Possibilistic Graphical Models and How to Learn Them from Data

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

Contents

- **Possibility Theory**
  - Axiomatic Approach
  - Semantical Considerations
- **Graphical Models / Inference Networks**
  - relational
  - probabilistic
  - possibilistic
- **Learning Possibilistic Graphical Models from Data**
  - Computing Maximum Projections
  - Naive Possibilistic Classifiers
  - Learning the Structure of Graphical Models
- **Summary**

Possibility Theory: Axiomatic Approach

Definition: Let Ω be a (finite) sample space.
A **possibility measure** Π on Ω is a function $\Pi : 2^\Omega \to [0,1]$ satisfying

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

- Similar to Kolmogorov's axioms of probability theory.
- From the axioms follows $\Pi(E_1 \cap E_2) \leq \min\{\Pi(E_1), \Pi(E_2)\}$.
- Attributes are introduced as random variables (as in probability theory).
- $\Pi(A = a)$ is an abbreviation of $\Pi(\{\omega \in \Omega \mid A(\omega) = a\})$
- If an event E is possible without restriction, then $\Pi(E) = 1$.
  If an event E is impossible, then $\Pi(E) = 0$.

Possibility Theory and the Context Model

Interpretation of Degrees of Possibility [Gebhardt and Kruse 1993]

- Let Ω be the (nonempty) set of all possible states of the world, ω_0 the actual (but unknown) state.

- Let $C = \{c_1, \ldots, c_n\}$ be a set of contexts (observers, frame conditions etc.) and $(C, 2^C, P)$ a finite probability space (context weights).

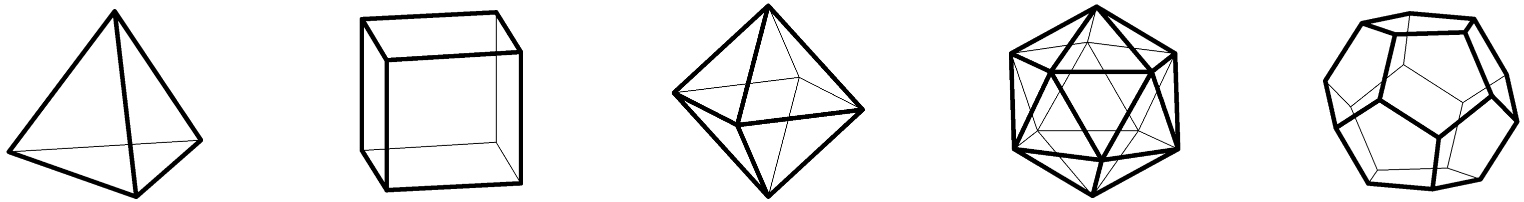
- Let $\Gamma : C \to 2^\Omega$ be a set-valued mapping, which assigns to each context the **most specific correct set-valued specification of** ω_0. The sets $\Gamma(c)$ are called the **focal sets** of Γ.

- Γ is a **random set** (i.e., a set-valued random variable) [Nguyen 1978]. The **basic possibility assignment** induced by Γ is the mapping

$$\begin{aligned} \pi : \Omega &\to [0, 1] \\ \pi(\omega) &\mapsto P(\{c \in C \mid \omega \in \Gamma(c)\}). \end{aligned}$$

Example: Dice and Shakers

shaker 1	shaker 2	shaker 3	shaker 4	shaker 5



tetrahedron	hexahedron	octahedron	icosahedron	dodecahedron
1 – 4	1 – 6	1 – 8	1 – 10	1 – 12

numbers	degree of possibility
1 – 4	$\frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = 1$
5 – 6	$\frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{4}{5}$
7 – 8	$\frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5}$
9 – 10	$\frac{1}{5} + \frac{1}{5} = \frac{2}{5}$
11 – 12	$\frac{1}{5} = \frac{1}{5}$

From the Context Model to Possibility Measures

Definition: Let $\Gamma : C \to 2^\Omega$ be a random set.
The **possibility measure** induced by Γ is the mapping

$$\begin{aligned} \Pi : 2^\Omega &\to [0,1], \\ E &\mapsto P(\{c \in C \mid E \cap \Gamma(c) \neq \emptyset\}). \end{aligned}$$

