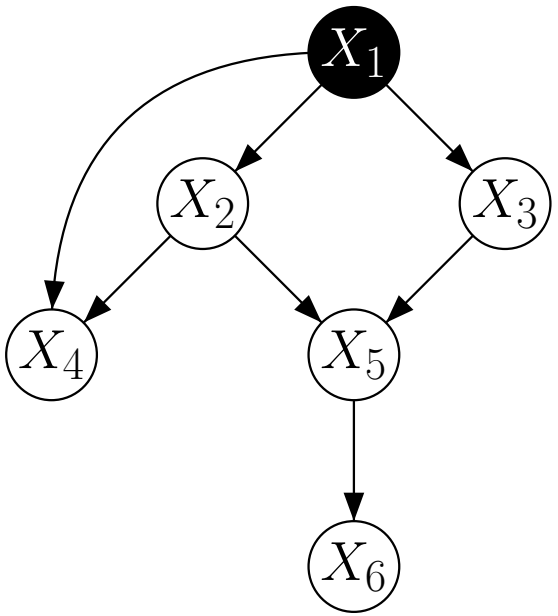
d) Two edges meeting head-to-head in Y are blocked if neither Y nor its successors are in S.

Relation to Conditional independence

If $S \subseteq V \backslash \{X, Y\}$ d-separates X and Y in a Belief network (V, E, P) then X and Y are conditionally independent given S:

$$P(X, Y \mid S) = P(X \mid S) \cdot P(Y \mid S)$$

Application to the previous example:



Paths: $\pi_1 = \langle X_2 - X_1 - X_3 \rangle$, $\pi_2 = \langle X_2 - X_5 - X_3 \rangle$
$\pi_3 = \langle X_2 - X_4 - X_1 - X_3 \rangle$, $S = \{X_1\}$

π_1 $X_2 \leftarrow X_1 \rightarrow X_3$ tail-to-tail
$X_1 \in S \Rightarrow \pi_1$ is blocked by S

π_2 $X_2 \rightarrow X_5 \leftarrow X_3$ head-to-head
$X_5, X_6 \notin S \Rightarrow \pi_2$ is blocked by S

π_3 $X_4 \leftarrow X_1 \rightarrow X_3$ tail-to-tail
$X_2 \rightarrow X_4 \leftarrow X_1$ head-to-head
both connections are blocked $\Rightarrow \pi_3$ is blocked

Example (cont.)

Answer: X_2 and X_3 are d-separated via $\{X_1\}$. Therefore X_2 and X_3 become conditionally independent given X_1.

$S = \{X_1, X_4\} \quad \Rightarrow \quad X_2$ and X_3 are d-separated by S

$S = \{X_1, X_6\} \quad \Rightarrow \quad X_2$ and X_3 are *not* d-separated by S

Algebraic structure of CI statements

Question: Is it possible to use a formal scheme to infer new conditional independence (CI) statements from a set of initial CIs?

Repetition

Let $(\Omega, \mathcal{E}, P)$ be a probability space and W, X, Y, Z disjoint subsets of variables. If X and Y are conditionally independent given Z we write:

$$X \perp\!\!\!\perp_P Y \mid Z$$

Often, the following (equivalent) notation is used:

$$I_P(X \mid Z \mid Y) \quad \text{or} \quad I_P(X, Y \mid Z)$$

If the underlying space is known the index P is omitted.

(Semi-)Graphoid Axioms

Definition: Let V be a set of (mathematical) objects and $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ a three-place relation of subsets of V. Furthermore, let W, X, Y, and Z be four disjoint subsets of V. The four statements

symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Rightarrow (Y \perp\!\!\!\perp X \mid Z)$

decomposition: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Rightarrow (W \perp\!\!\!\perp Y \mid Z) \wedge (X \perp\!\!\!\perp Y \mid Z)$

weak union: $(W \cup X \perp\!\!\!\perp Y \mid Z) \Rightarrow (X \perp\!\!\!\perp Y \mid Z \cup W)$

