Probabilistic Graphical Models

Christian Borgelt

Dept. of Knowledge Processing and Language Engineering
Otto-von-Guericke-University of Magdeburg
Universitätsplatz 2, D-39106 Magdeburg, Germany

E-mail: `borgelt@iws.cs.uni-magdeburg.de`
WWW: `http://fuzzy.cs.uni-magdeburg.de/~borgelt`

Contents

- Inference Networks / Graphical Models
- Relational Networks: Decomposition and Reasoning
- Probabilistic Networks: Decomposition and Reasoning
- Conditional Independence Graphs
- Graphs and Decompositions
- Evidence Propagation in Graphs
- Danish Jersey Cattle Blood Type Determination: An Example
- Learning Relational Networks from Data
- Learning Probabilistic Networks from Data
- Probabilistic Networks and Causality
- Fault Analysis at DaimlerChrysler: An Application

Inference Networks / Graphical Models

- **Decomposition:** Under certain conditions a distribution δ (e.g. a probability distribution) on a multi-dimensional domain, which encodes *prior* or *generic knowledge* about this domain, can be decomposed into a set $\{\delta_1, \ldots, \delta_s\}$ of (overlapping) distributions on lower-dimensional subspaces.

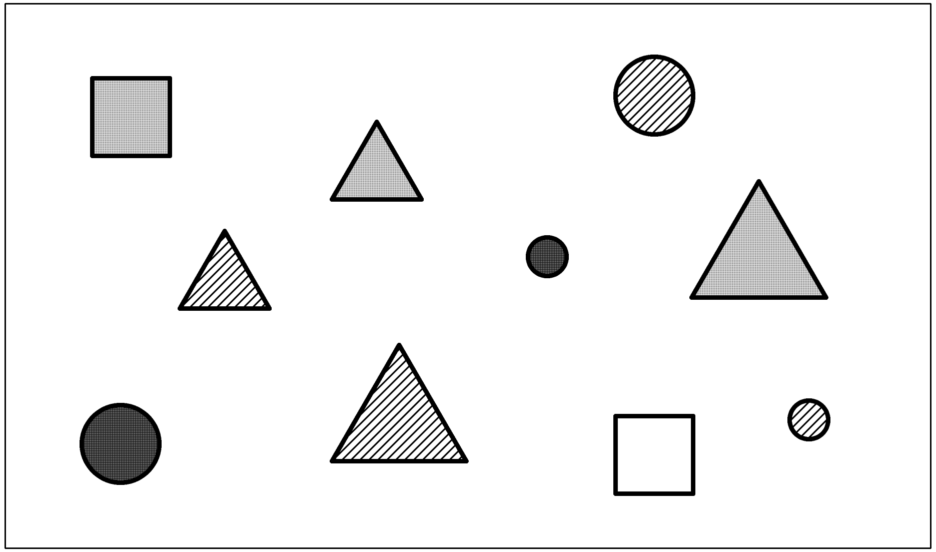
- **Simplified Reasoning:** If such a decomposition is possible, it is sufficient to know the distributions on the subspaces to draw all inferences in the domain under consideration that can be drawn using the original distribution δ.

- Since such a decomposition is usually represented as a network and since it is used to draw inferences, it can be called an **inference network**. The edges of the network indicate the paths along which evidence has to be propagated.

- Another popular name is **graphical model**, where "graphical" indicates that it is based on a *graph* in the sense of graph theory.

A Simple Example

Example World

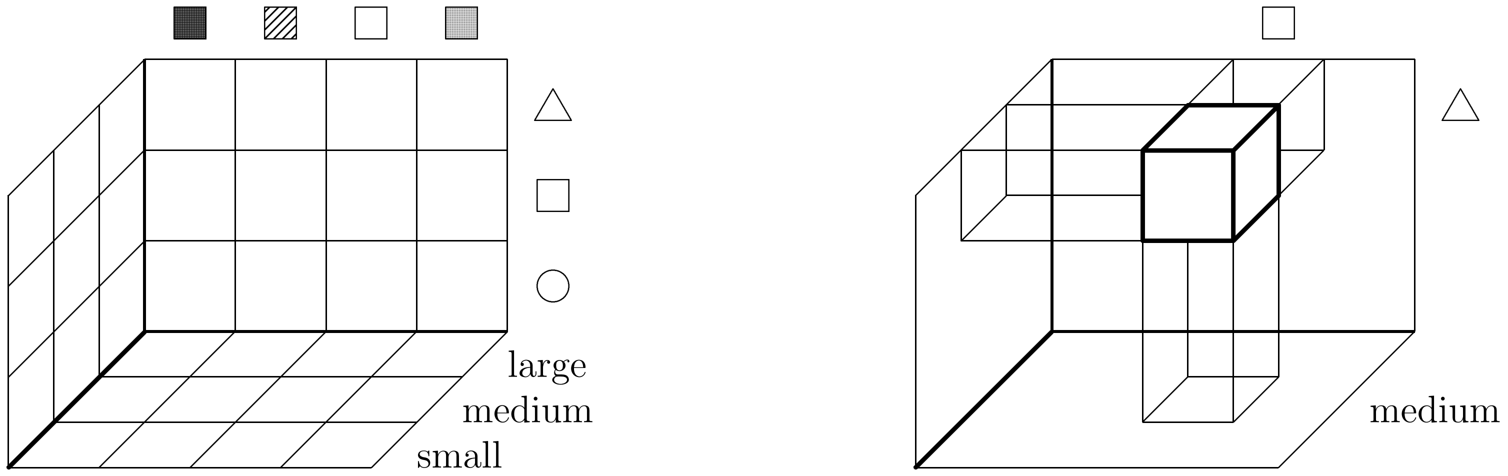


Relation

color	shape	size
■	○	small
■	○	medium
▨	○	small
▨	○	medium
▨	△	medium
▨	△	large
□	□	medium
░	□	medium
░	△	medium
░	△	large

- 10 simple geometrical objects, 3 attributes.
- One object is chosen at random and examined.
- Inferences are drawn about the unobserved attributes.

The Reasoning Space



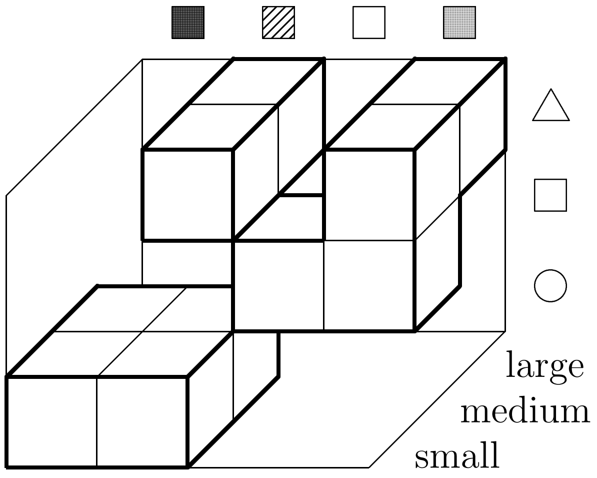
- The reasoning space consists of a finite set Ω of world states.
- The world states are described by a set of attributes A_i, whose domains $\{a_1^{(i)}, \ldots, a_{k_i}^{(i)}\}$ can be seen as sets of propositions or events.
- The events in a domain are mutually exclusive and exhaustive.
- The reasoning space is assumed to contain the true, but unknown state ω_0.

The Relation in the Reasoning Space

Relation

color	shape	size
■	○	small
■	○	medium
▨	○	small
▨	○	medium
▨	△	medium
▨	△	large
□	□	medium
░	□	medium
░	△	medium
░	△	large

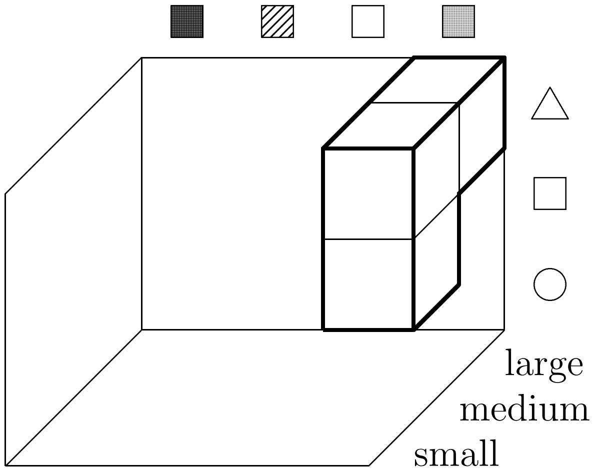
Relation in the Reasoning Space

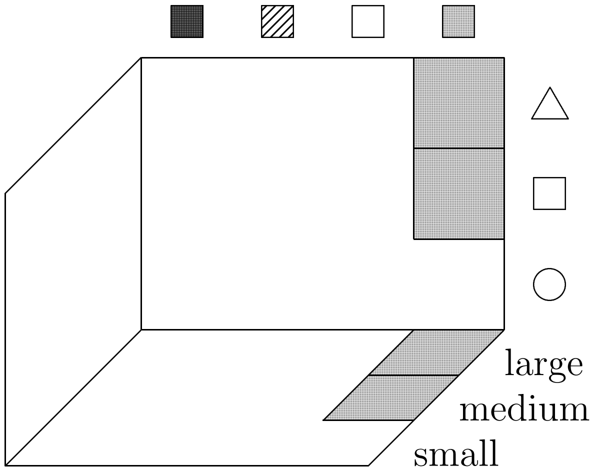


Each cube represents one tuple.

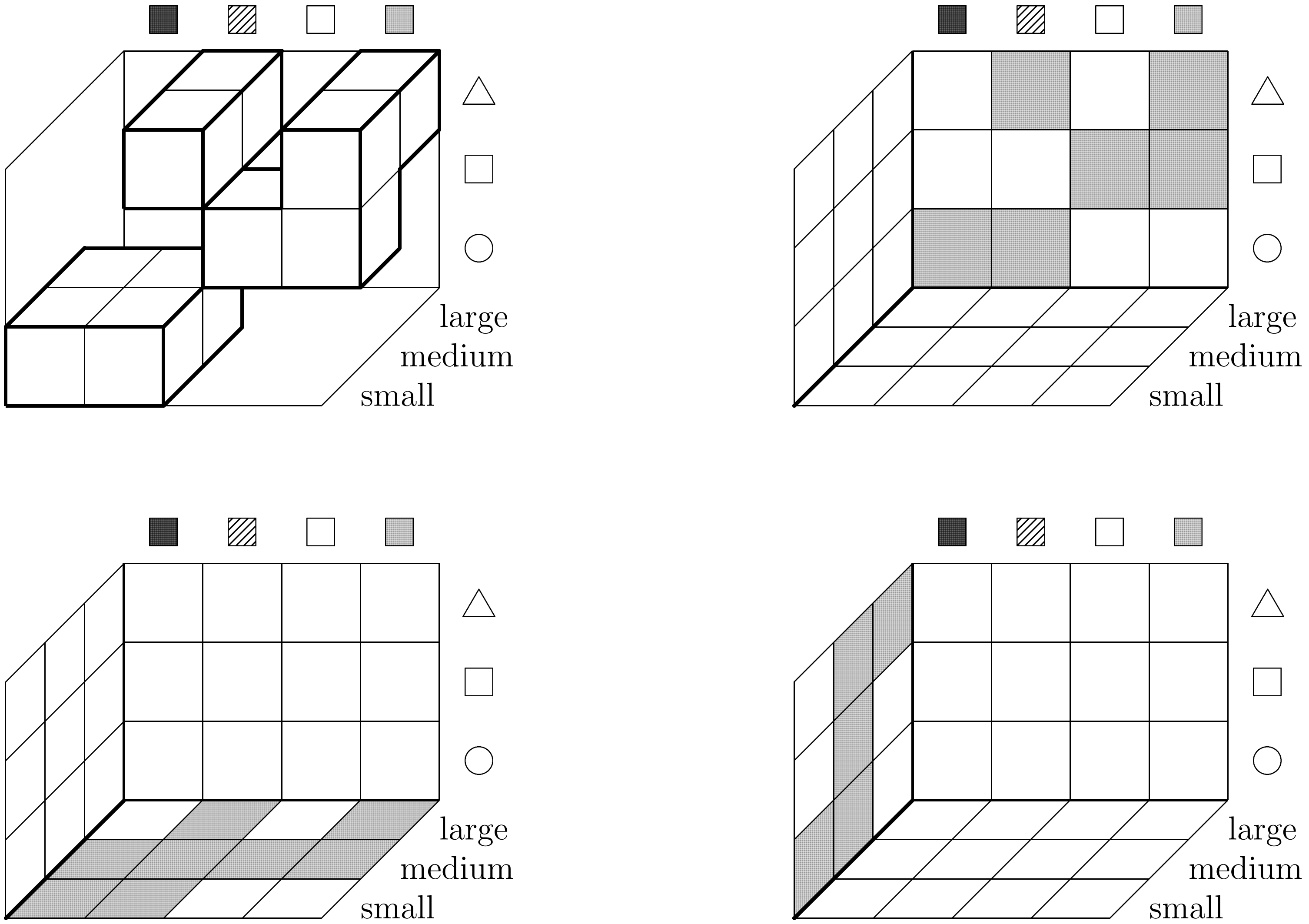
Reasoning

- Let it be known (e.g. from an observation) that the given object is green. This information considerably reduces the space of possible value combinations.
- From the prior knowledge it follows that the given object must be
  - either a triangle or a square and
  - either medium or large.

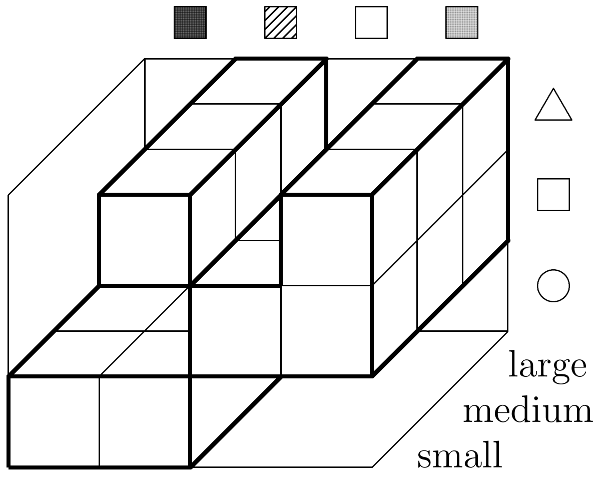




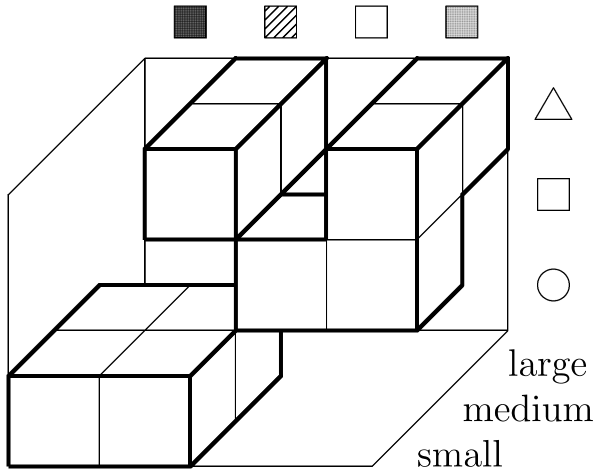
Prior Knowledge and Its Projections

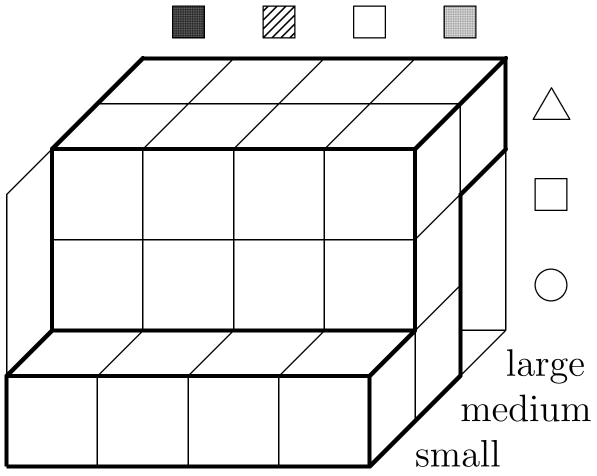


Cylindrical Extensions and Their Intersection



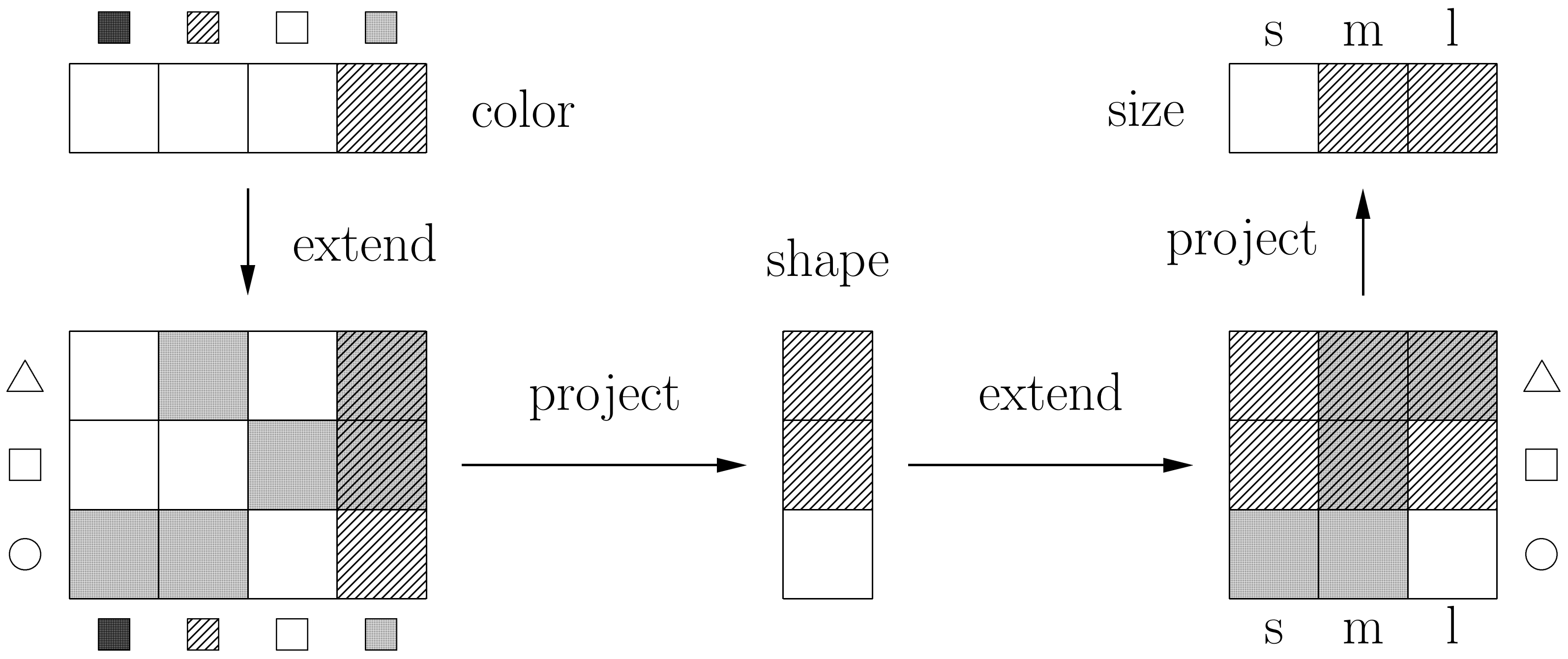
Intersecting the cylindrical extensions of the projection to the subspace formed by color and shape and of the projection to the subspace formed by shape and size yields the original three-dimensional relation.