Problem: From the given interpretation it follows only:

$$\forall E \subseteq \Omega : \quad \max_{\omega \in E} \pi(\omega) \;\leq\; \Pi(E) \;\leq\; \min\Big\{1, \sum_{\omega \in E} \pi(\omega)\Big\}.$$

	1	2	3	4	5
$c_1 : \frac{1}{2}$			●		
$c_2 : \frac{1}{4}$		●	●	●	
$c_3 : \frac{1}{4}$	●	●	●	●	●
π	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{4}$

	1	2	3	4	5
$c_1 : \frac{1}{2}$			●		
$c_2 : \frac{1}{4}$	●	●			
$c_3 : \frac{1}{4}$				●	●
π	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

From the Context Model to Possibility Measures (cont.)

Attempts to solve the indicated problem:

- Require the focal sets to be **consonant**:
  Definition: Let $\Gamma : C \to 2^\Omega$ be a random set with $C = \{c_1, \ldots, c_n\}$. The focal sets $\Gamma(c_i)$, $1 \leq i \leq n$, are called **consonant**, iff there exists a sequence $c_{i_1}, c_{i_2}, \ldots, c_{i_n}$, $1 \leq i_1, \ldots, i_n \leq n$, $\forall 1 \leq j < k \leq n : i_j \neq i_k$, so that
  $$\Gamma(c_{i_1}) \subseteq \Gamma(c_{i_2}) \subseteq \ldots \subseteq \Gamma(c_{i_n}).$$
  $\rightarrow$ mass assignment theory [Baldwin *et al.* 1995]

  Problem: The "voting model" is not sufficient to justify consonance.

- Use the lower bound as the "most pessimistic" choice. [Gebhardt 1997]

  Problem: Basic possibility assignments represent negative information, the lower bound is actually the *most optimistic* choice.

- Justify the lower bound from decision making purposes.
  [Borgelt 1995, Borgelt 2000]

From the Context Model to Possibility Measures (cont.)

- Assume that in the end we have to decide on a single event.
- Each event is described by the values of a set of attributes.
- Then it can be useful to assign to a set of events the degree of possibility of the "most possible" event in the set.

Example:

$\sum$					
	36	18	18	28	
28	0	0	0	28	28
18	18	0	0	0	18
18	18	0	0	0	18
36	0	18	18	0	18
	18	18	18	28	max

0	40	0	40
40	0	0	40
0	0	20	20
40	40	20	max

Possibility Distributions

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

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

- 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.

Conditional Possibility and Independence

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

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

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

Definition: Let Ω be a (finite) sample space, Π a 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 possibilistically independent** given C, written $A \perp\!\!\!\perp_\Pi B \mid C$, iff

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

- Similar to the corresponding notions of probability theory.

Graphical Models / Inference Networks

- **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.

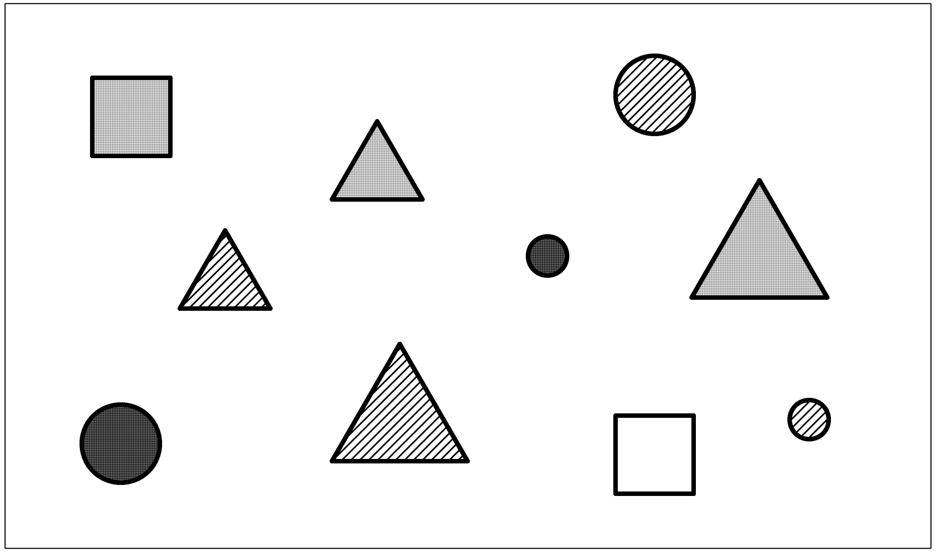
- **Simplified Reasoning:** If such a decomposition is possible, it is sufficient to know the distributions on the subspaces to draw all inferences in the domain under consideration that can be drawn using the original distribution δ.