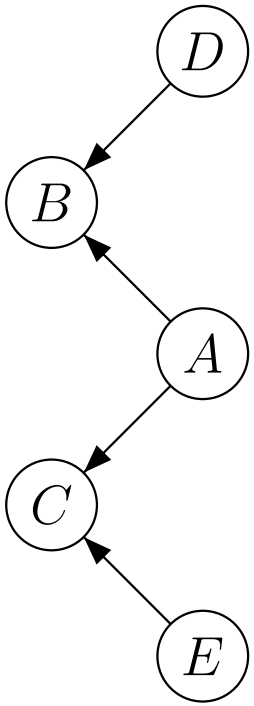
contraction: $(X \perp\!\!\!\perp Y \mid Z \cup W) \wedge (W \perp\!\!\!\perp Y \mid Z) \Rightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

are called the **semi-graphoid axioms**. A three-place relation $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ that satisfies the semi-graphoid axioms for all W, X, Y, and Z is called a **semi-graphoid**.
The above four statements together with

intersection: $(W \perp\!\!\!\perp Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp Y \mid Z \cup W) \Rightarrow (W \cup X \perp\!\!\!\perp Y \mid Z)$

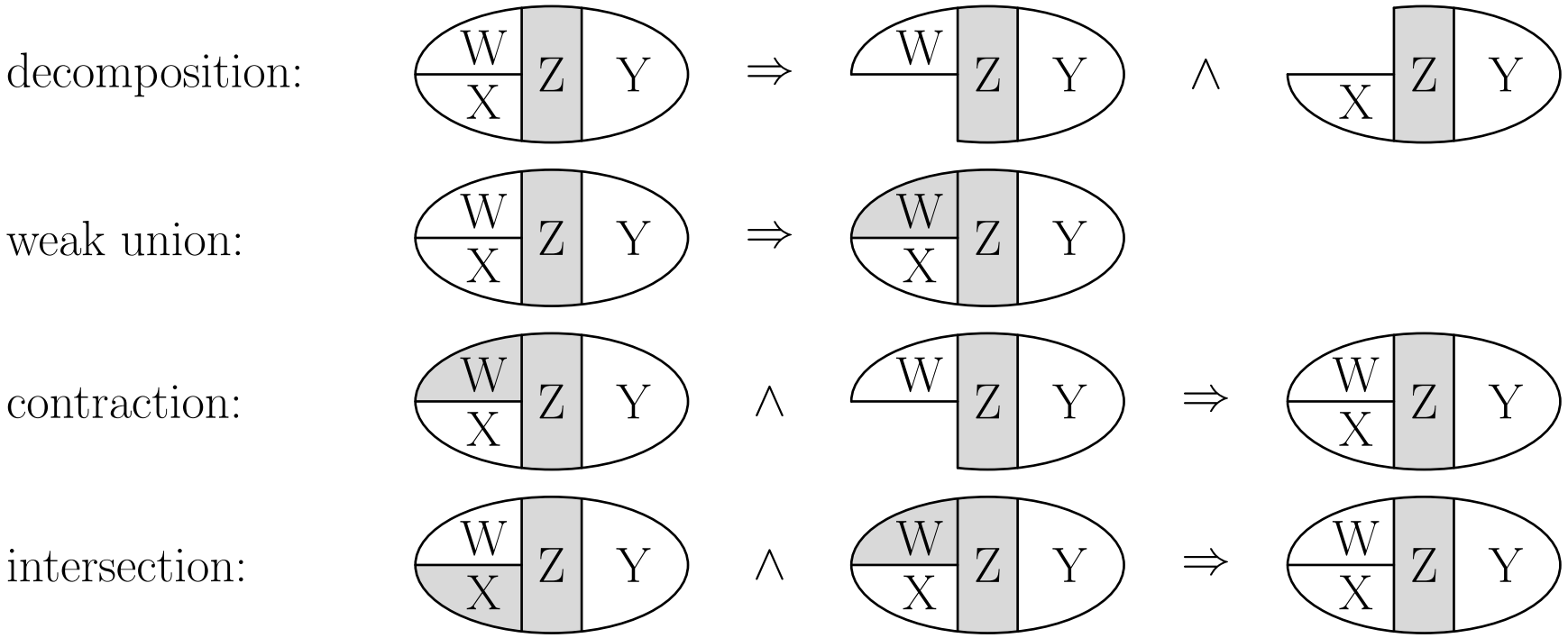
are called the **graphoid axioms**. A three-place relation $(\cdot \perp\!\!\!\perp \cdot \mid \cdot)$ that satisfies the graphoid axioms for all W, X, Y, and Z is called a **graphoid**.