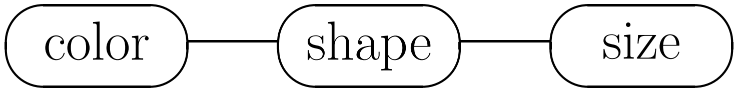


Reasoning with Projections

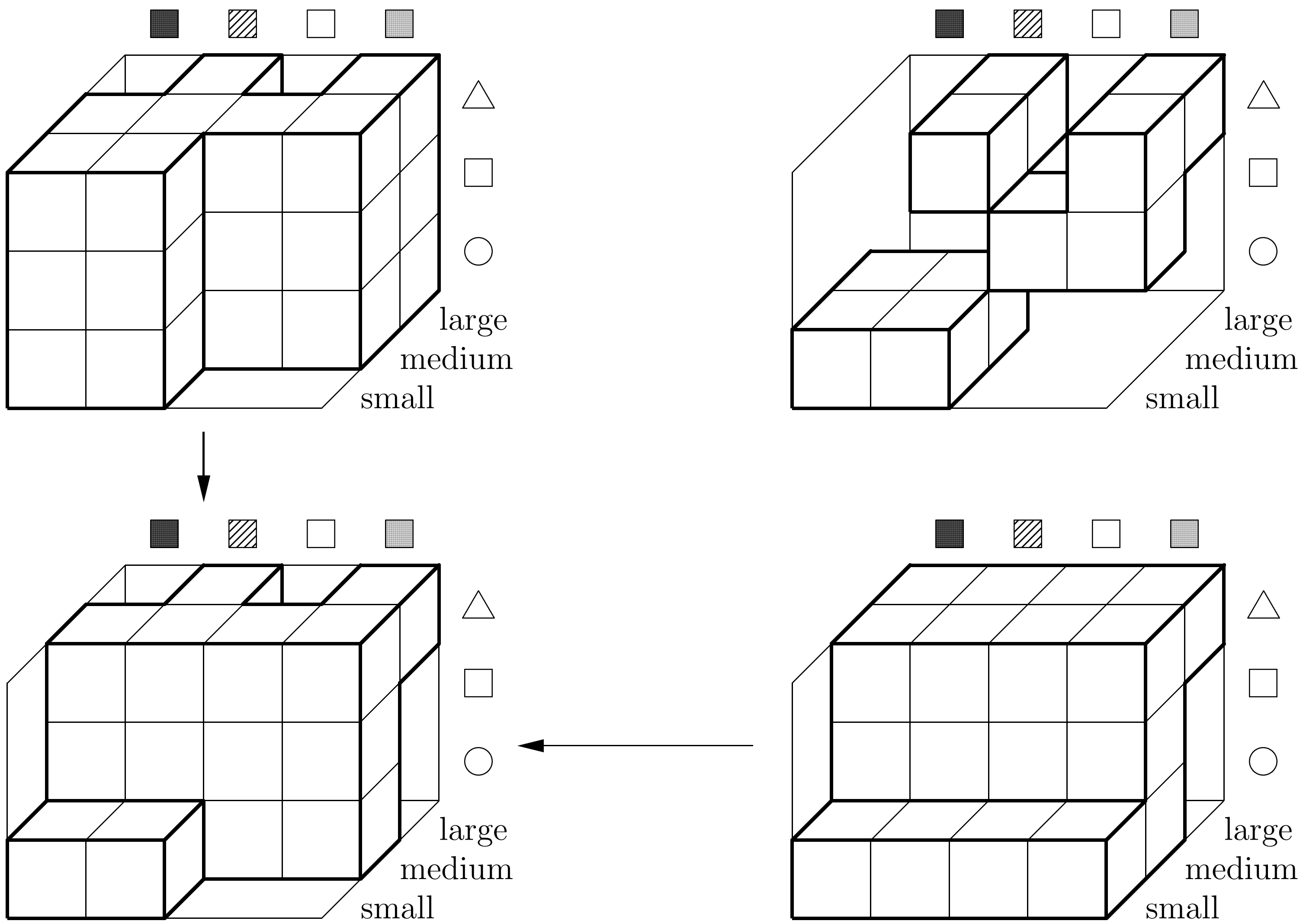
The same result can be obtained using only the projections to the subspaces without reconstructing the original three-dimensional space:



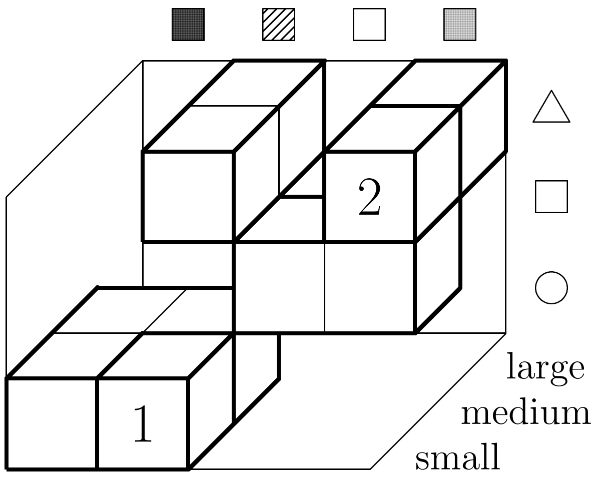
This justifies a network representation:

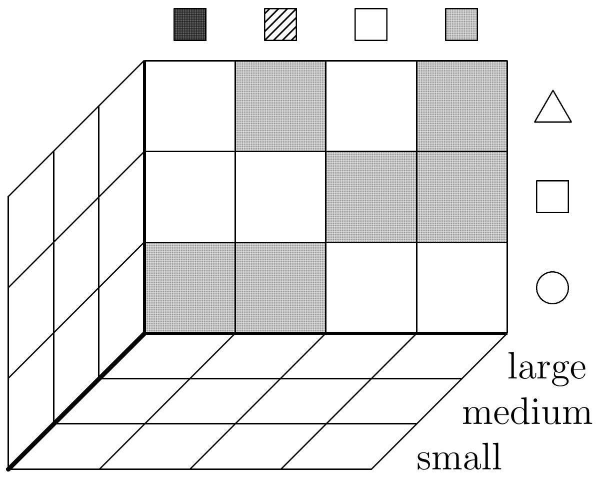


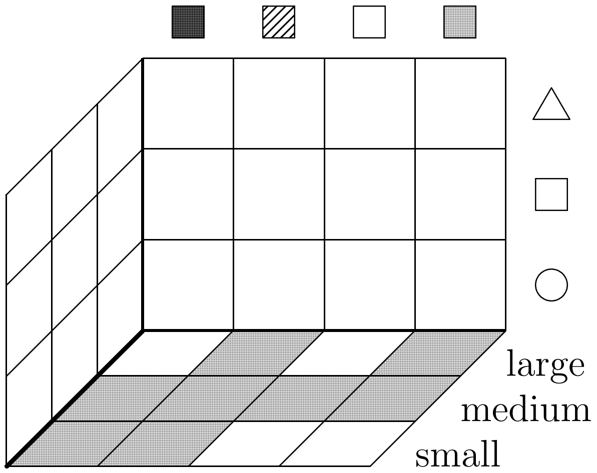
Using other Projections

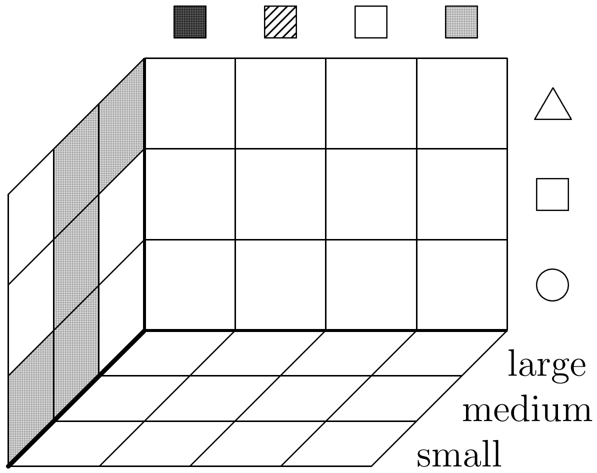


Is Decomposition Always Possible?









Possibility-Based Formalization

Definition: Let Ω be a (finite) sample space.
A **discrete possibility measure** R on Ω is a function $R : 2^{\Omega} \to \{0, 1\}$ satisfying

1. $R(\emptyset) = 0$ and
2. $\forall E_1, E_2 \subseteq \Omega : R(E_1 \cup E_2) = \max\{R(E_1), R(E_2)\}$.

- Similar to Kolmogorov's axioms of probability theory.
- If an event E can occur (if it is possible), then $R(E) = 1$, otherwise (if E cannot occur/is impossible) then $R(E) = 0$.
- $R(\Omega) = 1$ is not required, because this would exclude the empty relation.
- From the axioms it follows $R(E_1 \cap E_2) \leq \min\{R(E_1), R(E_2)\}$.
- Attributes are introduced as random variables (as in probability theory).
- $R(A = a)$ is an abbreviation of $R(\{\omega \mid A(\omega) = a\})$

Possibility-Based Formalization (continued)

Definition: Let $X = \{A_1, \ldots, A_n\}$ be a set of attributes defined on a (finite) sample space Ω with respective domains $\mathrm{dom}(A_i)$, $i = 1, \ldots, n$. A **relation** r_X over X is the restriction of a discrete possibility measure R on Ω to the set of all events that can be defined by stating values for all attributes in X. That is, $r_X = R|_{\mathcal{E}_X}$, where

$$\begin{aligned}
\mathcal{E}_X &= \Big\{ E \in 2^\Omega \;\Big|\; \exists a_1 \in \mathrm{dom}(A_1) : \ldots \exists a_n \in \mathrm{dom}(A_n) : \\
&\qquad E \mathrel{\widehat{=}} \bigwedge_{A_j \in X} A_j = a_j \Big\} \\
&= \Big\{ E \in 2^\Omega \;\Big|\; \exists a_1 \in \mathrm{dom}(A_1) : \ldots \exists a_n \in \mathrm{dom}(A_n) : \\
&\qquad E = \Big\{ \omega \in \Omega \;\Big|\; \bigwedge_{A_j \in X} A_j(\omega) = a_j \Big\} \Big\}.
\end{aligned}$$

- Corresponds to the notion of a probability distribution.
- Advantage of this formalization: No index transformation functions are needed for projections, there are just fewer terms in the conjunctions.

Possibility-Based Formalization (continued)

Definition: Let $U = \{A_1, \ldots, A_n\}$ be a set of attributes and r_U a relation over U. Furthermore, let $\mathcal{M} = \{M_1, \ldots, M_m\} \subseteq 2^U$ be a set of nonempty (but not necessarily disjoint) subsets of U satisfying

$$\bigcup_{M \in \mathcal{M}} M = U.$$

r_U is called **decomposable** w.r.t. $\mathcal{M}$ iff

$$\begin{aligned} &\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) : \\ &\quad r_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \min_{M \in \mathcal{M}} \Big\{ r_M\Big(\bigwedge_{A_i \in M} A_i = a_i\Big)\Big\}. \end{aligned}$$

If r_U is decomposable w.r.t. $\mathcal{M}$, the set of relations

$$\mathcal{R}_{\mathcal{M}} = \{r_{M_1}, \ldots, r_{M_m}\} = \{r_M \mid M \in \mathcal{M}\}$$

is called the **decomposition** of r_U.

Conditional Possibility and Independence

Definition: Let Ω be a (finite) sample space, R a discrete possibility measure on Ω, and $E_1, E_2 \subseteq \Omega$ events. Then

$$R(E_1 \mid E_2) = R(E_1 \cap E_2)$$

is called the **conditional possibility** of E_1 given E_2.

Definition: Let Ω be a (finite) sample space, R a discrete possibility measure on Ω, and A, B, and C attributes with respective domains $\mathrm{dom}(A)$, $\mathrm{dom}(B)$, and $\mathrm{dom}(C)$. A and B are called **conditionally relationally independent** given C, written $A \perp\!\!\!\perp_R B \mid C$, iff

$$\begin{aligned}&\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : \\ &\qquad R(A = a, C = c \mid B = b) = \min\{R(A = a \mid B = b), R(C = c \mid B = b)\}.\end{aligned}$$

- Similar to the corresponding notions of probability theory.

Relational Evidence Propagation, Step 1

$$
\begin{aligned}
&R(B = b \mid A = a_{\text{obs}}) \\
&\quad = \quad R\Big(\bigvee_{a \in \text{dom}(A)} A = a, B = b, \bigvee_{c \in \text{dom}(C)} C = c \,\Big|\, A = a_{\text{obs}} \Big) \\
&\quad \overset{(1)}{=} \quad \max_{a \in \text{dom}(A)} \{ \max_{c \in \text{dom}(C)} \{ R(A = a, B = b, C = c \mid A = a_{\text{obs}}) \} \} \\
&\quad \overset{(2)}{=} \quad \max_{a \in \text{dom}(A)} \{ \max_{c \in \text{dom}(C)} \{ \min\{ R(A = a, B = b, C = c), R(A = a \mid A = a_{\text{obs}}) \} \} \} \\
&\quad \overset{(3)}{=} \quad \max_{a \in \text{dom}(A)} \{ \max_{c \in \text{dom}(C)} \{ \min\{ R(A = a, B = b), R(B = b, C = c), \\
&\qquad\qquad\qquad\qquad\qquad R(A = a \mid A = a_{\text{obs}}) \} \} \} \\
&\quad = \quad \max_{a \in \text{dom}(A)} \{ \min\{ R(A = a, B = b), R(A = a \mid A = a_{\text{obs}}), \\
&\qquad\qquad\qquad \underbrace{\max_{c \in \text{dom}(C)} \{ R(B = b, C = c) \} \} \}}_{= R(B=b) \geq R(A=a, B=b)} \\
&\quad = \quad \max_{a \in \text{dom}(A)} \{ \min\{ R(A = a, B = b), R(A = a \mid A = a_{\text{obs}}) \} \}.
\end{aligned}
$$

Relational Evidence Propagation, Step 1 (continued)

(1) holds because of the second axiom a discrete possibility measure has to satisfy.

(3) holds because of the fact that the relation R_{ABC} can be decomposed w.r.t. the set $\mathcal{M} = \{\{A, B\}, \{B, C\}\}$.

(2) holds, since in the first place

$$\begin{aligned} R(A = a, B = b, C = c | A = a_{obs}) &= R(A = a, B = b, C = c, A = a_{obs}) \\ &= \begin{cases} R(A = a, B = b, C = c), & \text{if } a = a_{\text{obs}}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

and secondly

$$\begin{aligned} R(A = a \mid A = a_{\text{obs}}) &= R(A = a, A = a_{\text{obs}}) \\ &= \begin{cases} R(A = a), & \text{if } a = a_{\text{obs}}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

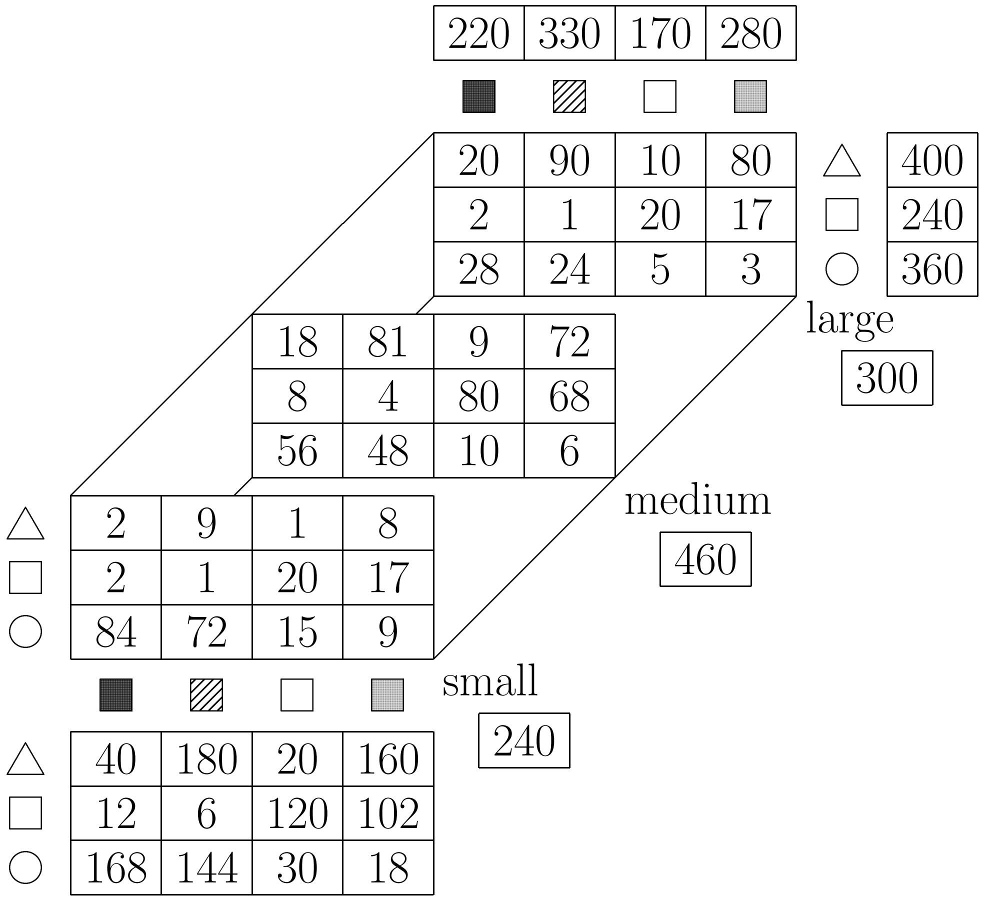
and therefore, since trivially $R(A = a) \geq R(A = a, B = b, C = c)$,

$$\begin{aligned} & R(A = a, B = b, C = c \mid A = a_{obs}) \\ & \quad = \min\{R(A = a, B = b, C = c), R(A = a \mid A = a_{\text{obs}})\}. \end{aligned}$$