- 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

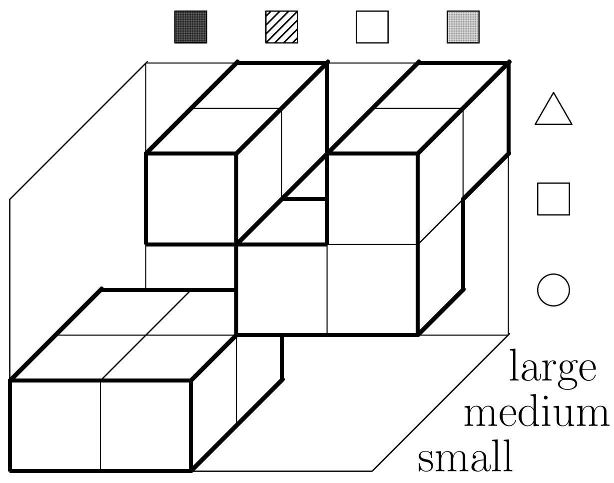
- 10 simple geometric objects, 3 attributes
- One object is chosen at random and examined.
- Inferences are drawn about the unobserved attributes.

The Reasoning Space

Relation

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

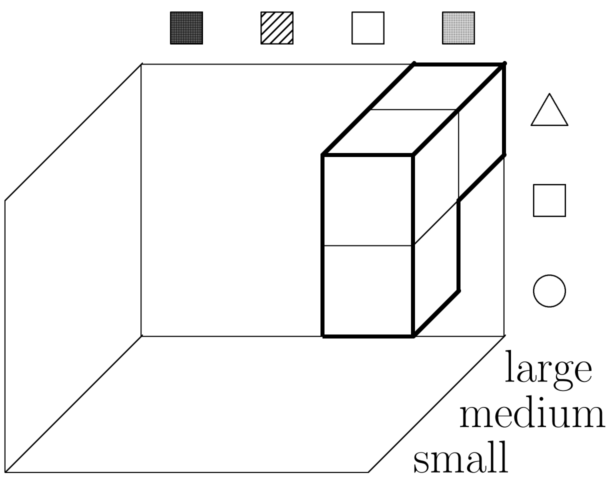
Geometric Interpretation

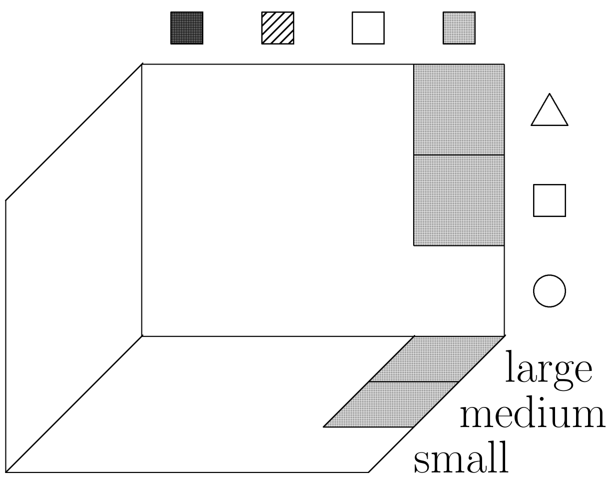


Each cube represents one tuple.