Example



$$
\begin{aligned}
& D \perp\!\!\!\perp A, C \mid \emptyset \quad \wedge \quad B \perp\!\!\!\perp C \mid A, D \\
\overset{\text{w. union}}{\Longrightarrow} \quad & D \perp\!\!\!\perp C \mid A \quad \wedge \quad B \perp\!\!\!\perp C \mid A, D \\
\overset{\text{symm.}}{\Longleftrightarrow} \quad & C \perp\!\!\!\perp D \mid A \quad \wedge \quad C \perp\!\!\!\perp B \mid A, D \\
\overset{\text{contr.}}{\Longrightarrow} \quad & C \perp\!\!\!\perp B, D \mid A \\
\overset{\text{decomp.}}{\Longrightarrow} \quad & C \perp\!\!\!\perp B \mid A \\
\overset{\text{symm.}}{\Longleftrightarrow} \quad & B \perp\!\!\!\perp C \mid A
\end{aligned}
$$

Illustration of the (Semi-)Graphoid Axioms



Similar to the properties of **separation in graphs**.
Idea: **Represent conditional independence by separation in graphs.**

Separation in Graphs

Definition: Let $G = (V, E)$ be an undirected graph and X, Y, and Z three disjoint subsets of nodes. Z **u-separates** X and Y in G, written $\langle X \mid Z \mid Y \rangle_G$, iff all paths from a node in X to a node in Y contain a node in Z. A path that contains a node in Z is called **blocked** (by Z), otherwise it is called **active**.

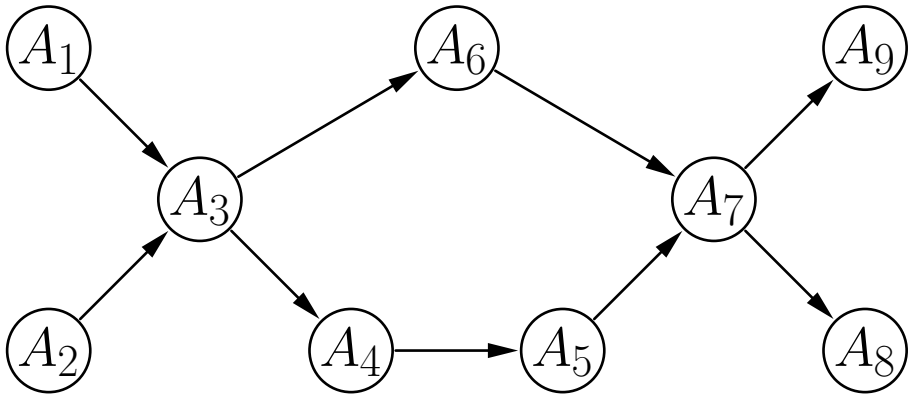
Definition: Let $\vec{G} = (V, \vec{E})$ be a directed acyclic graph and X, Y, and Z three disjoint subsets of nodes. Z **d-separates** X and Y in $\vec{G}$, written $\langle X \mid Z \mid Y \rangle_{\vec{G}}$, iff there is *no* path from a node in X to a node in Y along which the following two conditions hold:

1. every node with converging edges either is in Z or has a descendant in Z,
2. every other node is not in Z.

A path satisfying the two conditions above is said to be **active**,
otherwise it is said to be **blocked** (by Z).