Relational Evidence Propagation, Step 2

$$
\begin{aligned}
&R(C = c \mid A = a_{\text{obs}}) \\
&\quad = \quad R\Big(\bigvee_{a\in\text{dom}(A)} A = a, \bigvee_{b\in\text{dom}(B)} B = b, C = c \,\Big|\, A = a_{\text{obs}}\Big) \\
&\quad \overset{(1)}{=} \quad \max_{a\in\text{dom}(A)}\{\max_{b\in\text{dom}(B)}\{R(A = a, B = b, C = c \mid A = a_{\text{obs}})\}\} \\
&\quad \overset{(2)}{=} \quad \max_{a\in\text{dom}(A)}\{\max_{b\in\text{dom}(B)}\{\min\{R(A = a, B = b, C = c), R(A = a \mid A = a_{\text{obs}})\}\}\} \\
&\quad \overset{(3)}{=} \quad \max_{a\in\text{dom}(A)}\{\max_{b\in\text{dom}(B)}\{\min\{R(A = a, B = b), R(B = b, C = c), \\
&\qquad\qquad\qquad R(A = a \mid A = a_{\text{obs}})\}\}\} \\
&\quad = \quad \max_{b\in\text{dom}(B)}\{\min\{R(B = b, C = c), \\
&\qquad\qquad \underbrace{\max_{a\in\text{dom}(A)}\{\min\{R(A = a, B = b), R(A = a \mid A = a_{\text{obs}})\}\}}_{=R(B=b|A=a_{\text{obs}})}\}\} \\
&\quad = \quad \max_{b\in\text{dom}(B)}\{\min\{R(B = b, C = c), R(B = b \mid A = a_{\text{obs}})\}\}.
\end{aligned}
$$

A Probability Distribution



220	330	170	280
■	▨	□	▒

large

	■	▨	□	▒
△	20	90	10	80
□	2	1	20	17
○	28	24	5	3

△	400
□	240
○	360

large 300

medium

	■	▨	□	▒
△	18	81	9	72
□	8	4	80	68
○	56	48	10	6

medium 460

small

	■	▨	□	▒
△	2	9	1	8
□	2	1	20	17
○	84	72	15	9

small 240

	■	▨	□	▒
△	40	180	20	160
□	12	6	120	102
○	168	144	30	18

all numbers in
parts per 1000

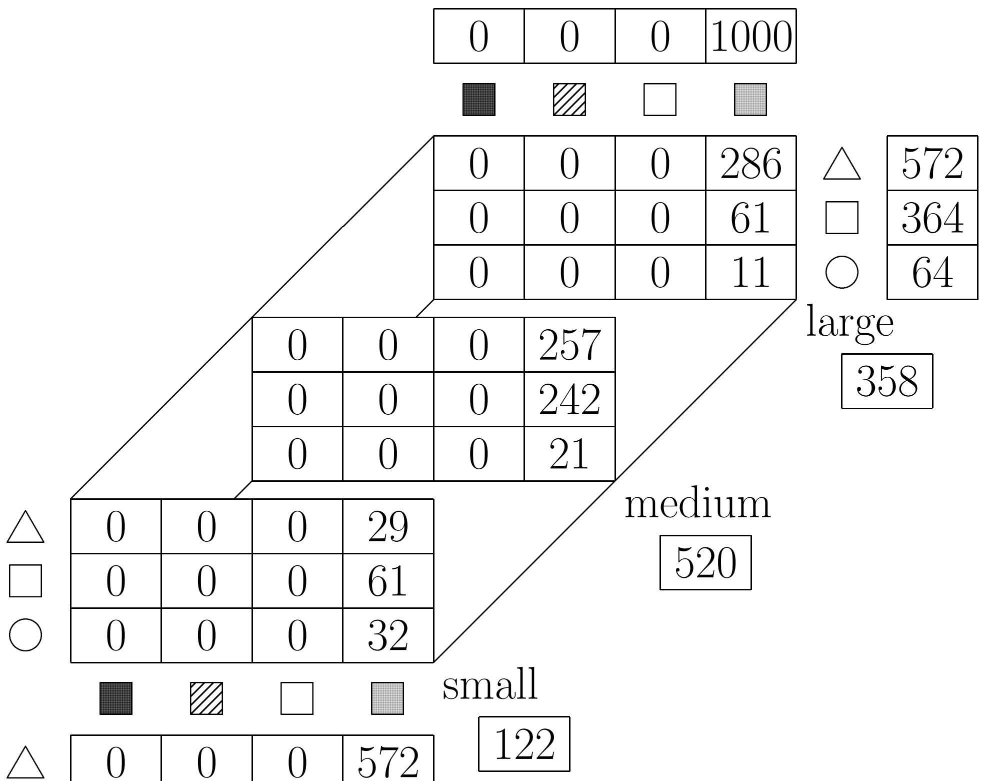
	s	m	l
△	20	180	200
□	40	160	40
○	180	120	60

	■	▨	□	▒
large	50	115	35	100
medium	82	133	99	146
small	88	82	36	34

- The numbers state the probability of the corresponding value combination.

Reasoning

all numbers in
parts per 1000



■	▨	□	░
0	0	0	1000

large

	■	▨	□	░	
△	0	0	0	286	572
□	0	0	0	61	364
○	0	0	0	11	64

358

medium

	■	▨	□	░
△	0	0	0	257
□	0	0	0	242
○	0	0	0	21

520

small

	■	▨	□	░
△	0	0	0	29
□	0	0	0	61
○	0	0	0	32

122

	■	▨	□	░
△	0	0	0	572
□	0	0	0	364
○	0	0	0	64

	s	m	l
△	29	257	286
□	61	242	61
○	32	21	11

	■	▨	□	░
large	0	0	0	358
medium	0	0	0	531
small	0	0	0	111

- Using the information that the given object is green.

Probabilistic Decomposition

- As for relational networks, the three-dimensional probability distribution can be decomposed into projections to subspaces, namely the marginal distribution on the subspace formed by color and shape and the marginal distribution on the subspace formed by shape and size.
- The original probability distribution can be reconstructed from the marginal distributions using the following formulae $\forall i, j, k :$

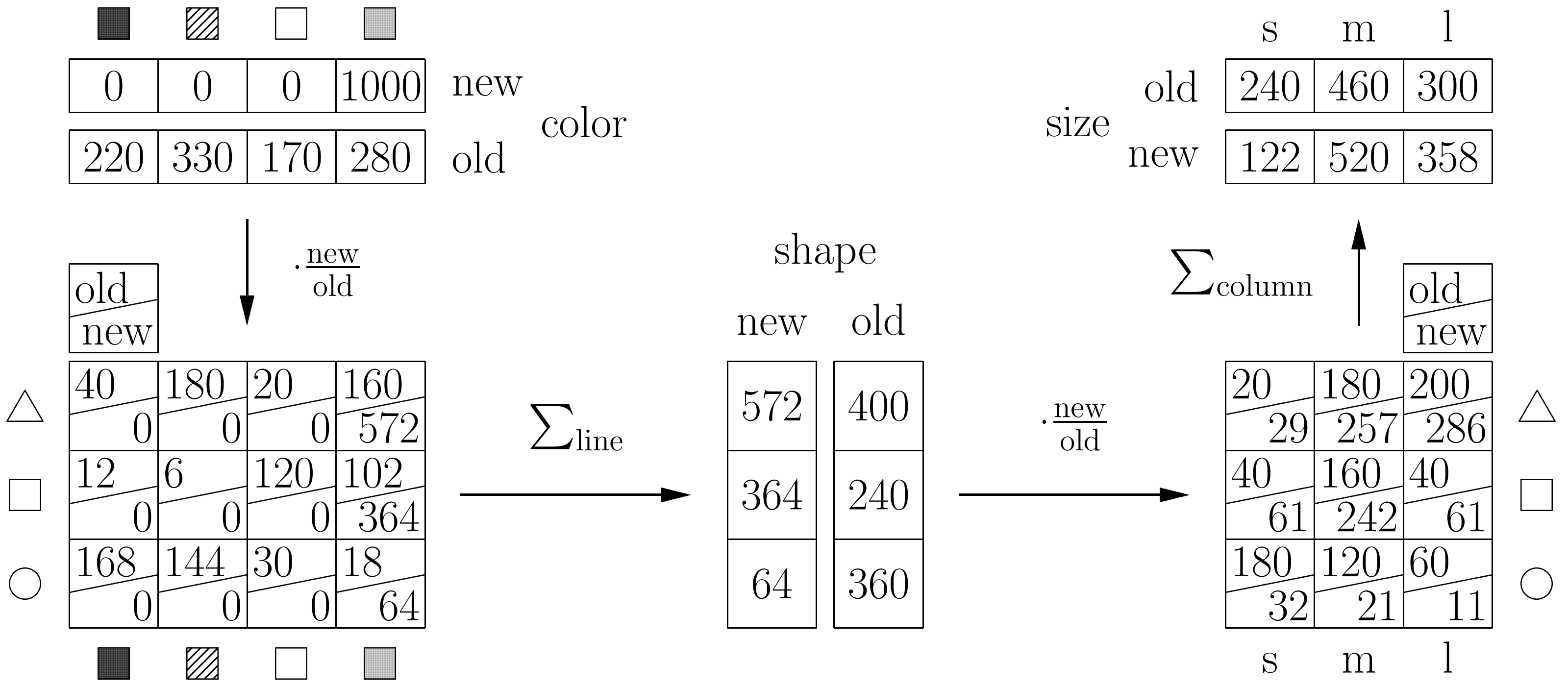
$$\begin{aligned} P(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}, \omega_k^{(\text{size})}) &= P(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}) \cdot P(\omega_k^{(\text{size})} \mid \omega_j^{(\text{shape})}) \\ &= P(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}) \cdot \frac{P(\omega_j^{(\text{shape})}, \omega_k^{(\text{size})})}{P(\omega_j^{(\text{shape})})} \end{aligned}$$

- These equations express the *conditional independence* of attributes *color* and *size* given the attribute *shape*, since they only hold if $\forall i, j, k :$

$$P(\omega_k^{(\text{size})} \mid \omega_j^{(\text{shape})}) = P(\omega_k^{(\text{size})} \mid \omega_i^{(\text{color})}, \omega_j^{(\text{shape})})$$

Reasoning with Projections

Again the same result can be obtained using only projections to subspaces (marginal distributions):



color

	■	▨	□	▩
new	0	0	0	1000
old	220	330	170	280

$\downarrow \cdot \frac{\text{new}}{\text{old}}$

old / new

	■	▨	□	▩
△	40 0	180 0	20 0	160 572
□	12 0	6 0	120 0	102 364
○	168 0	144 0	30 0	18 64

$\xrightarrow{\sum_{\text{line}}}$

shape

new	old
572	400
364	240
64	360

$\xrightarrow{\cdot \frac{\text{new}}{\text{old}}}$

old / new

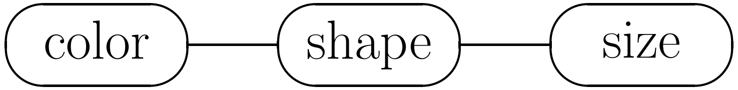
	s	m	l
△	20 29	180 257	200 286
□	40 61	160 242	40 61
○	180 32	120 21	60 11

$\uparrow \sum_{\text{column}}$

size

	s	m	l
old	240	460	300
new	122	520	358

This justifies a network representation:



Probabilistic Decomposition (continued)

Definition: Let $U = \{A_1, \ldots, A_n\}$ be a set of attributes and p_U a probability distribution over U. Furthermore, let $\mathcal{M} = \{M_1, \ldots, M_m\} \subseteq 2^U$ be a set of nonempty (but not necessarily disjoint) subsets of U satisfying

$$\bigcup_{M \in \mathcal{M}} M = U.$$

p_U is called **decomposable** or **factorizable** w.r.t. $\mathcal{M}$ iff it can be written as a product of m nonnegative functions $\phi_M : \mathcal{E}_M \to \mathbb{R}_0^+$, $M \in \mathcal{M}$, i.e., iff

$$\begin{aligned} &\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) : \\ &\qquad p_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \prod_{M \in \mathcal{M}} \phi_M\Big(\bigwedge_{A_i \in M} A_i = a_i\Big). \end{aligned}$$

If p_U is decomposable w.r.t. $\mathcal{M}$ the set of functions

$$\Phi_{\mathcal{M}} = \{\phi_{M_1}, \ldots, \phi_{M_m}\} = \{\phi_M \mid M \in \mathcal{M}\}$$

is called the **decomposition** or the **factorization** of p_U. The functions in $\Phi_{\mathcal{M}}$ are called the **factor potentials** of p_U.

Conditional Probability and Independence

Definition: Let Ω be a (finite) sample space, P a probability measure on Ω, and $E_1, E_2 \subseteq \Omega$ events with $P(E_2) > 0$. Then

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

is called the **conditional probability** of E_1 given E_2.

Definition: Let Ω be a (finite) sample space, P a probability measure on Ω, and A, B, and C attributes with respective domains $\mathrm{dom}(A)$, $\mathrm{dom}(B)$, and $\mathrm{dom}(C)$. A and B are called **conditionally probabilistically independent** given C, written $A \perp\!\!\!\perp_P B \mid C$, iff

$$\begin{aligned}&\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : \\ &\qquad P(A = a, B = b \mid C = c) = P(A = a \mid C = c) \cdot P(B = b \mid C = c)\end{aligned}$$

Equivalent Formula:

$$\begin{aligned}&\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \forall c \in \mathrm{dom}(C) : \\ &\qquad P(A = a \mid B = b, C = c) = P(A = a \mid C = c)\end{aligned}$$

Probabilistic Decomposition (continued)

Chain Rule of Probability:

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$P\Big(\textstyle\bigwedge_{i=1}^{n} A_i = a_i\Big) = \displaystyle\prod_{i=1}^{n} P\Big(A_i = a_i \,\Big|\, \textstyle\bigwedge_{j=1}^{i-1} A_j = a_j\Big)$$

- The chain rule of probability is valid in general (or at least for strictly positive distributions).

Chain Rule Factorization:

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$P\Big(\textstyle\bigwedge_{i=1}^{n} A_i = a_i\Big) = \displaystyle\prod_{i=1}^{n} P\Big(A_i = a_i \,\Big|\, \textstyle\bigwedge_{A_j \in \mathrm{parents}(A_i)} A_j = a_j\Big)$$

- Conditional independence statements are used to “cancel” conditions.

Probabilistic Evidence Propagation, Step 1

$$
\begin{aligned}
&P(B = b \mid A = a_{\text{obs}}) \\
&\quad = \quad P\Big(\bigvee_{a \in \text{dom}(A)} A = a, B = b, \bigvee_{c \in \text{dom}(C)} C = c \Big| A = a_{\text{obs}}\Big) \\
&\quad \overset{(1)}{=} \quad \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} P(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\quad \overset{(2)}{=} \quad \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a_i)} \\
&\quad \overset{(3)}{=} \quad \sum_{a \in \text{dom}(A)} \sum_{c \in \text{dom}(C)} \frac{P(A = a, B = b) P(B = b, C = c)}{P(B = b)} \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)} \\
&\quad = \quad \sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)} \underbrace{\sum_{c \in \text{dom}(C)} P(C = c \mid B = b)}_{=1} \\
&\quad = \quad \sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)}.
\end{aligned}
$$

Probabilistic Evidence Propagation, Step 1 (continued)

(1) holds because of Kolmogorov's axioms.

(3) holds because of the fact that the distribution p_{ABC} can be decomposed w.r.t. the set $\mathcal{M} = \{\{A, B\}, \{B, C\}\}$.