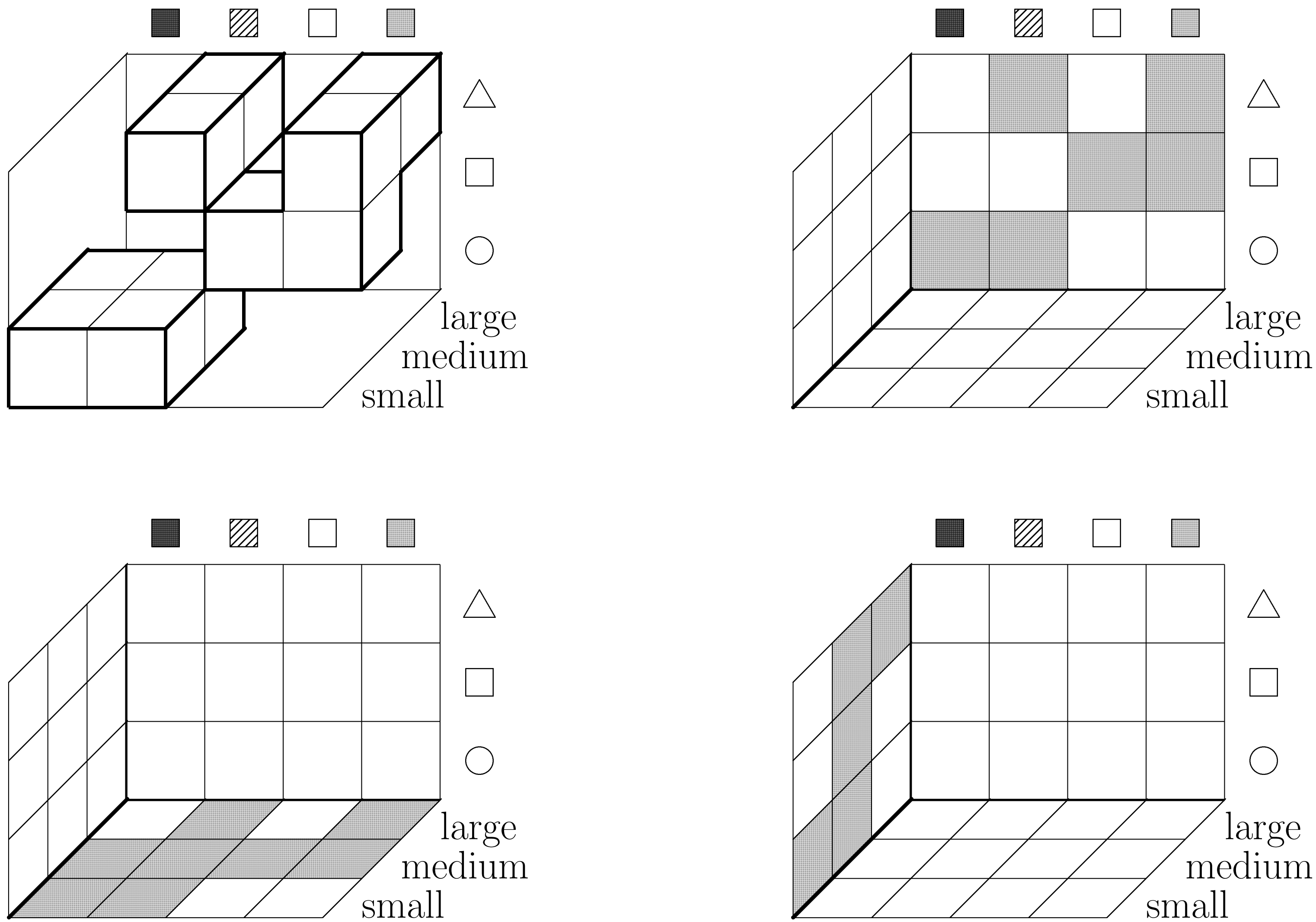
Reasoning

- 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.