Separation in Directed Acyclic Graphs

Example Graph:



Valid Separations:

$\langle \{A_1\} \mid \{A_3\} \mid \{A_4\} \rangle$
$\langle \{A_3\} \mid \{A_4, A_6\} \mid \{A_7\} \rangle$
$\langle \{A_8\} \mid \{A_7\} \mid \{A_9\} \rangle$
$\langle \{A_1\} \mid \emptyset \mid \{A_2\} \rangle$

Invalid Separations:

$\langle \{A_1\} \mid \{A_4\} \mid \{A_2\} \rangle$
$\langle \{A_4\} \mid \{A_3, A_7\} \mid \{A_6\} \rangle$
$\langle \{A_1\} \mid \{A_6\} \mid \{A_7\} \rangle$
$\langle \{A_1\} \mid \{A_4, A_9\} \mid \{A_5\} \rangle$

Conditional (In)Dependence Graphs

Definition: Let $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ be a three-place relation representing the set of conditional independence statements that hold in a given distribution δ over a set U of attributes. An undirected graph $G = (U, E)$ over U is called a **conditional dependence graph** or a **dependence map** w.r.t. δ, iff for all disjoint subsets $X, Y, Z \subseteq U$ of attributes

$$X \perp\!\!\!\perp_\delta Y \mid Z \;\Rightarrow\; \langle X \mid Z \mid Y \rangle_G,$$

i.e., if G captures by u-separation all (conditional) independences that hold in δ and thus represents only valid (conditional) dependences. Similarly, G is called a **conditional independence graph** or an **independence map** w.r.t. δ, iff for all disjoint subsets $X, Y, Z \subseteq U$ of attributes

$$\langle X \mid Z \mid Y \rangle_G \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y \mid Z,$$

i.e., if G captures by u-separation only (conditional) independences that are valid in δ. G is said to be a **perfect map** of the conditional (in)dependences in δ, if it is both a dependence map and an independence map.

Conditional (In)Dependence Graphs

Definition: A conditional dependence graph is called **maximal** w.r.t. a distribution δ (or, in other words, a **maximal dependence map** w.r.t. δ) iff no edge can be added to it so that the resulting graph is still a conditional dependence graph w.r.t. the distribution δ.

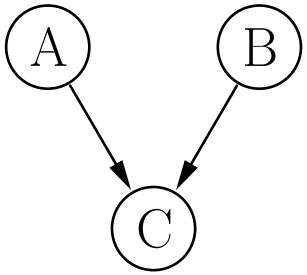
Definition: A conditional independence graph is called **minimal** w.r.t. a distribution δ (or, in other words, a **minimal independence map** w.r.t. δ) iff no edge can be removed from it so that the resulting graph is still a conditional independence graph w.r.t. the distribution δ.

Conditional independence graphs are sometimes required to be minimal.

However, this requirement is not necessary for a conditional independence graph to be usable for evidence propagation.

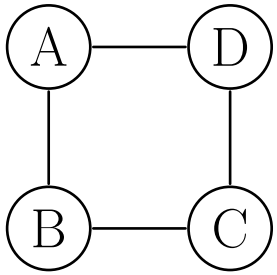
The disadvantage of a non-minimal conditional independence graph is that evidence propagation may be more costly computationally than necessary.

Perfect directed map, no perfect undirected map:



p_{ABC}	$A = a_1$		$A = a_2$	
	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	$4/24$	$3/24$	$3/24$	$2/24$
$C = c_2$	$2/24$	$3/24$	$3/24$	$4/24$

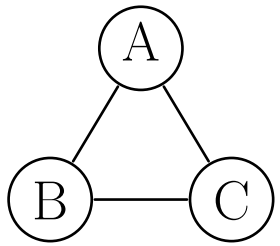
Perfect undirected map, no perfect directed map:



p_{ABCD}		$A = a_1$		$A = a_2$	
		$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	$D = d_1$	$1/47$	$1/47$	$1/47$	$2/47$
	$D = d_2$	$1/47$	$1/47$	$2/47$	$4/47$
$C = c_2$	$D = d_1$	$1/47$	$2/47$	$1/47$	$4/47$
	$D = d_2$	$2/47$	$4/47$	$4/47$	$16/47$

Limitations of Graph Representations

There are also probability distributions for which
there exists neither a directed nor an undirected perfect map:



p_{ABC}	$A = a_1$		$A = a_2$	
	$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	$2/12$	$1/12$	$1/12$	$2/12$
$C = c_2$	$1/12$	$2/12$	$2/12$	$1/12$

In such cases either not all dependences or not all independences

can be captured by a graph representation.
In such a situation one usually decides to neglect some of the independence information, that is, to use only a (minimal) conditional independence graph.