(2) holds, since in the first place

$$\begin{aligned} P(A = a, B = b, C = c \,|\, A = a_{obs}) &= \frac{P(A = a, B = b, C = c, A = a_{obs})}{P(A = a_{\text{obs}})} \\ &= \begin{cases} \dfrac{P(A = a, B = b, C = c)}{P(A = a_{\text{obs}})}, & \text{if } a = a_{\text{obs}}, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

and secondly

$$P(A = a, A = a_{\text{obs}}) = \begin{cases} P(A = a), & \text{if } a = a_{\text{obs}}, \\ 0, & \text{otherwise}, \end{cases}$$

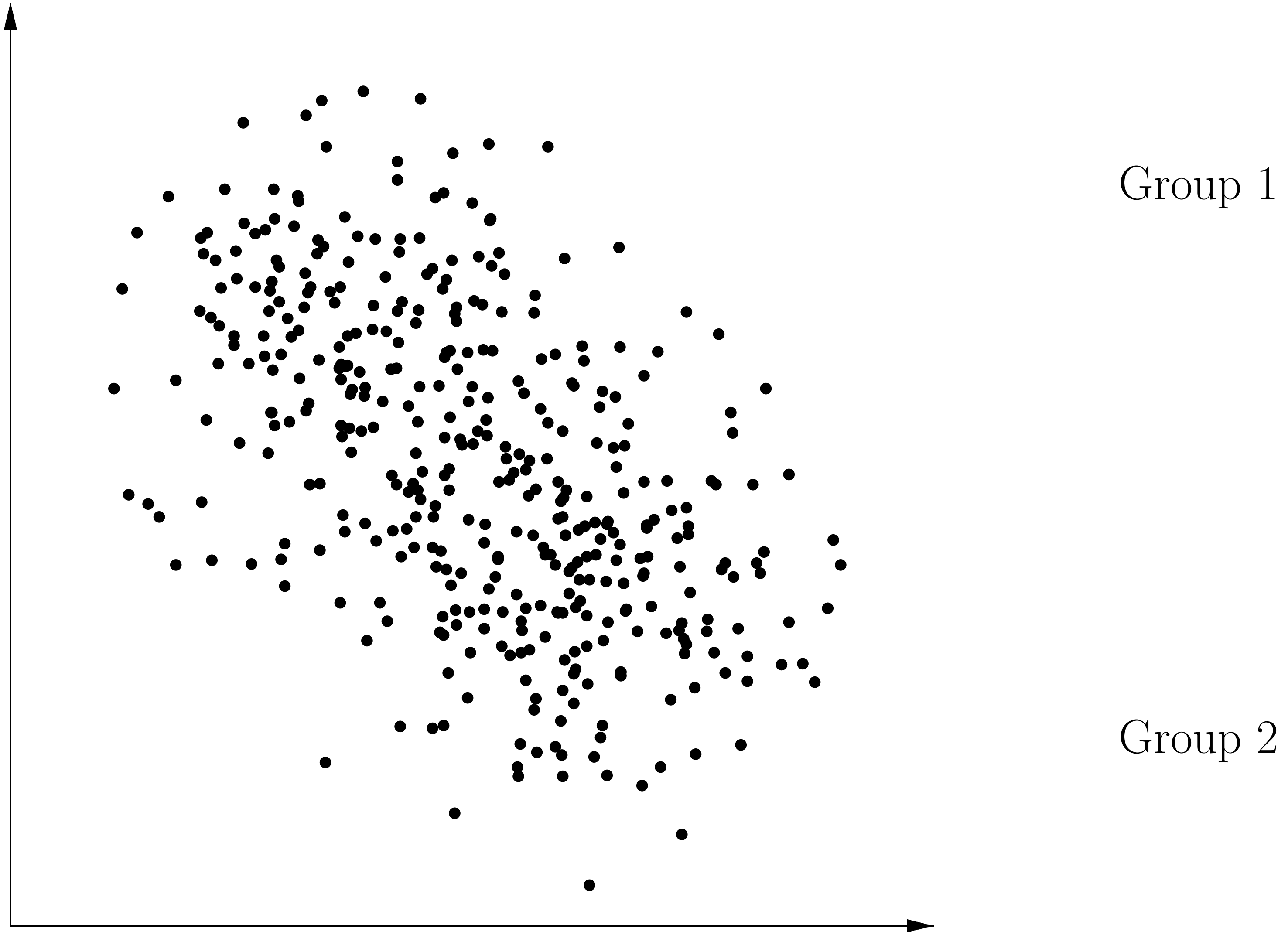
and therefore

$$\begin{aligned} &P(A = a, B = b, C = c \mid A = a_{obs}) \\ &\quad = P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)}. \end{aligned}$$

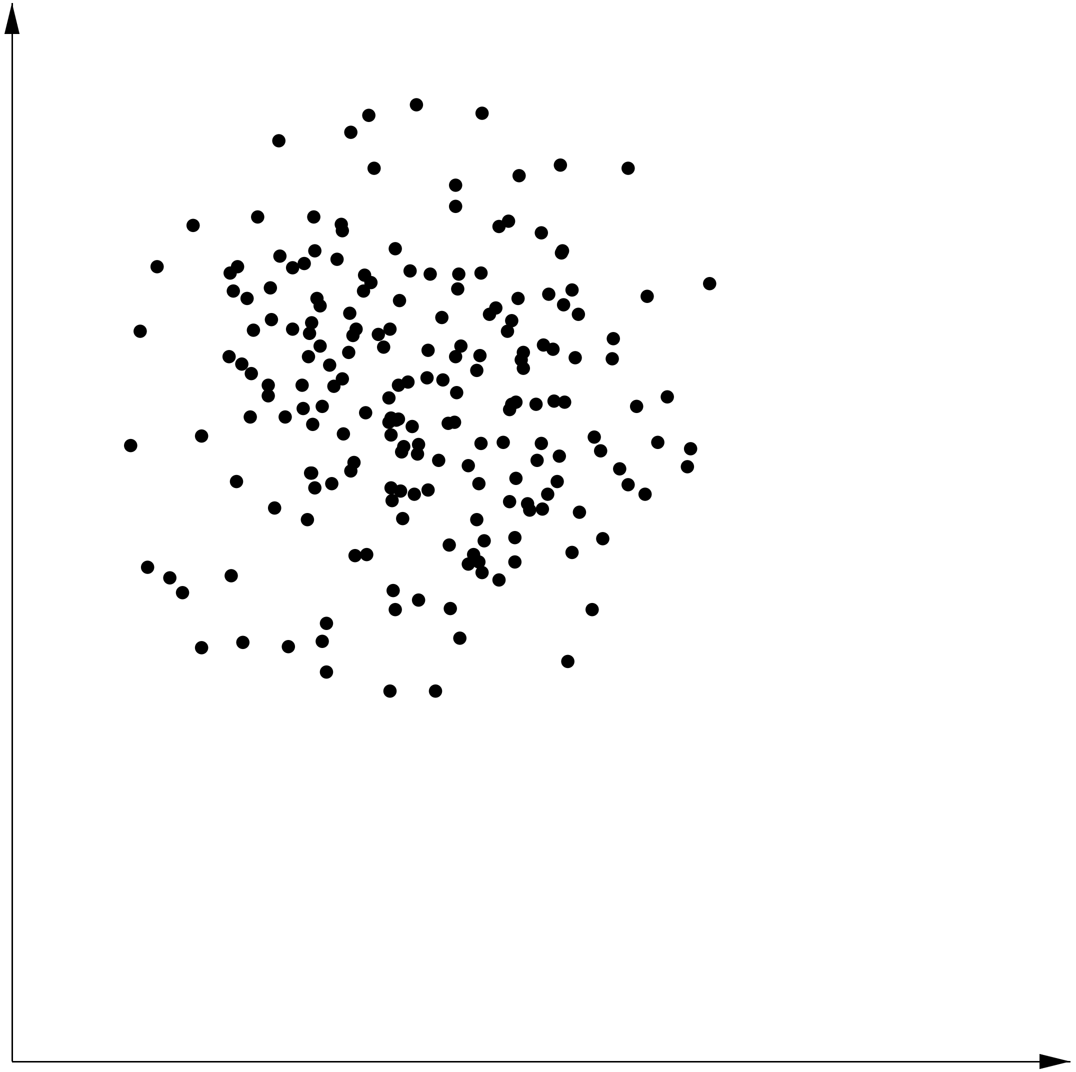
Probabilistic Evidence Propagation, Step 2

$$
\begin{aligned}
&P(C = c \mid A = a_{\text{obs}}) \\
&\quad= P\Big(\bigvee_{a \in \text{dom}(A)} A = a, \bigvee_{b \in \text{dom}(B)} B = b, C = c \Big| A = a_{\text{obs}} \Big) \\
&\quad\overset{(1)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(A = a, B = b, C = c \mid A = a_{\text{obs}}) \\
&\quad\overset{(2)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(A = a, B = b, C = c) \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)} \\
&\quad\overset{(3)}{=} \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} \frac{P(A = a, B = b) P(B = b, C = c)}{P(B = b)} \cdot \frac{P(A = a \mid A = a_{\text{obs}})}{P(A = a)} \\
&\quad= \sum_{b \in \text{dom}(B)} \frac{P(B = b, C = c)}{P(B = b)} \underbrace{\sum_{a \in \text{dom}(A)} P(A = a, B = b) \cdot \frac{R(A = a \mid A = a_{\text{obs}})}{P(A = a)}}_{= P(B = b \mid A = a_{\text{obs}})} \\
&\quad= \sum_{b \in \text{dom}(B)} P(B = b, C = c) \cdot \frac{P(B = b \mid A = a_{\text{obs}})}{P(B = b)}.
\end{aligned}
$$

Conditional Independence: An Example

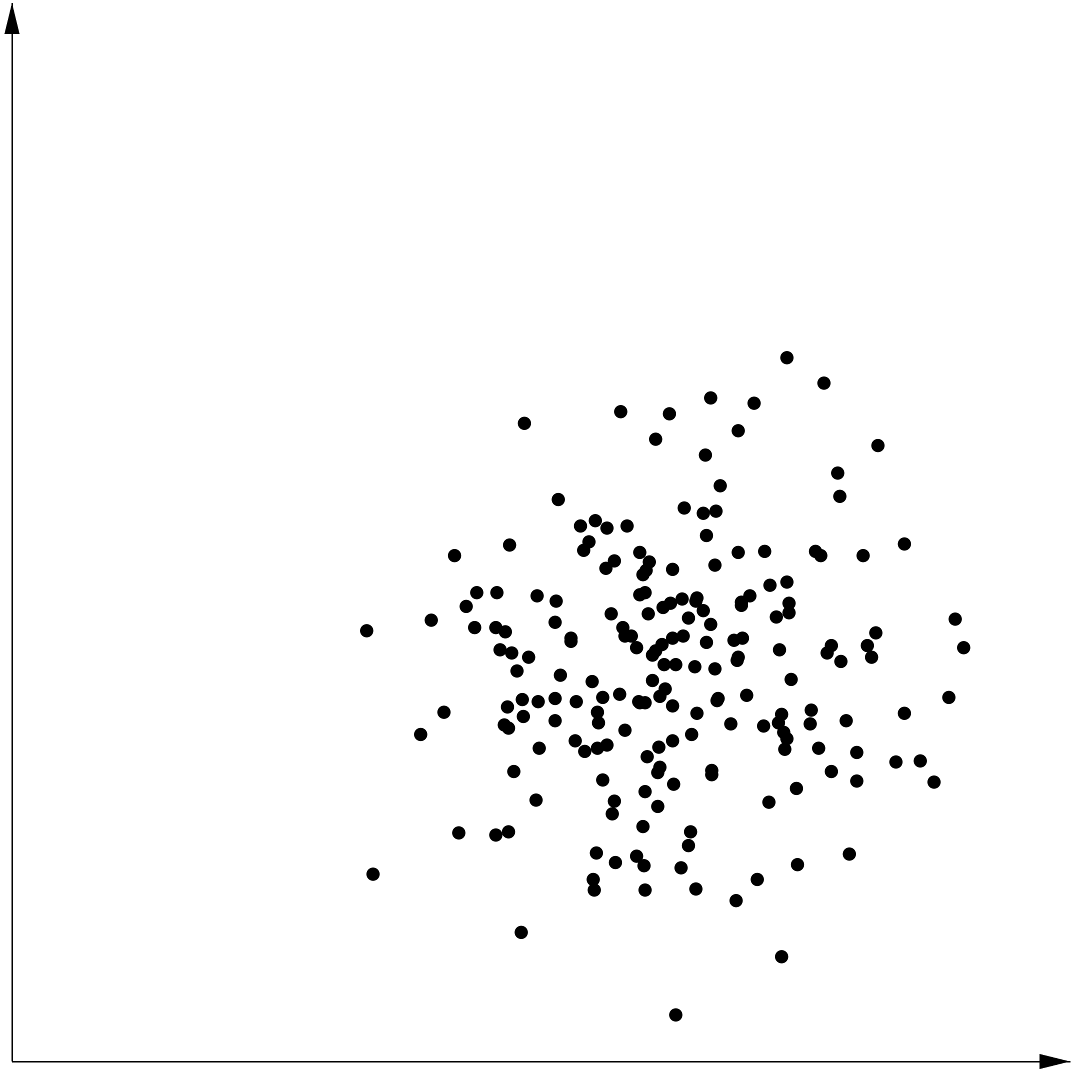


Conditional Independence: An Example



Group 1

Conditional Independence: An Example



Group 2

Axioms of Conditional Independence

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

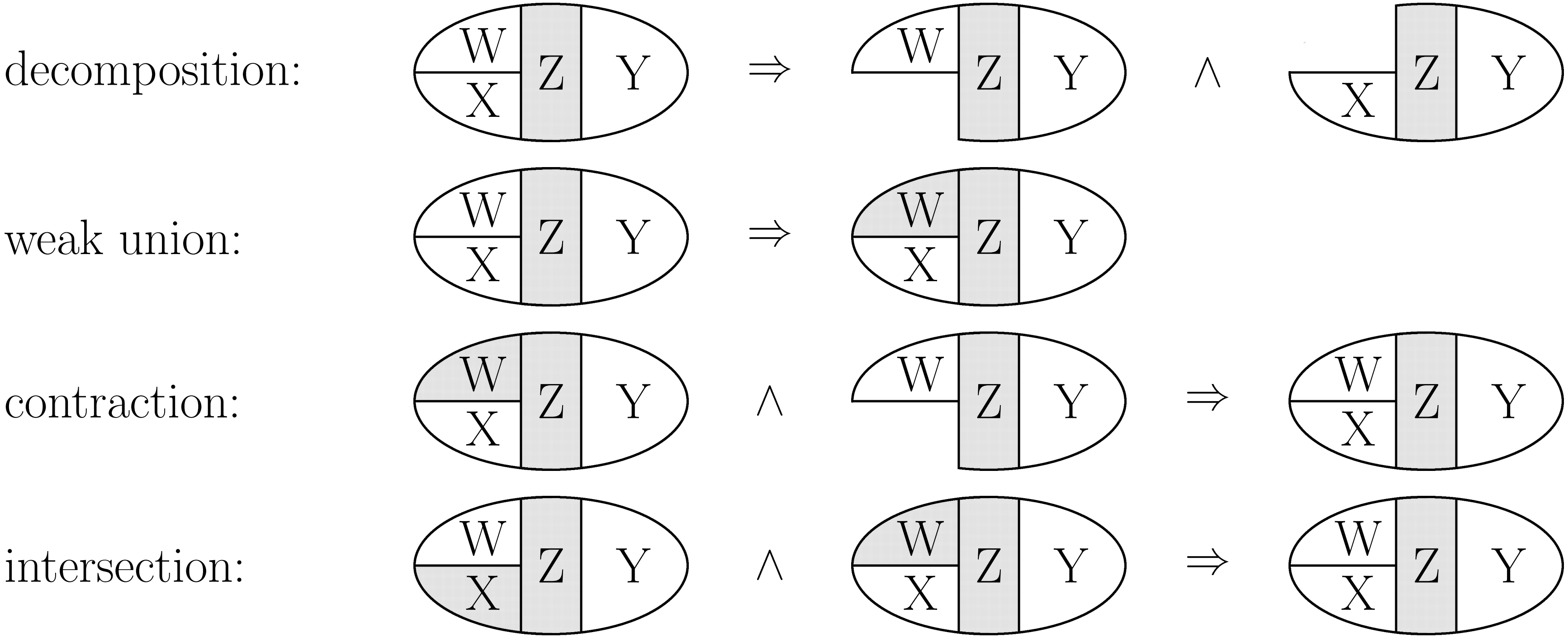
contraction: $(X \perp\!\!\!\perp Y \mid Z \cup W) \wedge (W \perp\!\!\!\perp Y \mid Z) \;\Rightarrow\; (W \cup X \perp\!\!\!\perp Y \mid Z)$

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**.

Illustration of the (Semi-)Graphoid Axioms



decomposition:

weak union:

contraction:

intersection:

- 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**.

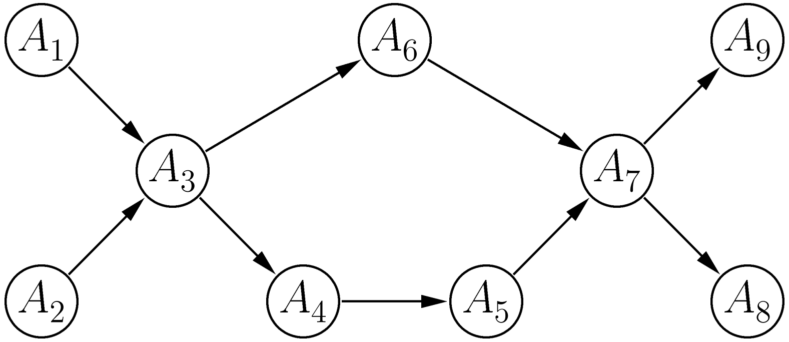
Definition: Let $\vec{G} = (V, \vec{E})$ be a directed acyclic graph and X, Y, and Z three disjoint subsets of nodes. Z **d-separates** X and Y in $\vec{G}$, written $\langle X \mid Z \mid Y \rangle_{\vec{G}}$, iff there is no path from a node in X to a node in Y along which the following two conditions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

$\langle\{A_1\} \mid \{A_3\} \mid \{A_4\}\rangle$ $\quad$ $\langle\{A_8\} \mid \{A_7\} \mid \{A_9\}\rangle$

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

Invalid Separations:

$\langle\{A_1\} \mid \{A_4\} \mid \{A_2\}\rangle$ $\quad$ $\langle\{A_1\} \mid \{A_6\} \mid \{A_7\}\rangle$

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

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

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

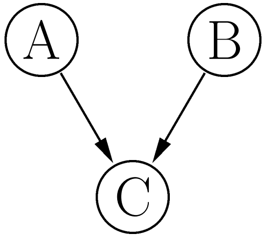
i.e., if G captures by u-separation all (conditional) independences that hold in δ and thus represents only valid (conditional) dependences. Similarly, G is called a **conditional independence graph** or an **independence map** w.r.t. δ iff for all disjoint subsets $X, Y, Z \subseteq U$ of attributes

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

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.