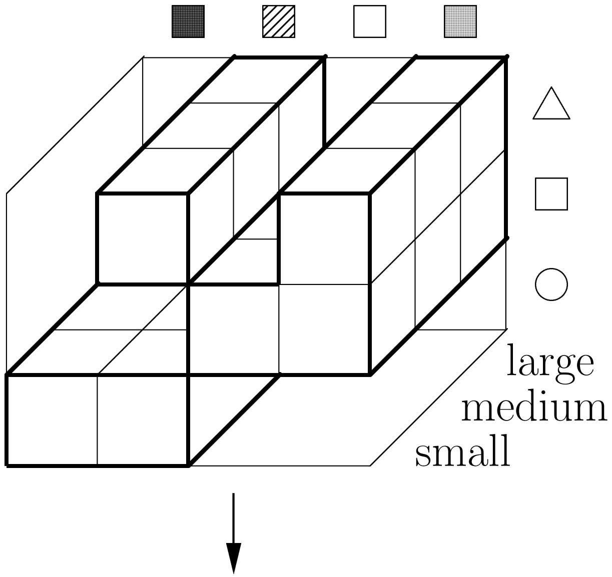




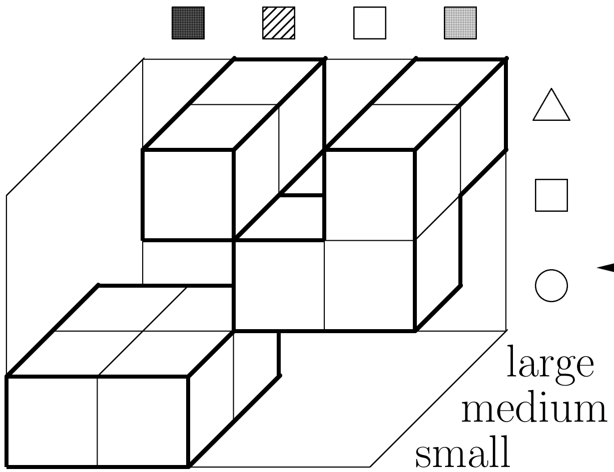
Prior Knowledge and Its Projections

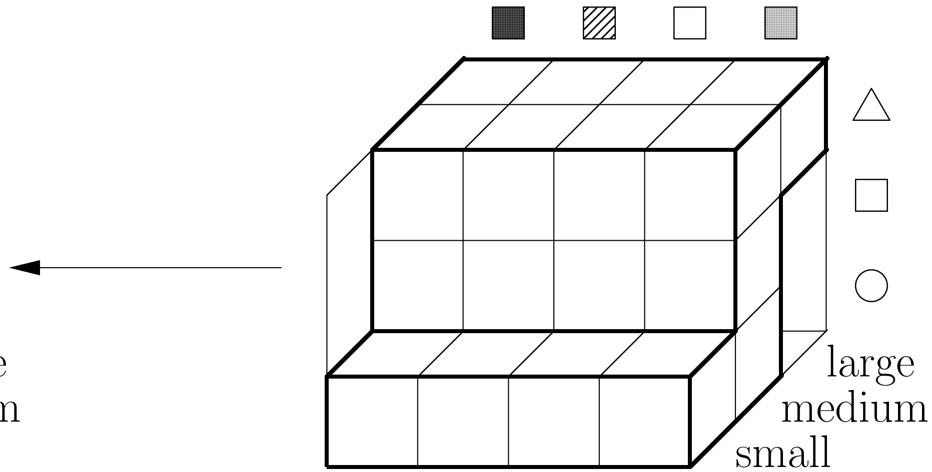


Cylindrical Extensions and Their Intersection