This is sufficient for correct evidence propagation,
the existence of a perfect map is not required.

Markov Properties of Undirected Graphs

Definition: An undirected graph $G = (U, E)$ over a set U of attributes is said to have (w.r.t. a distribution δ) the

pairwise Markov property,

iff in δ any pair of attributes which are nonadjacent in the graph are conditionally independent given all remaining attributes, i.e., iff

$$\forall A, B \in U, A \neq B: \quad (A, B) \notin E \;\Rightarrow\; A \perp\!\!\!\perp_\delta B \mid U - \{A, B\},$$

local Markov property,

iff in δ any attribute is conditionally independent of all remaining attributes given its neighbors, i.e., iff

$$\forall A \in U: \quad A \perp\!\!\!\perp_\delta U - \text{closure}(A) \mid \text{boundary}(A),$$

global Markov property,

iff in δ any two sets of attributes which are u-separated by a third are conditionally independent given the attributes in the third set, i.e., iff

$$\forall X, Y, Z \subseteq U: \quad \langle X \mid Z \mid Y \rangle_G \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y \mid Z.$$

Markov Properties of Directed Acyclic Graphs

Definition: A directed acyclic graph $\vec{G} = (U, \vec{E})$ over a set U of attributes is said to have (w.r.t. a distribution δ) the

pairwise Markov property,

iff in δ any attribute is conditionally independent of any non-descendant not among its parents given all remaining non-descendants, i.e., iff

$$\forall A, B \in U: \; B \in \text{non-descs}(A) - \text{parents}(A) \;\Rightarrow\; A \perp\!\!\!\perp_\delta B \mid \text{non-descs}(A) - \{B\},$$

local Markov property,

iff in δ any attribute is conditionally independent of all remaining non-descendants given its parents, i.e., iff

$$\forall A \in U: \quad A \perp\!\!\!\perp_\delta \text{non-descs}(A) - \text{parents}(A) \mid \text{parents}(A),$$

global Markov property,

iff in δ any two sets of attributes which are d-separated by a third are conditionally independent given the attributes in the third set, i.e., iff

$$\forall X, Y, Z \subseteq U: \quad \langle X \mid Z \mid Y \rangle_{\vec{G}} \;\Rightarrow\; X \perp\!\!\!\perp_\delta Y \mid Z.$$

Equivalence of Markov Properties

Theorem: If a three-place relation $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ representing the set of conditional independence statements that hold in a given joint distribution δ over a set U of attributes satisfies the graphoid axioms, then the pairwise, the local, and the global Markov property of an undirected graph $G = (U, E)$ over U are equivalent.

Theorem: If a three-place relation $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ representing the set of conditional independence statements that hold in a given joint distribution δ over a set U of attributes satisfies the semi-graphoid axioms, then the local and the global Markov property of a directed acyclic graph $\vec{G} = (U, \vec{E})$ over U are equivalent.

If $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ satisfies the graphoid axioms, then the pairwise, the local, and the global Markov property are equivalent.

Markov Equivalence of Graphs

Can two distinct graphs represent the exactly the same set of conditional independence statements?

The answer is relevant for learning graphical models from data, because it determines whether we can expect a unique graph as a learning result or not.

Definition: Two (directed or undirected) graphs $G_1 = (U, E_1)$ and $G_2 = (U, E_2)$ with the same set U of nodes are called **Markov equivalent** iff they satisfy the same set of node separation statements (with d-separation for directed graphs and u-separation for undirected graphs), or formally, iff

$$\forall X, Y, Z \subseteq U : \quad \langle X \mid Z \mid Y \rangle_{G_1} \;\Leftrightarrow\; \langle X \mid Z \mid Y \rangle_{G_2}.$$

No two different *undirected* graphs can be Markov equivalent.