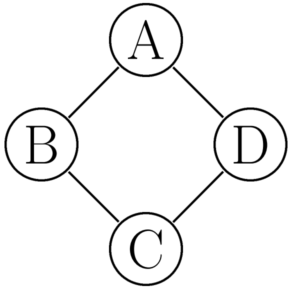
Limitations of Graph Representations

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

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

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

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

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

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

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

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

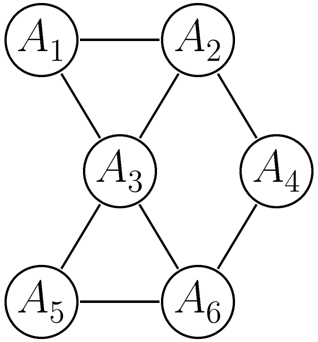
Theorem: If a three-place relation $(\cdot \perp\!\!\!\perp_\delta \cdot \mid \cdot)$ representing the set of conditional independence statements that hold in a given joint distribution δ over a set U of attributes satisfies the graphoid axioms, then the pairwise, the local, and the global Markov property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

Undirected Graphs and Decompositions

Definition: A probability distribution p_U over a set U of attributes is called **decomposable** or **factorizable w.r.t. an undirected graph** $G = (U, E)$ over U, 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 attributes, such that the subgraphs of G induced by the sets $M \in \mathcal{M}$ are the maximal cliques of G. Then there must exist functions $\phi_M : \mathcal{E}_M \to \mathbb{R}_0^+$, $M \in \mathcal{M}$,

$$\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :$$
$$p_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \prod_{M \in \mathcal{M}} \phi_M\Big(\bigwedge_{A_i \in M} A_i = a_i\Big).$$

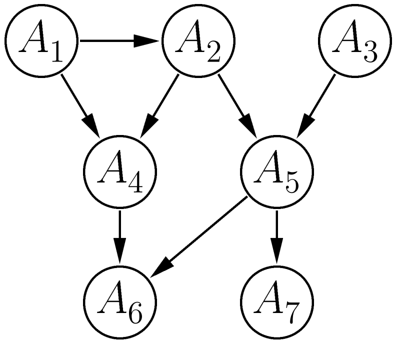


$$\begin{aligned} p_U(A_1 = a_1, \ldots, A_6 = a_6) &= \phi_{A_1A_2A_3}(A_1 = a_1, A_2 = a_2, A_3 = a_3) \\ &\cdot \phi_{A_3A_5A_6}(A_3 = a_3, A_5 = a_5, A_6 = a_6) \\ &\cdot \phi_{A_2A_4}(A_2 = a_2, A_4 = a_4) \\ &\cdot \phi_{A_4A_6}(A_4 = a_4, A_6 = a_6). \end{aligned}$$

Directed Acyclic Graphs and Decompositions

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

$$\begin{aligned}&\forall a_1 \in \mathrm{dom}(A_1) : \ldots \forall a_n \in \mathrm{dom}(A_n) :\\ &\qquad p_U\Big(\bigwedge_{A_i \in U} A_i = a_i\Big) = \prod_{A_i \in U} P\Big(A_i = a_i \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}_{\vec{G}}(A_i)} A_j = a_j\Big).\end{aligned}$$



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

Conditional Independence Graphs and Decompositions

Theorem: Let p_U be a strictly positive probability distribution on a set U of (discrete) attributes. An undirected graph $G = (U, E)$ is a conditional independence graph w.r.t. p_U, if and only if p_U is factorizable w.r.t. G.

Theorem: Let p_U be a probability distribution on a set U of (discrete) attributes. A directed acyclic graph $\vec{G} = (U, \vec{E})$ is a conditional independence graph w.r.t. p_U, if and only if p_U is factorizable w.r.t. $\vec{G}$.

Definition: A **Markov network** is an undirected conditional independence graph of a probability distribution p_U 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.

Naive Bayes Classifiers

- Try to compute $P(C = c_i \mid \omega) = P(C = c_i \mid A_1 = a_1, \ldots, A_n = a_n)$.
- Predict the class with the highest conditional probability.

Bayes' Rule:

$$P(C = c_i \mid \omega) = \frac{P(A_1 = a_1, \ldots, A_n = a_n \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_n = a_n)} \qquad \leftarrow p_0$$

Chain Rule of Probability:

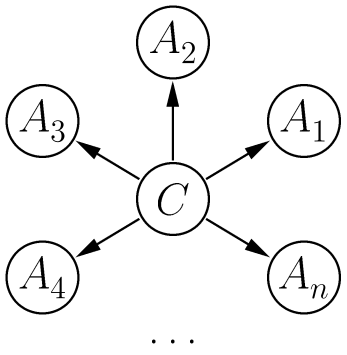
$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{j=1}^{n} P(A_j = a_j \mid A_1 = a_1, \ldots, A_{j-1} = a_{j-1}, C = c_i)$$

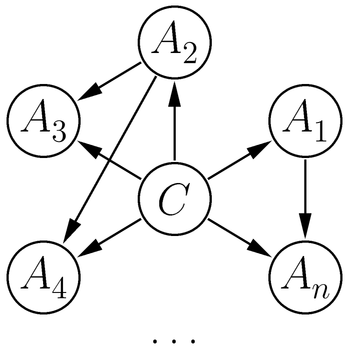
Conditional Independence Assumption:

$$P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{j=1}^{n} P(A_j = a_j \mid C = c_i)$$

Star-like Bayesian Networks

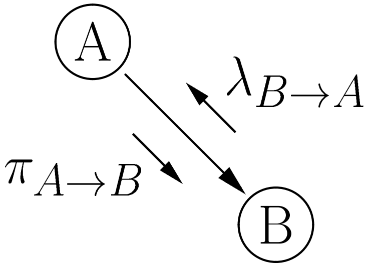
- A naive Bayes classifier is a Bayesian network with a star-like structure.
- The class attribute is the only unconditioned attribute.
- All other attributes are conditioned on the class only.





$$P(C = c_i, \omega) = P(C = c_i \mid \omega) \cdot p_0 = P(C = c_i) \cdot \prod_{j=1}^{n} P(A_j = a_j \mid C = c_i)$$

Evidence Propagation in Polytrees



Idea: Node processors communicating by message passing: π-messages are sent from parent to child and λ-messages are sent from child to parent.

Derivation of the Propagation Formulae

Computation of Marginal Distribution:

$$P(A_g = a_g) \;=\; \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j = a_j \Big)$$

Chain Rule Factorization w.r.t. the Polytree:

$$P(A_g = a_g) \;=\; \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j \Big)$$

Evidence Propagation in Polytrees (continued)

Decomposition w.r.t. Subgraphs:

$$\begin{aligned}
P(A_g = a_g) \quad = \quad & \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \operatorname{dom}(A_i)}} \Big(P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_g)} A_j = a_j \Big) \\
& \cdot \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_k)} A_j = a_j \Big) \\
& \cdot \prod_{A_k \in U_-(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_k)} A_j = a_j \Big)\Big).
\end{aligned}$$

Attribute sets underlying subgraphs:

$$U_B^A(C) = \{C\} \cup \{D \in U \mid D \underset{\vec{G}'}{\sim} C, \vec{G}' = (U, E - \{(A, B)\})\},$$

$$U_+(A) = \bigcup_{C \in \text{parents}(A)} U_A^C(C), \qquad U_+(A, B) = \bigcup_{C \in \text{parents}(A) - \{B\}} U_A^C(C),$$

$$U_-(A) = \bigcup_{C \in \text{children}(A)} U_C^A(C), \qquad U_-(A, B) = \bigcup_{C \in \text{children}(A) - \{B\}} U_A^C(C).$$

Evidence Propagation in Polytrees (continued)

Terms that are independent of a summation variable can be moved out of the corresponding sum. This yields a decomposition into two main factors:

$$\begin{aligned}
P(A_g = a_g) &= \Big(\sum_{\substack{\forall A_i \in \text{parents}(A_g):\\ a_i \in \text{dom}(A_i)}} P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_g)} A_j = a_j\Big) \\
&\qquad \cdot \Big[\sum_{\substack{\forall A_i \in U_+^*(A_g):\\ a_i \in \text{dom}(A_i)}} \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_k)} A_j = a_j\Big)\Big]\Big) \\
&\qquad \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_g):\\ a_i \in \text{dom}(A_i)}} \prod_{A_k \in U_-(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_k)} A_j = a_j\Big)\Big] \\
&= \pi(A_g = a_g) \cdot \lambda(A_g = a_g),
\end{aligned}$$

where $U_+^*(A_g) = U_+(A_g) - \text{parents}(A_g)$.

Evidence Propagation in Polytrees (continued)

$$
\begin{aligned}
&\sum_{\substack{\forall A_i \in U_+^*(A_g):\\ a_i \in \mathrm{dom}(A_i)}} \;\prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\\
=\;& \prod_{A_p \in \mathrm{parents}(A_g)} \Big(\sum_{\substack{\forall A_i \in \mathrm{parents}(A_p):\\ a_i \in \mathrm{dom}(A_i)}} P\Big(A_p = a_p \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_p)} A_j = a_j\Big)\\
&\quad \cdot \Big[\sum_{\substack{\forall A_i \in U_+^*(A_p):\\ a_i \in \mathrm{dom}(A_i)}} \;\prod_{A_k \in U_+(A_p)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]\Big)\\
&\quad \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_g):\\ a_i \in \mathrm{dom}(A_i)}} \;\prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]\\
=\;& \prod_{A_p \in \mathrm{parents}(A_g)} \pi(A_p = a_p)\\
&\quad \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_g):\\ a_i \in \mathrm{dom}(A_i)}} \;\prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]
\end{aligned}
$$

Evidence Propagation in Polytrees (continued)

$$\sum_{\substack{\forall A_i \in U_+^*(A_g):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \pi(A_p = a_p)$$

$$\cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_g):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]$$

$$= \prod_{A_p \in \mathrm{parents}(A_g)} \pi_{A_p \to A_g}(A_p = a_p)$$

$$\pi(A_g = a_g) = \sum_{\substack{\forall A_i \in \mathrm{parents}(A_g):\\ a_i \in \mathrm{dom}(A_i)}} P(A_g = a_g \mid \bigwedge_{A_j \in \mathrm{parents}(A_g)} A_j = a_j) \cdot \prod_{A_p \in \mathrm{parents}(A_g)} \pi_{A_p \to A_g}(A_p = a_p)$$

Evidence Propagation in Polytrees (continued)

$$\begin{aligned}
\lambda(A_g = a_g) &= \sum_{\substack{\forall A_i \in U_-(A_g):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big) \\
&= \prod_{A_c \in \mathrm{children}(A_g)} \sum_{a_c \in \mathrm{dom}(A_c)} \\
&\qquad \underbrace{\begin{aligned}
&\Big(\sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_g\}:\\ a_i \in \mathrm{dom}(A_i)}} P\Big(A_c = a_c \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_c)} A_j = a_j\Big) \\
&\cdot \Big[\sum_{\substack{\forall A_i \in U_+^*(A_c, A_g):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_+(A_c, A_g)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]\Big) \\
&\cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_c):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_c)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]
\end{aligned}}_{= \lambda(A_c = a_c)} \\
&= \prod_{A_c \in \mathrm{children}(A_g)} \lambda_{A_c \to A_g}(A_g = a_g)
\end{aligned}$$

Propagation Formulae without Evidence

$$
\begin{aligned}
&\pi_{A_p \to A_c}(A_p = a_p)\\
&\quad = \quad \pi(A_p = a_p) \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p,A_c):\\ a_i \in \mathrm{dom}(A_i)}} \prod_{A_k \in U_-(A_p,A_c)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_k)} A_j = a_j\Big)\Big]\\
&\quad = \quad \frac{P(A_p = a_p)}{\lambda_{A_c \to A_p}(A_p = a_p)}
\end{aligned}
$$

$$
\begin{aligned}
&\lambda_{A_c \to A_p}(A_p = a_p)\\
&\quad = \sum_{a_c \in \mathrm{dom}(A_c)} \lambda(A_c = a_c) \sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_p\}:\\ a_i \in \mathrm{dom}(A_k)}} P\Big(A_c = a_c \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_c)} A_j = a_j\Big)\\
&\qquad\qquad \cdot \prod_{A_k \in \mathrm{parents}(A_c) - \{A_p\}} \pi_{A_k \to A_p}(A_k = a_k)
\end{aligned}
$$

Evidence Propagation in Polytrees (continued)

Evidence: The attributes in a set X_{obs} are observed.

$$P\Big(A_g = a_g \,\Big|\, \bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}\Big)$$

$$= \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \text{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j = a_j \,\Big|\, \bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}\Big)$$

$$= \alpha \sum_{\substack{\forall A_i \in U - \{A_g\}: \\ a_i \in \text{dom}(A_i)}} P\Big(\bigwedge_{A_j \in U} A_j = a_j\Big) \prod_{A_k \in X_{\text{obs}}} P\left(A_k = a_k \,\middle|\, A_k = a_k^{(\text{obs})}\right),$$

where $\quad \alpha = \dfrac{1}{P\left(\bigwedge_{A_k \in X_{\text{obs}}} A_k = a_k^{(\text{obs})}\right)}$

Propagation Formulae with Evidence

$$
\begin{aligned}
&\pi_{A_p \to A_c}(A_p = a_p) \\
&\quad = \quad P\left(A_p = a_p \,\middle|\, A_p = a_p^{(\text{obs})}\right) \cdot \pi(A_p = a_p) \\
&\qquad \cdot \Big[\sum_{\substack{\forall A_i \in U_-(A_p, A_c): \\ a_i \in \text{dom}(A_i)}} \; \prod_{A_k \in U_-(A_p, A_c)} P\Big(A_k = a_k \,\Big|\, \bigwedge_{A_j \in \text{parents}(A_k)} A_j = a_j\Big)\Big] \\
&\quad = \quad \begin{cases} \beta, & \text{if } a_p = a_p^{(\text{obs})}, \\ 0, & \text{otherwise}, \end{cases}
\end{aligned}
$$

- The value of β is not explicitly determined. Usually a value of 1 is used and the correct value is implicitly determined later by normalizing the resulting probability distribution for A_g.

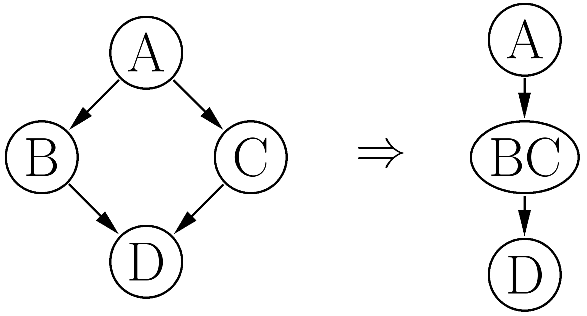
Propagation Formulae with Evidence

$$\begin{aligned}
&\lambda_{A_c \to A_p}(A_p = a_p) \\
&\quad = \sum_{a_c \in \mathrm{dom}(A_c)} P\left(A_c = a_c \,\middle|\, A_c = a_c^{(\mathrm{obs})}\right) \cdot \lambda(A_c = a_c) \\
&\qquad\qquad \cdot \sum_{\substack{\forall A_i \in \mathrm{parents}(A_c) - \{A_p\}: \\ a_i \in \mathrm{dom}(A_k)}} P\Big(A_c = a_c \,\Big|\, \bigwedge_{A_j \in \mathrm{parents}(A_c)} A_j = a_j\Big) \\
&\qquad\qquad\qquad\qquad \cdot \prod_{A_k \in \mathrm{parents}(A_c) - \{A_p\}} \pi_{A_k \to A_c}(A_k = a_k)
\end{aligned}$$

Propagation in Multiply Connected Networks

- Multiply connected networks pose a problem:
  - There are several ways on which information can travel from one attribute (node) to another.
  - As a consequence, the same evidence may be used twice to update the probability distribution of an attribute.
  - Since probabilistic update is not idempotent, multiple inclusion of the same evidence usually invalidates the result.
- General idea to solve this problem:

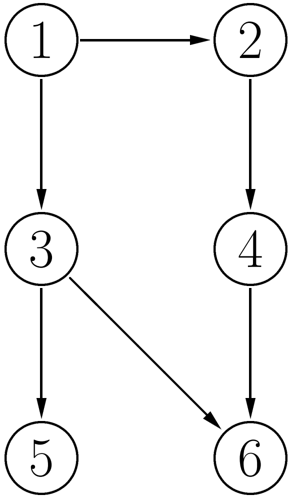
  Transform network into a singly connected structure.



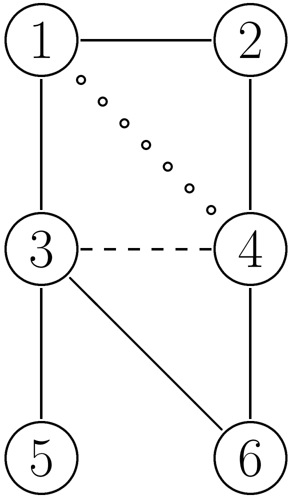
Merging attributes can make the polytree algorithm applicable in multiply connected networks.

Transformation into a Join Tree

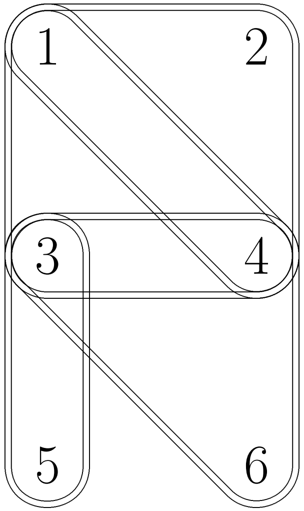
- **Goal:** Transform a graph into a singly connected structure



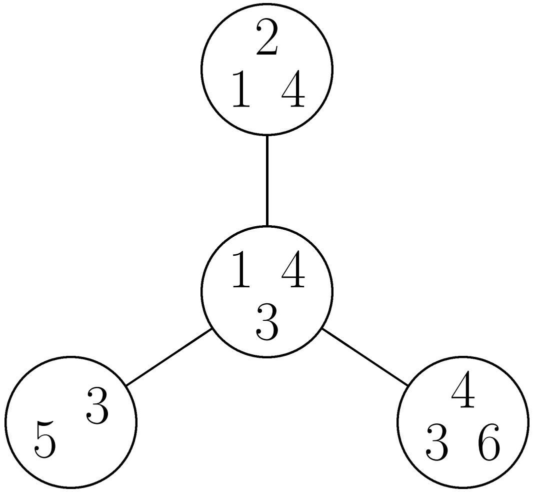
original graph



triangulated moral graph



cliques of the triangulated moral graph



join tree

Graph Triangulation

Algorithm: (graph triangulation)

Input: An undirected graph $G = (V, E)$.

Output: A triangulated undirected graph $G' = (V, E')$ with $E' \supseteq E$.

1. Compute an ordering of the nodes of the graph using *maximum cardinality search*, i.e., number the nodes from 1 to $n = |V|$, in increasing order, always assigning the next number to the node having the largest set of previously numbered neighbors (breaking ties arbitrarily).
2. From $i = n$ to $i = 1$ recursively fill in edges between any nonadjacent neighbors of the node numbered i having lower ranks than i (including neighbors linked to the node numbered i in previous steps). If no edges are added, then the original graph is chordal; otherwise the new graph is chordal.