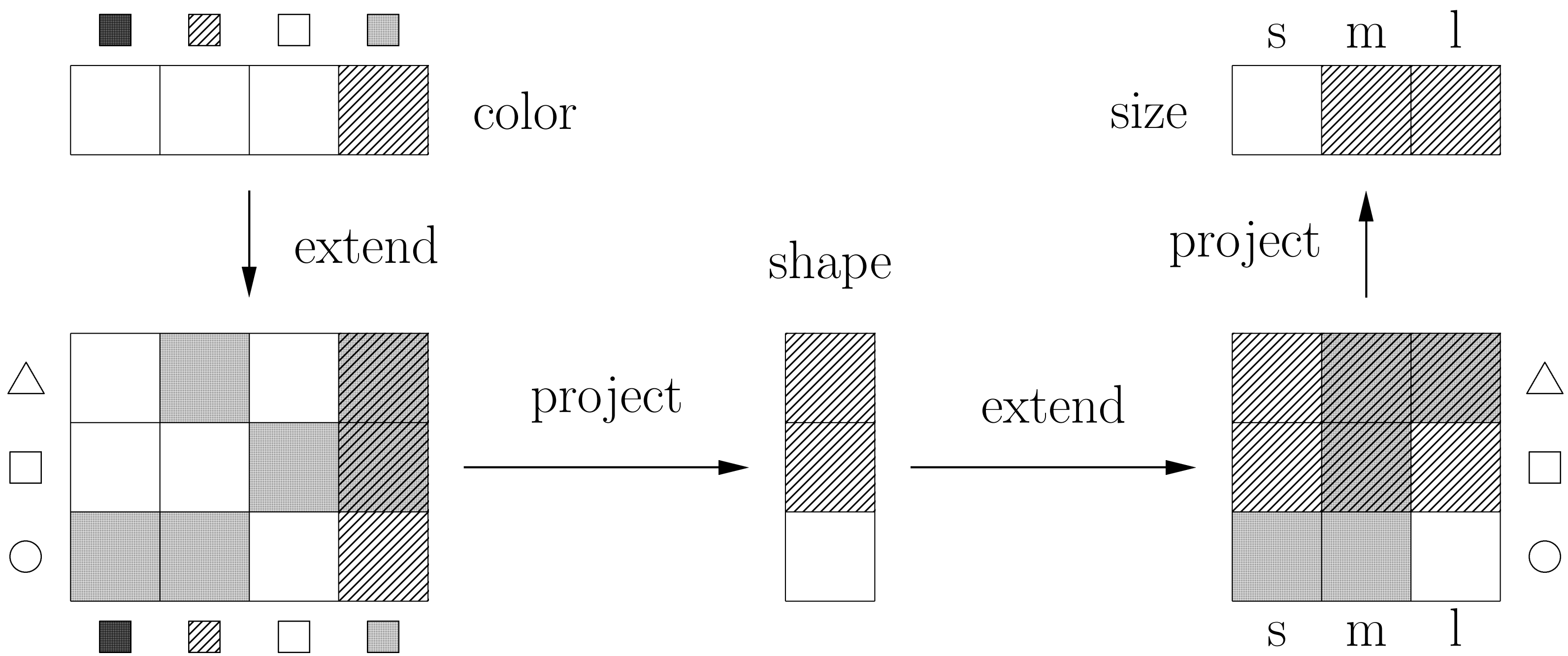
Intersecting the cylindrical extensions of the projection to the subspace formed by color and shape and of the projection to the subspace formed by shape and size yields the original three-dimensional relation.