The reason is that these two graphs, in order to be different, have to differ in at least one edge. However, the graph lacking this edge satisfies a node separation (and thus expresses a conditional independence) that is not statisfied (expressed) by the graph possessing the edge.

Markov Equivalence of Graphs

Definition: Let $\vec{G} = (U, \vec{E})$ be a directed graph.
The **skeleton** of $\vec{G}$ is the undirected graph $G = (V, E)$ where E contains the same edges as $\vec{E}$, but with their directions removed, or formally:

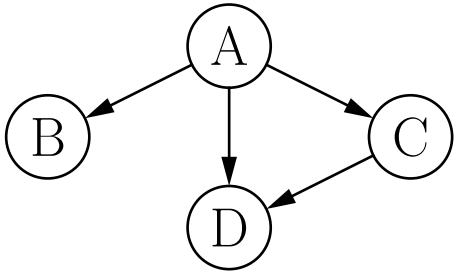
$$E = \{(A, B) \in U \times U \mid (A, B) \in \vec{E} \vee (B, A) \in \vec{E}\}.$$

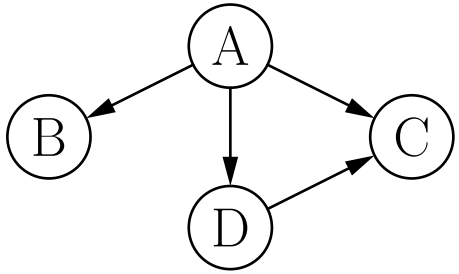
Definition: Let $\vec{G} = (U, \vec{E})$ be a directed graph and $A, B, C \in U$ three nodes of $\vec{G}$. The triple (A, B, C) is called a **v-structure** of $\vec{G}$ iff $(A, B) \in \vec{E}$ and $(C, B) \in \vec{E}$, but neither $(A, C) \in \vec{E}$ nor $(C, A) \in \vec{E}$, that is, iff $\vec{G}$ has converging edges from A and C at B, but A and C are unconnected.

Theorem: Let $\vec{G}_1 = (U, \vec{E}_1)$ and $\vec{G}_2 = (U, \vec{E}_2)$ be two directed acyclic graphs with the same node set U. The graphs $\vec{G}_1$ and $\vec{G}_2$ are Markov equivalent iff they possess the same skeleton and the same set of v-structures.

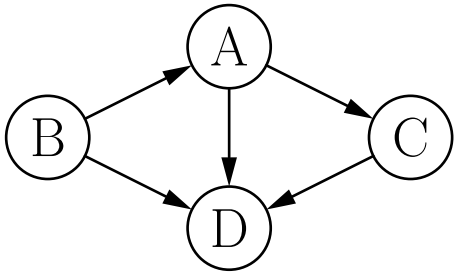
Intuitively:
Edge directions may be reversed if this does not change the set of v-structures.

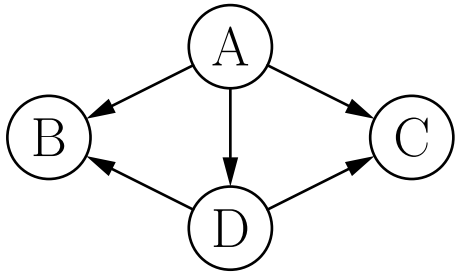
Markov Equivalence of Graphs





Graphs with the same skeleton, but converging edges at different nodes, which start from *connected* nodes, can be Markov equivalent.





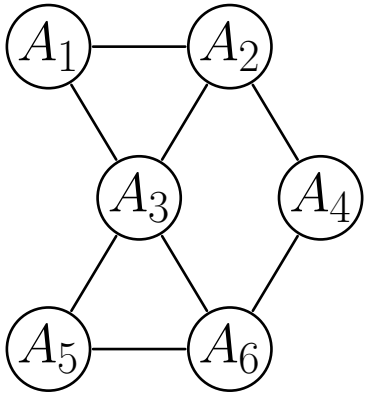
Of several edges that converge at a node only a subset may actually represent a v-structure. This v-structure, however, is relevant.