Join Tree Construction

Algorithm: (join tree construction)

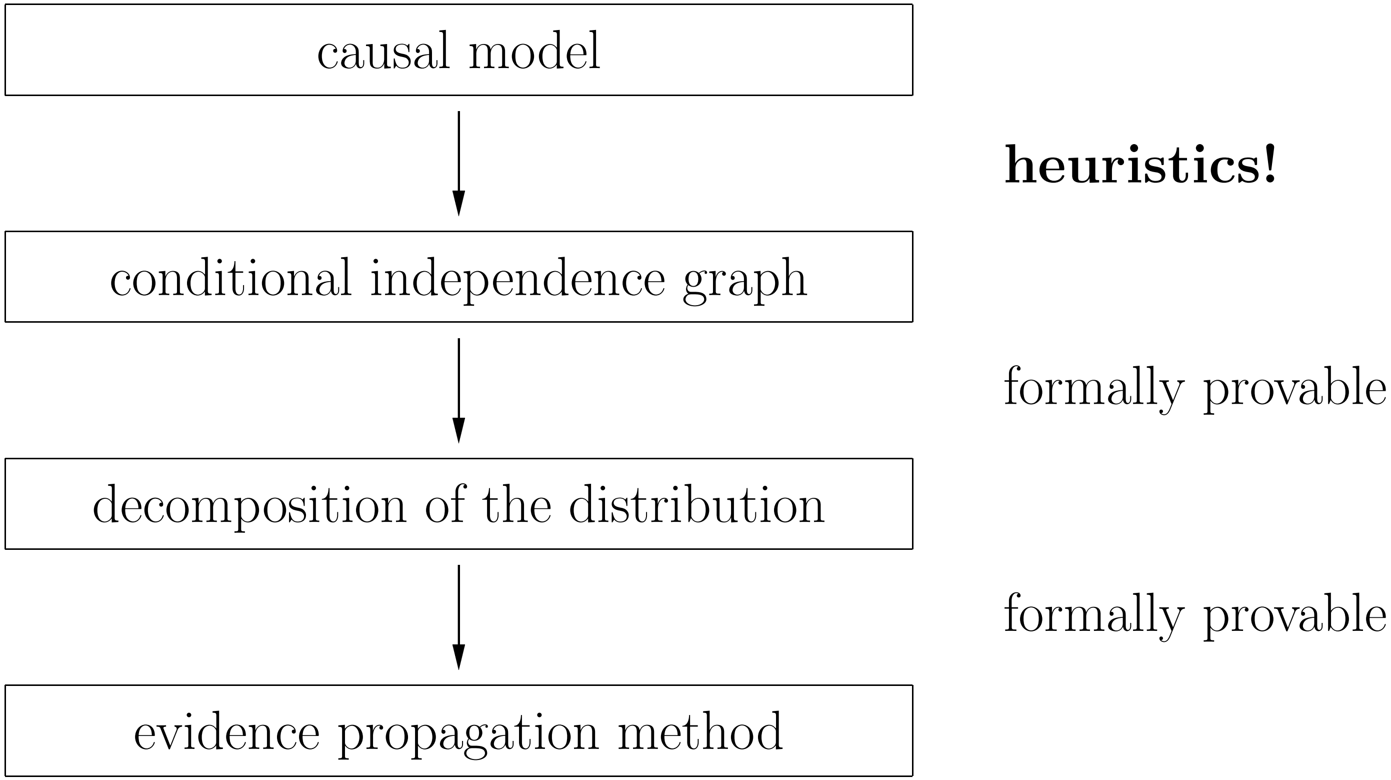
Input: A triangulated undirected graph $G = (V, E)$.

Output: A join tree $G' = (V', E')$ for G.

1. Determine a numbering of the nodes of G using maximum cardinality search.
2. Assign to each clique the maximum of the ranks of its nodes.
3. Sort the cliques in ascending order w.r.t. the numbers assigned to them.
4. Traverse the cliques in ascending order and for each clique C_i choose from the cliques $C_1, \ldots, C_{i-1}$ preceding it the clique with which it has the largest number of nodes in common (breaking ties arbitrarily).

Constructing a Graphical Model

Procedure based on human expert knowledge:



causal model

heuristics!

conditional independence graph

formally provable

decomposition of the distribution

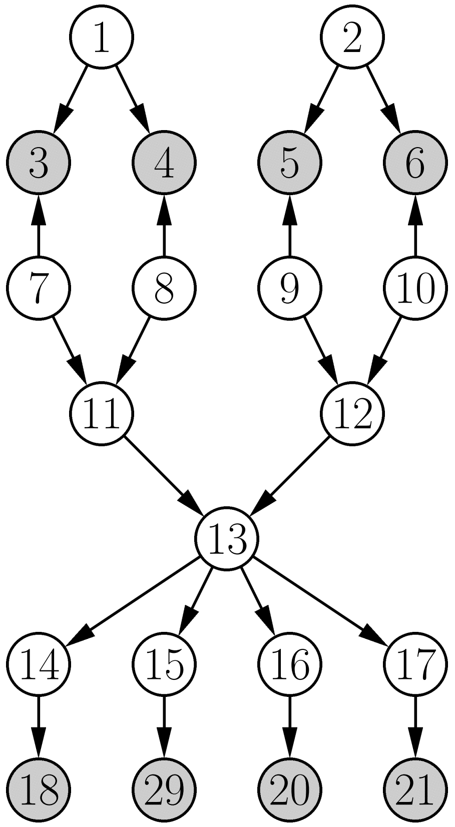
formally provable

evidence propagation method

- Problem: strong assumptions about the statistical effects of causal relations

Probabilistic Graphical Models: An Example

Danish Jersey Cattle Blood Type Determination



21 attributes:
1 – dam correct?
2 – sire correct?
3 – stated dam ph.gr. 1
4 – stated dam ph.gr. 2
5 – stated sire ph.gr. 1
6 – stated sire ph.gr. 2
7 – true dam ph.gr. 1
8 – true dam ph.gr. 2
9 – true sire ph.gr. 1
10 – true sire ph.gr. 2
11 – offspring ph.gr. 1
12 – offspring ph.gr. 2
13 – offspring genotype
14 – factor 40
15 – factor 41
16 – factor 42
17 – factor 43
18 – lysis 40
19 – lysis 41
20 – lysis 42
21 – lysis 43

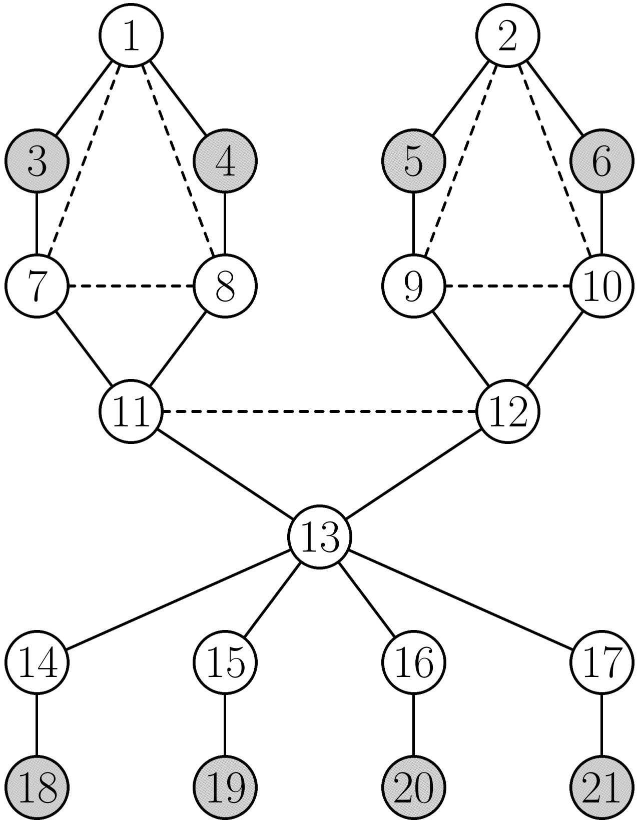
The grey nodes correspond to observable attributes.

Danish Jersey Cattle Blood Type Determination

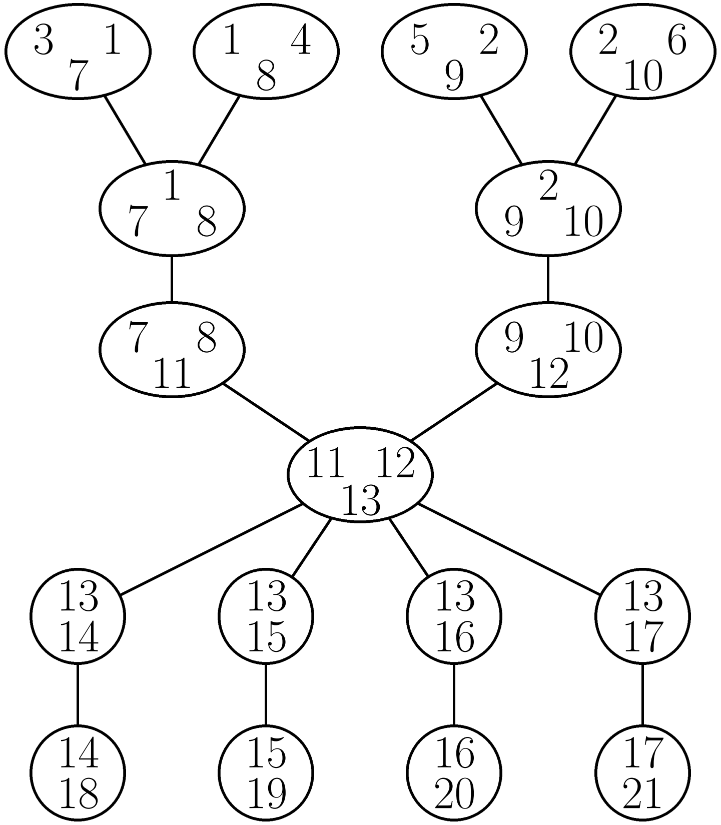
- Full 21-dimensional domain has $2^6 \cdot 3^{10} \cdot 6 \cdot 8^4 = 92\ 876\ 046\ 336$ possible states.
- Bayesian network requires only 306 conditional probabilities.
- Example of a conditional probability table (attributes 2, 5, and 9):

sire correct	true sire phenogroup 1	stated sire phenogroup 1		
		F1	V1	V2
yes	F1	1	0	0
yes	V1	0	1	0
yes	V2	0	0	1
no	F1	0.58	0.10	0.32
no	V1	0.58	0.10	0.32
no	V2	0.58	0.10	0.32

Danish Jersey Cattle Blood Type Determination



moral graph



join tree

Learning Graphical Models from Data

Given: A database of sample cases from a domain of interest.

Desired: A (good) graphical model of the domain of interest.

- **Quantitative or Parameter Learning**
  - The structure of the conditional independence graph is known.
  - Conditional or marginal distributions have to be estimated by standard statistical methods. (*parameter estimation*)
- **Qualitative or Structural Learning**
  - The structure of the conditional independence graph is not known.
  - A good graph has to be selected from the set of all possible graphs. (*model selection*)
  - Tradeoff between model complexity and model accuracy.

Inducing Naive Bayes Classifiers from Data

- Maximum likelihood model
  for each class and
  for each attribute

- Symbolic attributes:

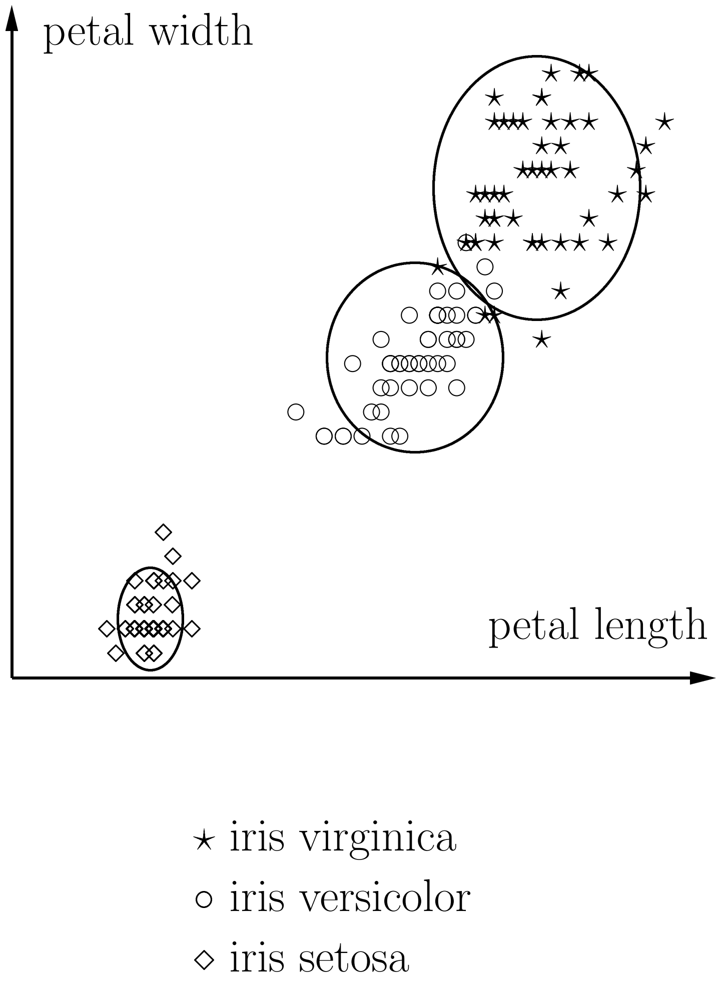
$$\hat{P}(A_j = a_j \mid C = c_i) = \frac{\#(A_j = a_j, C = c_i)}{\#(C = c_i)}$$

- Numeric attributes:

$$\hat{\mu}_j(c_i) = \frac{1}{\#(C = c_i)} \sum_{k=1}^{\#(C=c_i)} a_{j(k)}$$

$$\hat{\sigma}_j^2(c_i) = \frac{1}{\#(C = c_i)} \sum_{k=1}^{\#(C=c_i)} \big(a_j(k) - \hat{\mu}_j(c_i)\big)^2$$

(normal distribution assumption)



Learning the Structure of Graphical Models from Data

- **Test whether a distribution is decomposable w.r.t. a given graph.**

  This is the most direct approach. It is not bound to a graphical representation, but can also be carried out w.r.t. other representations of the set of subspaces to be used to compute the (candidate) decomposition of the given distribution.

- **Find an independence map by conditional independence tests.**

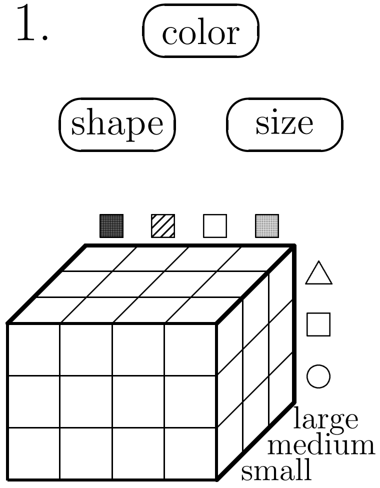
  This approach exploits the theorems that connect conditional independence graphs and graphs that represent decompositions. It has the advantage that a single conditional independence test, if it fails, can exclude several candidate graphs.

- **Find a suitable graph by measuring the strength of dependences.**

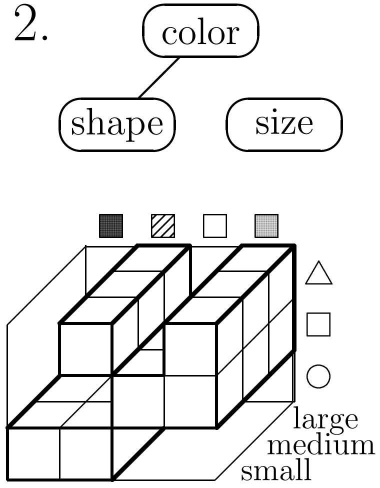
  This is a heuristic, but often highly successful approach, which is based on the frequently valid assumption that in a conditional independence graph an attribute is more strongly dependent on adjacent attributes than on attributes that are not directly connected to them.

Direct Test for Decomposability

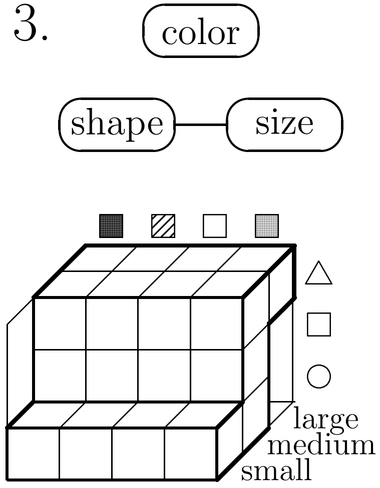
1.



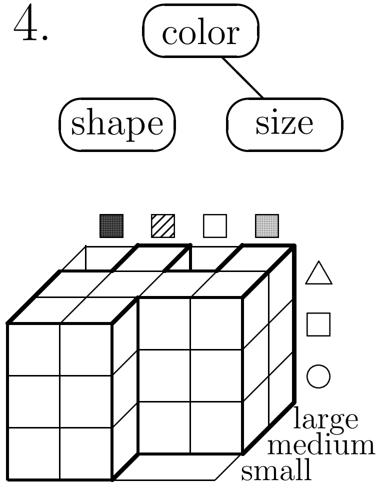
2.



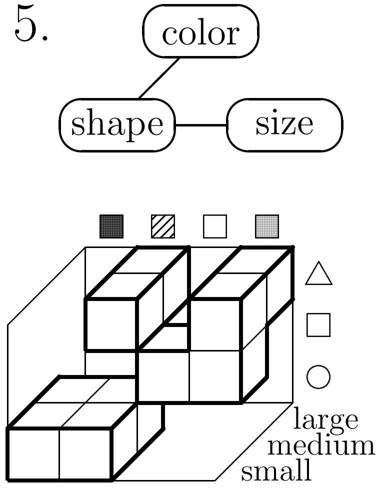
3.



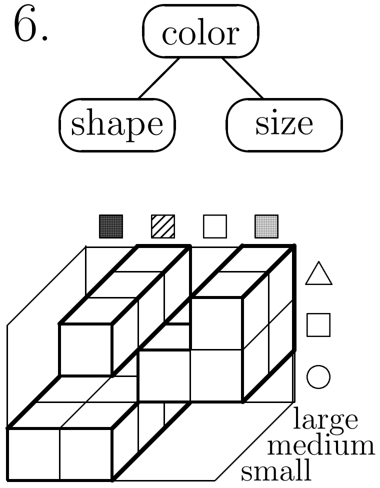
4.



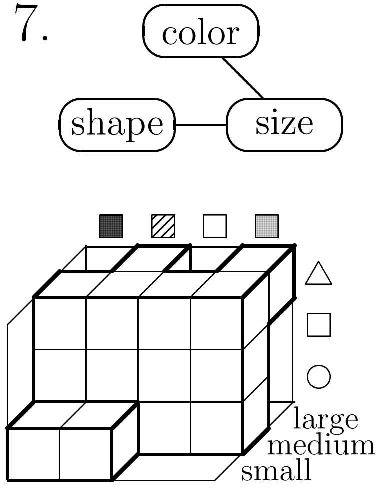
5.



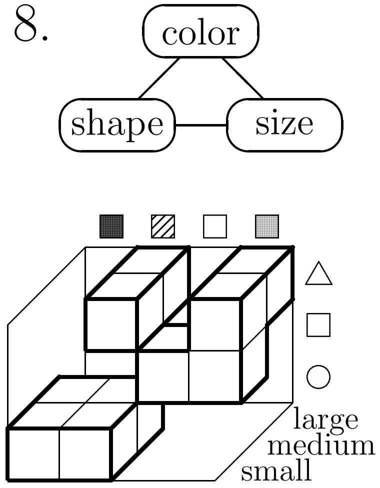
6.



7.



8.



Evaluation Measures and Search Methods

- An exhaustive search over all graphs is too expensive:
  - $2^{\binom{n}{2}}$ possible undirected graphs for n attributes.
  - $f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i)$ possible directed acyclic graphs.

- Therefore all learning algorithms consist of

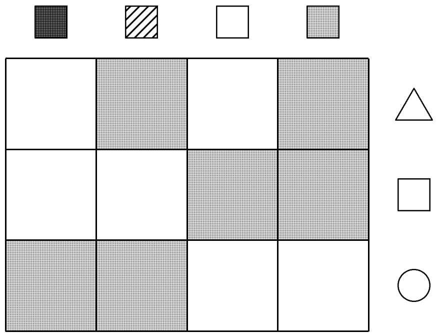
  an **evaluation measure** (scoring function), e.g.

  - Hartley information gain (relational networks)
  - Shannon information gain, K2 metric (probabilistic networks)

  and a (heuristic) **search method**, e.g.

  - guided random search (simulated annealing, genetic algorithms)
  - greedy search (K2 algorithm)
  - conditional independence search

Marginal Independence Tests



Hartley information needed to determine

coordinates:	$\log_2 4 + \log_2 3 = \log_2 12 \approx 3.58$
coordinate pair:	$\log_2 6 \approx 2.58$
gain:	$\log_2 12 - \log_2 6 = \log_2 2 = 1$

Definition: Let A and B be two attributes and R a discrete possibility measure with $\exists a \in \mathrm{dom}(A) : \exists b \in \mathrm{dom}(B) : R(A = a, B = b) = 1$. Then

$$\begin{aligned}
I_{\text{gain}}^{(\text{Hartley})}(A, B) &= \log_2 \Big(\sum_{a \in \mathrm{dom}(A)} R(A = a) \Big) + \log_2 \Big(\sum_{b \in \mathrm{dom}(B)} R(B = b) \Big) \\
&- \log_2 \Big(\sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} R(A = a, B = b) \Big) \\
&= \log_2 \frac{\Big(\sum_{a \in \mathrm{dom}(A)} R(A = a) \Big) \cdot \Big(\sum_{b \in \mathrm{dom}(B)} R(B = b) \Big)}{\sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} R(A = a, B = b)},
\end{aligned}$$

is called the **Hartley information gain** of A and B w.r.t. R.

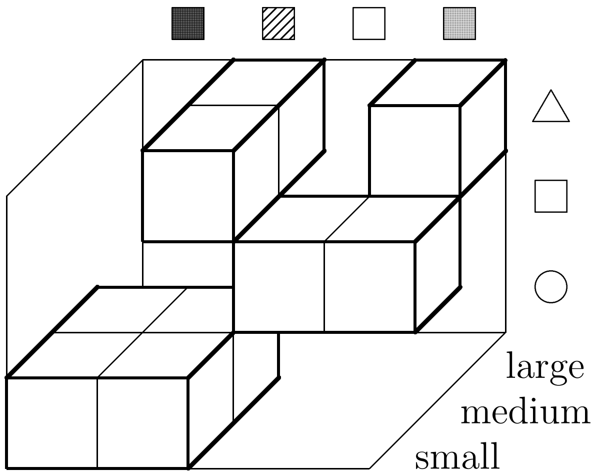
Conditional Independence Tests

- The Hartley information gain can be used directly to test for (approximate) **marginal independence**.

attributes	relative number of possible value combinations	Hartley information gain
color, shape	$\frac{6}{3\cdot 4} = \frac{1}{2} = 50\%$	$\log_2 3 + \log_2 4 - \log_2 6 = 1$
color, size	$\frac{8}{3\cdot 4} = \frac{2}{3} \approx 67\%$	$\log_2 3 + \log_2 4 - \log_2 8 \approx 0.58$
shape, size	$\frac{5}{3\cdot 3} = \frac{5}{9} \approx 56\%$	$\log_2 3 + \log_2 3 - \log_2 5 \approx 0.85$

- In order to test for (approximate) **conditional independence**:
  - Compute the Hartley information gain for each possible instantiation of the conditioning attributes.
  - Aggregate the result over all possible instantiations, for instance, by simply averaging them.

Conditional Independence Tests (continued)



A	Hartley information gain
a_1	$\log_2 1 + \log_2 2 - \log_2 2 = 0$
a_2	$\log_2 2 + \log_2 3 - \log_2 4 \approx 0.58$
a_3	$\log_2 1 + \log_2 1 - \log_2 1 = 0$
a_4	$\log_2 2 + \log_2 2 - \log_2 2 = 1$
	average: ≈ 0.40

B	Hartley information gain
b_1	$\log_2 2 + \log_2 2 - \log_2 4 = 0$
b_2	$\log_2 2 + \log_2 1 - \log_2 2 = 0$
b_3	$\log_2 2 + \log_2 2 - \log_2 3 \approx 0.42$
	average: ≈ 0.14

C	Hartley information gain
c_1	$\log_2 2 + \log_2 1 - \log_2 2 = 0$
c_2	$\log_2 4 + \log_2 3 - \log_2 5 \approx 1.26$
c_3	$\log_2 2 + \log_2 1 - \log_2 2 = 0$
	average: ≈ 0.42

Conditional Independence Tests (continued)

Algorithm: (conditional independence graph construction)

1. For each pair of attributes A and B, search for a set $S_{AB} \subseteq U \backslash \{A, B\}$ such that $A \perp\!\!\!\perp B \mid S_{AB}$ holds in $\widehat{P}$, i.e., A and B are independent in $\widehat{P}$ conditioned on S_{AB}. If there is no such S_{AB}, connect the attributes by an undirected edge.

2. For each pair of non-adjacent variables A and B with a common neighbour C (i.e., C is adjacent to A as well as to B), check whether $C \in S_{AB}$.

   - If it is, continue.
   - If it is not, add arrowheads pointing to C, i.e., $A \rightarrow C \leftarrow B$.

3. Recursively direct all undirected edges according to the rules:

   - If for two adjacent variables A and B there is a strictly directed path from A to B not including $A \rightarrow B$, then direct the edge towards B.
   - If there are three variables A, B, and C with A and B not adjacent, $B - C$, and $A \rightarrow C$, then direct the edge $C \rightarrow B$.

Measuring the Strengths of Marginal Dependences

- Learning a relational network consists in finding those subspace, for which the intersection of the cylindrical extensions of the projections to these subspaces approximates best the set of possible world states, i.e. contains as few additional states as possible.

- Since computing explicitly the intersection of the cylindrical extensions of the projections and comparing it to the original relation is too expensive, local evaluation functions are used, for instance:

subspace	color × shape	shape × size	size × color
possible combinations	12	9	12
occurring combinations	6	5	8
relative number	50%	56%	67%

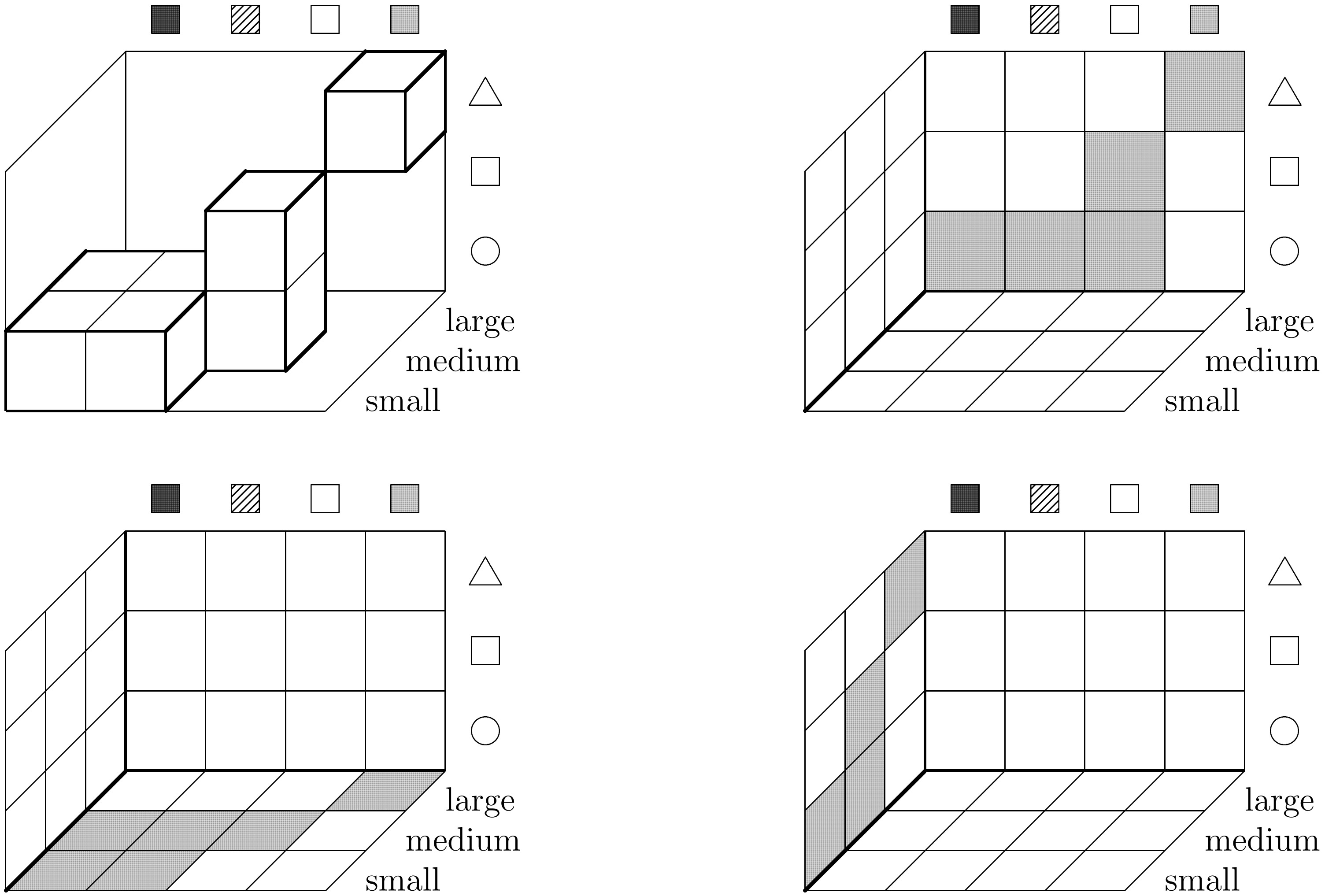
- The relational network can be obtained by interpreting the relative numbers as edge weights and constructing the minimal weight spanning tree.

Measuring the Strengths of Marginal Dependences

- **Optimum Weight Spanning Tree Construction**
  - Compute an evaluation measure on all possible edges (two-dimensional subspaces).
  - Use the Kruskal algorithm to determine an optimum weight spanning tree.

- **Greedy Parent Selection** (for directed graphs)
  - Define a topological order of the attributes (to restrict the search space).
  - Compute an evaluation measure on all single attribute hyperedges.
  - For each preceding attribute (w.r.t. the topological order):
    add it as a candidate parent to the hyperedge and
    compute the evaluation measure again.
  - Greedily select a parent according to the evaluation measure.
  - Repeat the previous two steps until no improvement results from them.

Measuring the Strengths of Marginal Dependences



Direct Test for Decomposability

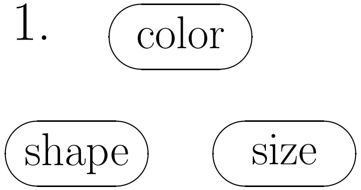
Definition: Let p_1 and p_2 be two strictly positive probability distributions on the same set $\mathcal{E}$ of events. Then

$$I_{\mathrm{KLdiv}}(p_1, p_2) = \sum_{E \in \mathcal{E}} p_1(E) \log_2 \frac{p_1(E)}{p_2(E)}$$

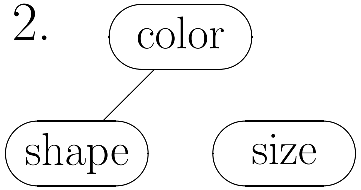
is called the **Kullback-Leibler information divergence** of p_1 and p_2.

- The Kullback-Leibler information divergence is non-negative.
- It is zero if and only if $p_1 \equiv p_2$.
- Therefore it is plausible that this measure can be used to assess the quality of the approximation of a given multi-dimensional distribution p_1 by the distribution p_2 that is represented by a given graph:
  The smaller the value of this measure, the better the approximation.

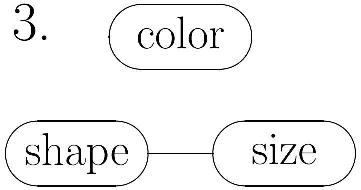
Direct Test for Decomposability (continued)

1.


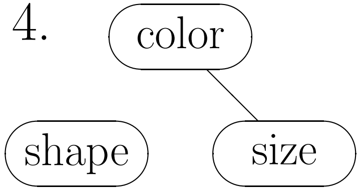
0.640
−5041

2.


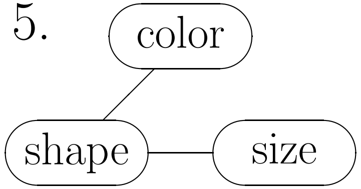
0.211
−4612

3.


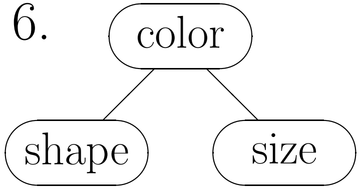
0.429
−4830

4.


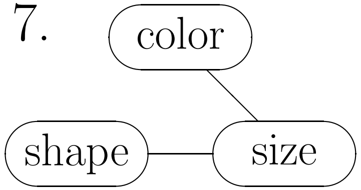
0.590
−4991

5.


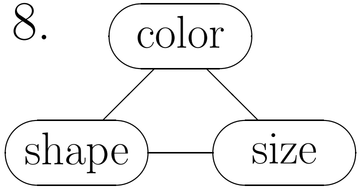
0
−4401

6.


0.161
−4563

7.


0.379
−4780

8.


0
−4401

Upper numbers: The Kullback-Leibler information divergence of the original distribution and its approximation.

Lower numbers: The binary logarithms of the probability of an example database (log-likelihood of data).

Evaluation Measures / Scoring Functions

Relational Networks

- Hartley Information Gain
- Conditional Hartley Information Gain

Probabilistic Networks

- χ^2-Measure
- Mutual Information / Cross Entropy / Information Gain
- (Symmetric) Information Gain Ratio
- (Symmetric/Modified) Gini Index
- Bayesian Measures (K2 metric, BDeu metric)
- Measures based on the Minimum Description Length Principle
- Other measures that are known from Decision Tree Induction

A Probabilistic Evaluation Measure

Mutual Information / Cross Entropy / Information Gain

Based on Shannon Entropy $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ (Shannon 1948)

$$\begin{aligned} I_{\text{gain}}(A,B) &= \qquad H(A) \qquad - \qquad H(A \mid B) \\ &= \overbrace{-\sum_{i=1}^{n_A} p_{i.} \log_2 p_{i.}} - \overbrace{\sum_{j=1}^{n_B} p_{.j} \left(-\sum_{i=1}^{n_A} p_{i|j} \log_2 p_{i|j} \right)} \end{aligned}$$

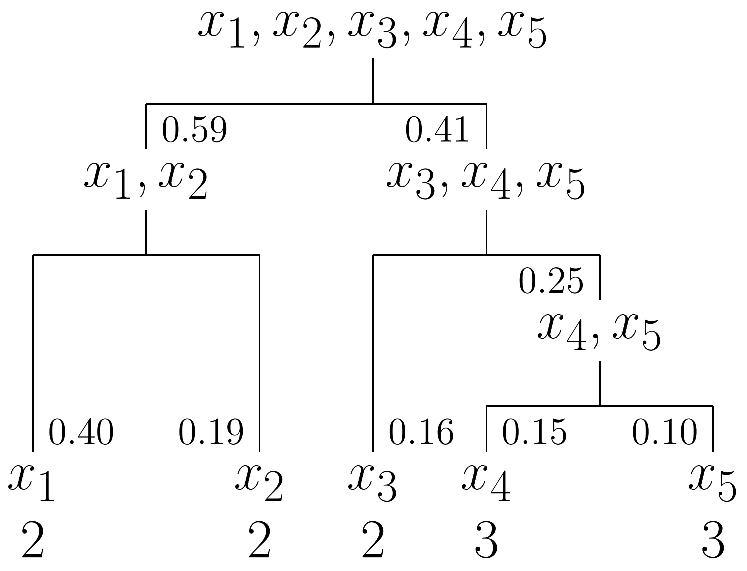
$H(A)$	Entropy of the distribution on attribute A
$H(A\|B)$	*Expected entropy* of the distribution on attribute A if the value of attribute B becomes known
$H(A) - H(A\|B)$	Expected reduction in entropy or *information gain*

Question/Coding Schemes

$$P(x_1) = 0.40, \quad P(x_2) = 0.19, \quad P(x_3) = 0.16, \quad P(x_4) = 0.15, \quad P(x_5) = 0.10$$

Shannon Entropy: 2.15 bit/symbol

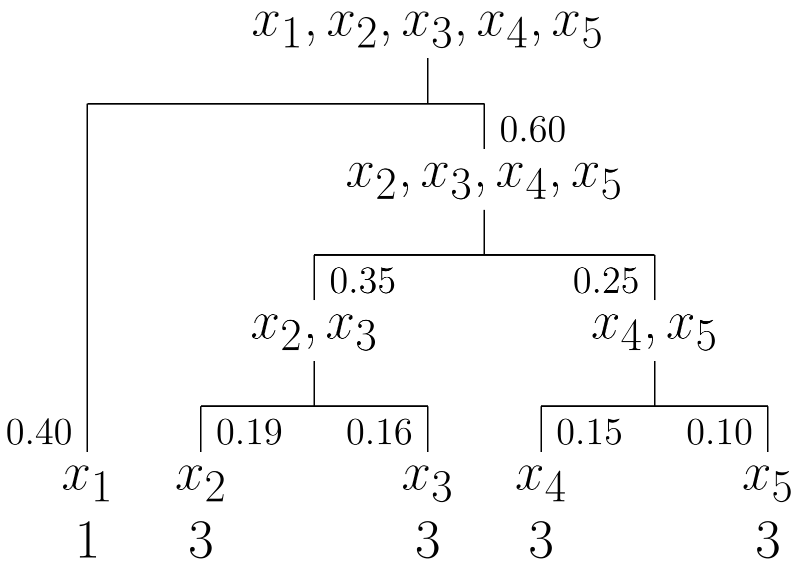
Shannon-Fano Coding (1948)



Average Code Length: 2.25 bit/symbol

Code Efficiency: 0.955

Huffman Coding (1952)



Average Code Length: 2.20 bit/symbol

Code Efficiency: 0.977

A Probabilistic Evaluation Measure

Mutual Information / Cross Entropy / Information Gain

- Mutual information is symmetric:

$$\begin{aligned}
I_{\text{gain}}(A,B) &= H_A - H_{A|B} \;=\; H_A + H_B - H_{AB} \\
&= -\sum_{a\in\text{dom}(A)} P(A=a)\log_2 P(A=a) \\
&\quad -\sum_{b\in\text{dom}(B)} P(B=b)\log_2 P(B=b) \\
&\quad +\sum_{a\in\text{dom}(A)}\sum_{b\in\text{dom}(B)} P(A=a,B=b)\log_2 P(A=a,B=b) \\
&= H_B - H_{B|A} \\
&= I_{\text{gain}}(B,A)
\end{aligned}$$

- Consequently, it can also be used for undirected graphs.