Definition: A probability distribution p_V over a set V of variables is called **decomposable** or **factorizable w.r.t. an undirected graph** $G = (V, E)$ iff it can be written as a product of nonnegative functions on the maximal cliques of G.

That is, let $\mathcal{M}$ be a family of subsets of variables, such that the subgraphs of G induced by the sets $M \in \mathcal{M}$ are the maximal cliques of G. Then there exist functions $\phi_M : \mathcal{E}_M \to \mathbb{R}_0^+$, $M \in \mathcal{M}$, $\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n)$:

$$p_V\left(\bigwedge_{A_i \in V} A_i = a_i\right) = \prod_{M \in \mathcal{M}} \phi_M\left(\bigwedge_{A_i \in M} A_i = a_i\right).$$

Example:

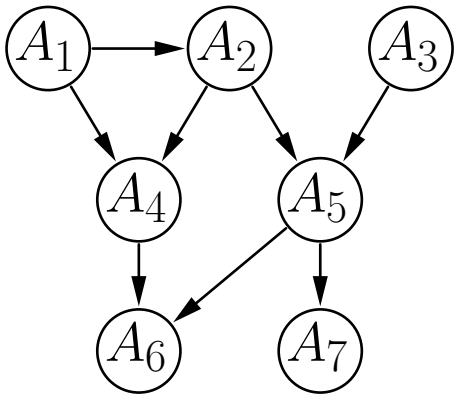


$$\begin{aligned} & p_V(A_1 = a_1, \ldots, A_6 = a_6) \\ = \;& \phi_{A_1A_2A_3}(A_1 = a_1, A_2 = a_2, A_3 = a_3) \\ \cdot \;& \phi_{A_3A_5A_6}(A_3 = a_3, A_5 = a_5, A_6 = a_6) \\ \cdot \;& \phi_{A_2A_4}(A_2 = a_2, A_4 = a_4) \\ \cdot \;& \phi_{A_4A_6}(A_4 = a_4, A_6 = a_6). \end{aligned}$$

Definition: A probability distribution p_U over a set U of attributes is called **decomposable** or **factorizable w.r.t. a directed acyclic graph** $\vec{G} = (U, \vec{E})$ over U, iff it can be written as a product of the conditional probabilities of the attributes given their parents in $\vec{G}$, i.e., iff

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$p_U\left(\bigwedge_{A_i \in U} A_i = a_i\right) = \prod_{A_i \in U} P\left(A_i = a_i \,\middle|\, \bigwedge_{A_j \in \mathrm{parents}_{\vec{G}}(A_i)} A_j = a_j\right).$$

Example:



$$\begin{aligned} & P(A_1 = a_1, \ldots, A_7 = a_7) \\ = \;& P(A_1 = a_1) \cdot P(A_2 = a_2 \mid A_1 = a_1) \cdot P(A_3 = a_3) \\ \cdot \;& P(A_4 = a_4 \mid A_1 = a_1, A_2 = a_2) \\ \cdot \;& P(A_5 = a_5 \mid A_2 = a_2, A_3 = a_3) \\ \cdot \;& P(A_6 = a_6 \mid A_4 = a_4, A_5 = a_5) \\ \cdot \;& P(A_7 = a_7 \mid A_5 = a_5). \end{aligned}$$

Core Theorem of Graphical Models:
Let p_V be a strictly positive probability distribution on a set V of (discrete) variables. A directed or undirected graph $G = (V, E)$ is a conditional independence graph w.r.t. p_V if and only if p_V is factorizable w.r.t. G.

Definition: A **Markov network** is an undirected conditional independence graph of a probability distribution p_V together with the family of positive functions ϕ_M of the factorization induced by the graph.

Definition: A **Bayesian network** is a directed conditional independence graph of a probability distribution p_U together with the family of conditional probabilities of the factorization induced by the graph.

Sometimes the conditional independence graph is required to be minimal, if it is to be used as the graph underlying a Markov or Bayesian network.
For correct evidence propagation it is not required that the graph is minimal. Evidence propagation may just be less efficient than possible.