Mutual Information for the Example

projection to subspace

	■	▨	□	▒
△	40	180	20	160
□	12	6	120	102
○	168	144	30	18

product of marginals

	■	▨	□	▒
△	88	132	68	112
□	53	79	41	67
○	79	119	61	101

mutual information

0.429 bit

projection to subspace

	s	m	l
△	20	180	200
□	40	160	40
○	180	120	60

product of marginals

	s	m	l
△	96	184	120
□	58	110	72
○	86	166	108

mutual information

0.211 bit

projection to subspace

	■	▨	□	▒
large	50	115	35	100
medium	82	133	99	146
small	88	82	36	34

product of marginals

	■	▨	□	▒
large	66	99	51	84
medium	101	152	78	129
small	53	79	41	67

mutual information

0.050 bit

Conditional Independence Tests

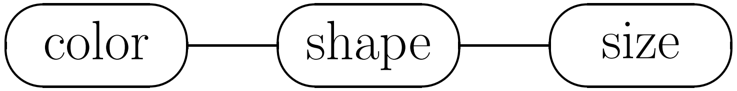
- There are no marginal independences, although the dependence of color and size is rather weak.

- Conditional independence tests may be carried out by summing the mutual information for all instantiations of the conditioning variables:

  $$I_{\mathrm{mut}}(A, B \mid C) = \sum_{c \in \mathrm{dom}(C)} P(c) \sum_{a \in \mathrm{dom}(A)} \sum_{b \in \mathrm{dom}(B)} P(a, b \mid c) \, \log_2 \frac{P(a, b \mid c)}{P(a \mid c) \, P(b \mid c)},$$

  where $P(c)$ is an abbreviation of $P(C = c)$ etc.

- Since $I_{\mathrm{mut}}(\mathrm{color}, \mathrm{size} \mid \mathrm{shape}) = 0$ indicates the only conditional independence, we get the following learning result:



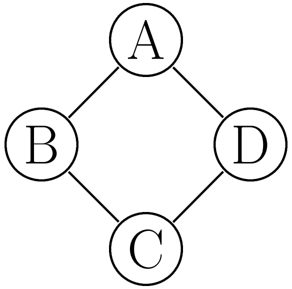
color

shape

size

Conditional Independence Tests (continued)

- The conditional independence graph construction algorithm presupposes that there is a **perfect map**. If there is no perfect map, the result may be invalid.



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

- **Independence tests of high order**, i.e., with a large number of conditions, may be necessary.

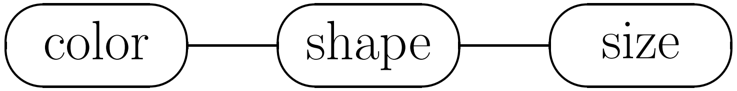
- There are approaches to mitigate these drawbacks.
  (E.g., the order is restricted and all tests of higher order are assumed to fail if all tests of lower order failed.)

Measuring the Strengths of Marginal Dependences

- Results for the simple example:

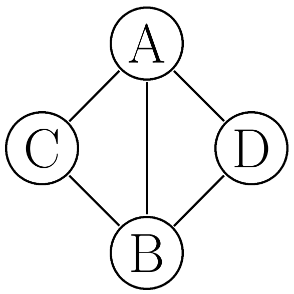
$$\begin{aligned} I_{\text{mut}}(\text{color}, \text{shape}) &= 0.429 \text{ bit} \\ I_{\text{mut}}(\text{shape}, \text{size}) &= 0.211 \text{ bit} \\ I_{\text{mut}}(\text{color}, \text{size}) &= 0.050 \text{ bit} \end{aligned}$$

- Applying the Kruskal algorithm yields as a learning result:



- It can be shown that this approach always yields the best possible spanning tree w.r.t. Kullback-Leibler information divergence (Chow and Liu 1968).
- For more complex graphs, the best graph need not be found (there are counterexamples, see next slide).

Measuring the Strengths of Marginal Dependences



p_{ABCD}		$A = a_1$		$A = a_2$	
		$B = b_1$	$B = b_2$	$B = b_1$	$B = b_2$
$C = c_1$	$D = d_1$	${}^{48}/_{250}$	${}^{2}/_{250}$	${}^{2}/_{250}$	${}^{27}/_{250}$
	$D = d_2$	${}^{12}/_{250}$	${}^{8}/_{250}$	${}^{8}/_{250}$	${}^{18}/_{250}$
$C = c_2$	$D = d_1$	${}^{12}/_{250}$	${}^{8}/_{250}$	${}^{8}/_{250}$	${}^{18}/_{250}$
	$D = d_2$	${}^{3}/_{250}$	${}^{32}/_{250}$	${}^{32}/_{250}$	${}^{12}/_{250}$

p_{AB}	a_1	a_2
b_1	0.3	0.2
b_2	0.2	0.3

p_{AC}	a_1	a_2
c_1	0.28	0.22
c_2	0.22	0.28

p_{AD}	a_1	a_2
d_1	0.28	0.22
d_2	0.22	0.28

p_{CD}	c_1	c_2
d_1	0.316	0.184
d_2	0.184	0.316

p_{BC}	b_1	b_2
c_1	0.28	0.22
c_2	0.22	0.28

p_{BD}	b_1	b_2
d_1	0.28	0.22
d_2	0.22	0.28

- Attributes C and D have the highest mutual information

Another Probabilistic Evaluation Measure: K2 Metric

- Idea: Compute the probability of a graph given the data (Bayesian approach)

$$P(\vec{G} \mid D) = \frac{1}{P(D)} \int_{\Theta} P(D \mid \vec{G}, \Theta) f(\Theta \mid \vec{G}) P(\vec{G}) \,\mathrm{d}\Theta$$

- Assumptions about data and parameter independence yield:

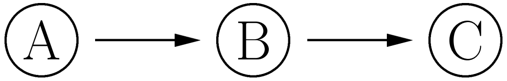
$$P(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \int_{\theta_{ijk}} \!\!\cdots\! \int \left(\prod_{i=1}^{n_k} \theta_{ijk}^{N_{ijk}} \right) f(\theta_{1jk}, \ldots, \theta_{n_k jk}) \,\mathrm{d}\theta_{1jk} \ldots \mathrm{d}\theta_{n_k jk}$$

- Choose $f(\theta_{1jk}, \ldots, \theta_{n_k jk}) = \text{const.}$ [Cooper and Herskovits 1992]

- Then the solution can be obtained via Dirichlet's integral:

$$K_2(\vec{G}, D) = \gamma \prod_{k=1}^{r} \prod_{j=1}^{m_k} \frac{(n_k - 1)!}{(N_{.jk} + n_k - 1)!} \prod_{i=1}^{n_k} N_{ijk}!$$

Simple Causal Structures and Alleged (In)Dependences

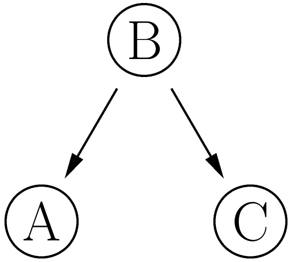


causal chain

Example:

A – accelerator pedal
B – fuel supply
C – engine speed

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

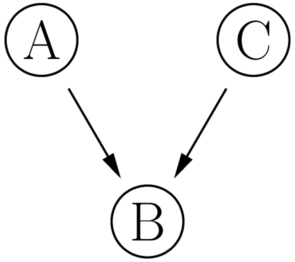


common cause

Example:

A – ice cream sales
B – temperature
C – bathing accidents

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



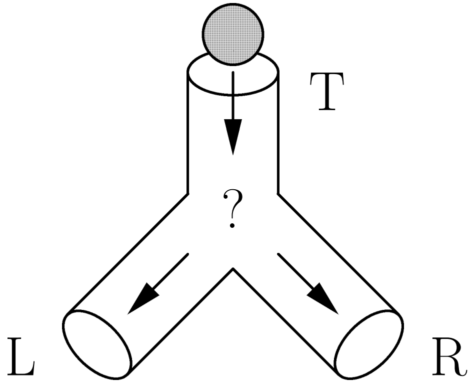
common effect

Example:

A – influenza
B – fever
C – measles

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

Common Cause Assumption (Causal Markov Assumption)



t	r	$\overline{r}$	Σ
l	0	$^1/_2$	$^1/_2$
$\bar{l}$	$^1/_2$	0	$^1/_2$
Σ	$^1/_2$	$^1/_2$	

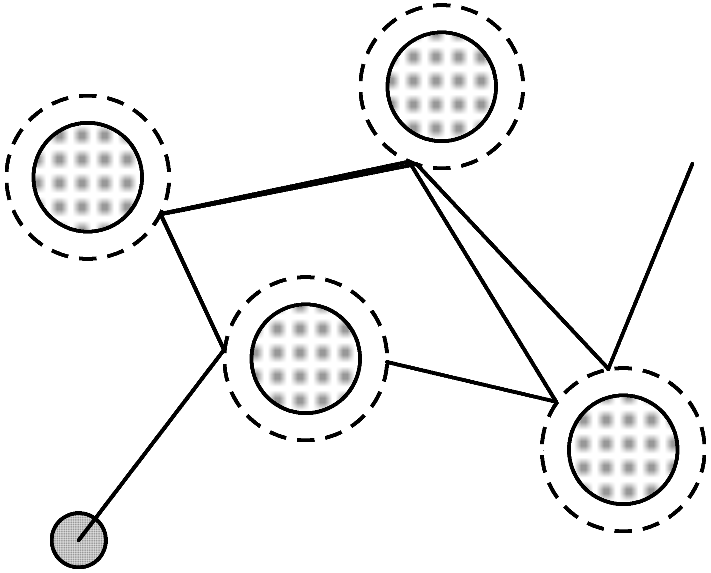
Y-shaped tube arrangement into which a ball is dropped (T). Since the ball can reappear *either* at the left outlet (L) *or* the right outlet (R) the corresponding variables are dependent.

Counter argument: The cause is insufficiently described. If the exact shape, position and velocity of the ball and the tubes are known, the outlet can be determined and the variables become independent.

Counter counter argument: Quantum mechanics states that location and momentum of a particle cannot both at the same time be measured with arbitrary precision.

Sensitive Dependence on the Initial Conditions

- *Sensitive dependence on the initial conditions* means that a small change of the initial conditions (e.g. a change of the initial position or velocity of a particle) causes a deviation that grows *exponentially* with time.
- Many physical systems show, for arbitrary initial conditions, a sensitive dependence on the initial conditions.



Example: Billiard with round (or generally convex) obstacles.

Initial imprecision: $\approx \frac{1}{100}$ degree

after four collisions: ≈ 100 degrees

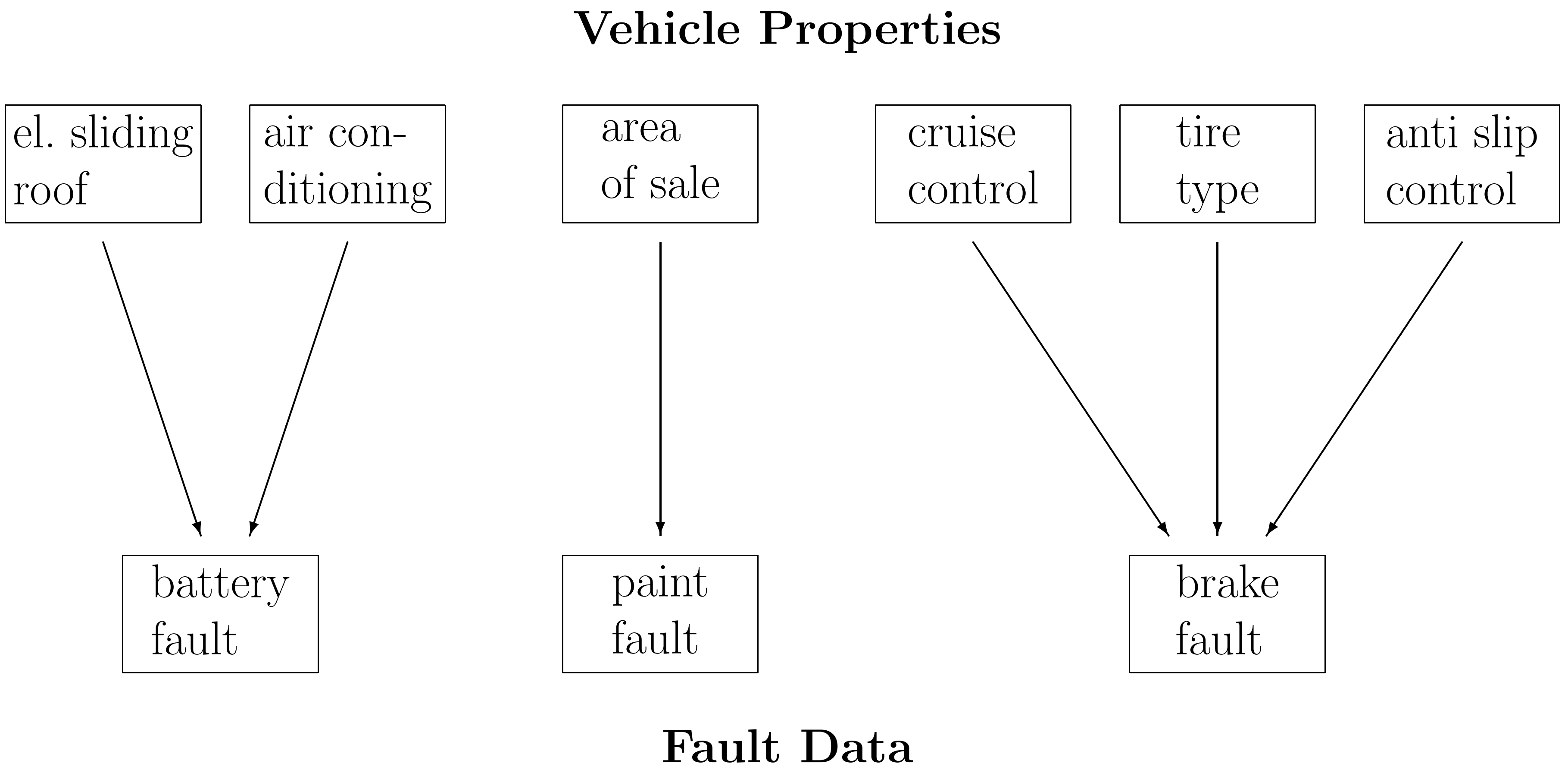
Fields of Application (DaimlerChrysler AG)

- **Improvement of Product Quality by Finding Weaknesses**
  - Learn decision trees or inference network
    for vehicle properties and faults.
  - Look for unusual conditional fault frequencies.
  - Find causes for these unusual frequencies.
  - Improve construction of vehicle.

- **Improvement of Error Diagnosis in Garages**
  - Learn decision trees or inference network
    for vehicle properties and faults.
  - Record properties of new faulty vehicle.
  - Test for the most probable faults.

A Simple Approach to Fault Analysis

- Check subnets consisting of an attribute and its parent attributes.
- Select subnets with highest deviation from an independent distribution.



Example Subnet

Influence of special equipment on battery faults:

(fictitious) frequency of battery faults		air conditioning with	air conditioning without
electrical sliding roof	with	8 %	3 %
	without	3 %	2 %

- Significant deviation from independent distribution.
- Hints to possible causes and improvements.
- Here: Larger battery may be required if an air conditioning system. *and* an electrical sliding roof are built in.

(The dependences and frequencies of this example are fictitious, true numbers are confidential.)

Summary

- **Decomposition:** Under certain conditions a distribution δ (e.g. a probability distribution) on a multi-dimensional domain, which encodes *prior* or *generic knowledge* about this domain, can be decomposed into a set $\{\delta_1, \ldots, \delta_s\}$ of (overlapping) distributions on lower-dimensional subspaces.

- **Simplified Reasoning:** If such a decomposition is possible, it is sufficient to know the distributions on the subspaces to draw all inferences in the domain under consideration that can be drawn using the original distribution δ.

- **Graphical Model:** The decomposition is represented by a graph (in the sense of graph theory). The edges of the graph indicate the paths along which evidence has to be propagated. Efficient and correct evidence propagation algorithms can be derived, which exploit the graph structure.

- **Learning from Data:** There are several highly successful approaches to learn graphical models from data, although all of them are based on heuristics.