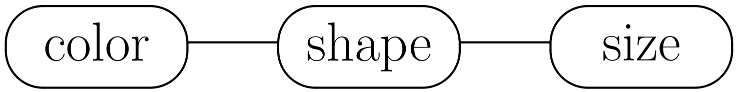


Reasoning with Projections

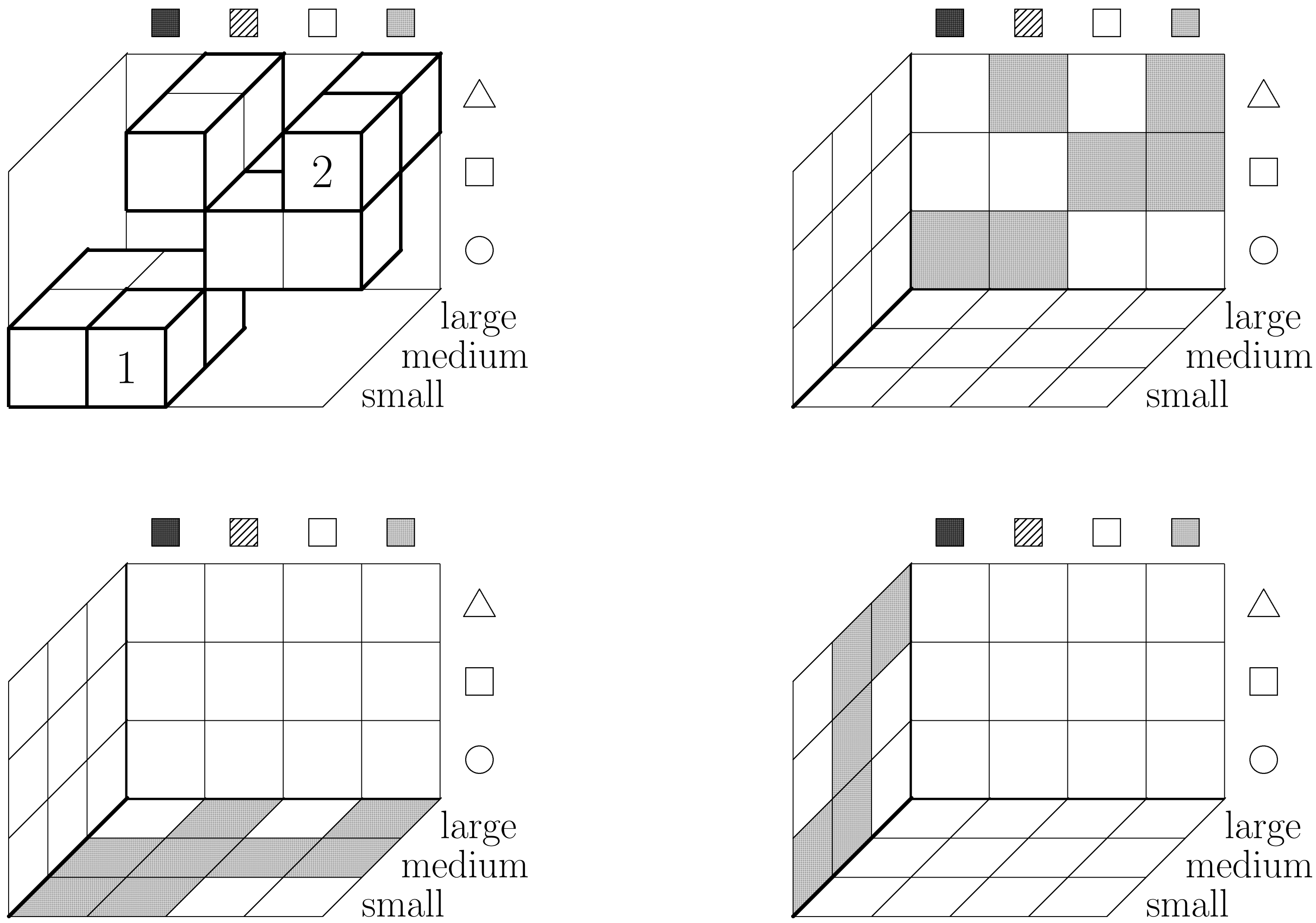
The same result can be obtained using only the projections to the subspaces without reconstructing the original three-dimensional space:



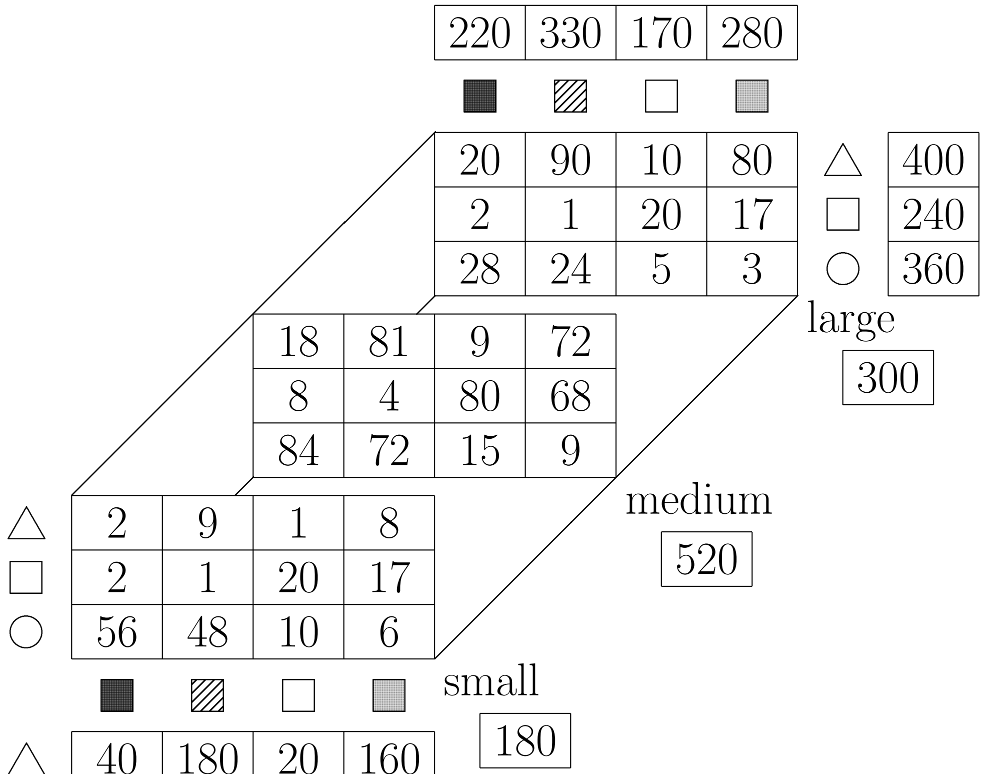
This justifies a network representation:



Is Decomposition Always Possible?



A Probability Distribution



all numbers in parts per 1000

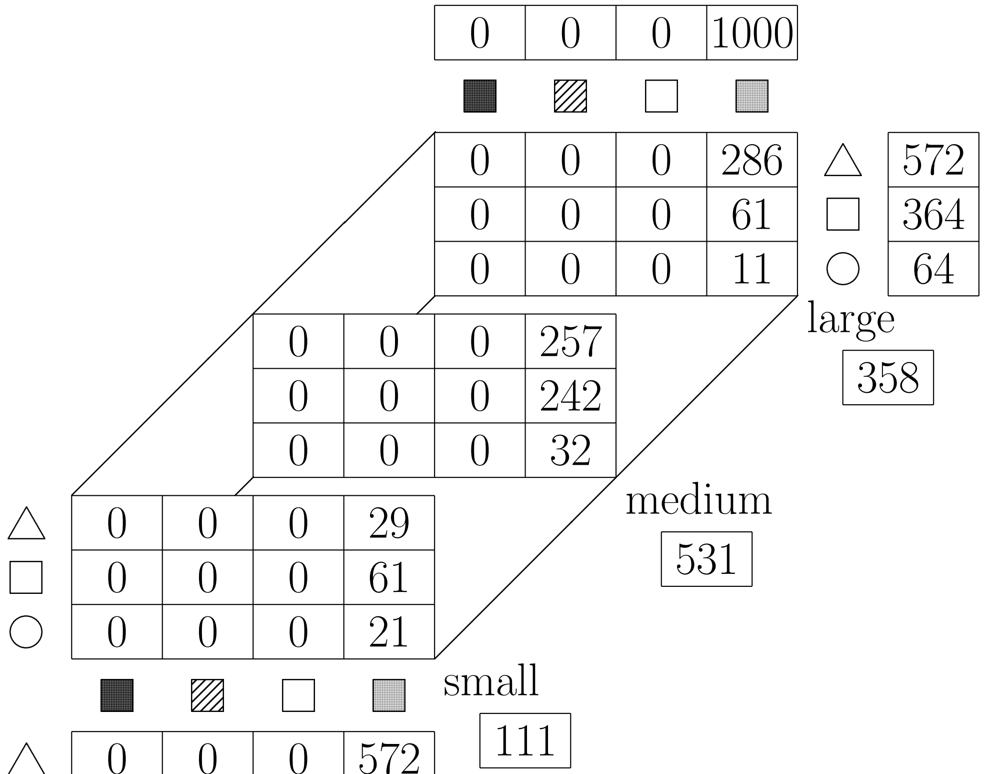
	■	▨	□	▩
△	40	180	20	160
□	12	6	120	102
○	168	144	30	18

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

	■	▨	□	▩
large	50	115	35	100
medium	110	157	104	149
small	60	58	31	31

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

Reasoning



all numbers in
parts per 1000

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

	s	m	l
△	29	257	286
□	61	242	61
○	21	32	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 color × shape and
  – the marginal distribution on the subspace shape × size.

- It can be reconstructed using the following formula:

$$\forall i,j,k:\; P(\omega_i^{(\text{color})},\omega_j^{(\text{shape})},\omega_k^{(\text{size})}) = P(\omega_i^{(\text{color})},\omega_j^{(\text{shape})})\cdot P(\omega_k^{(\text{size})} \mid \omega_j^{(\text{shape})})$$

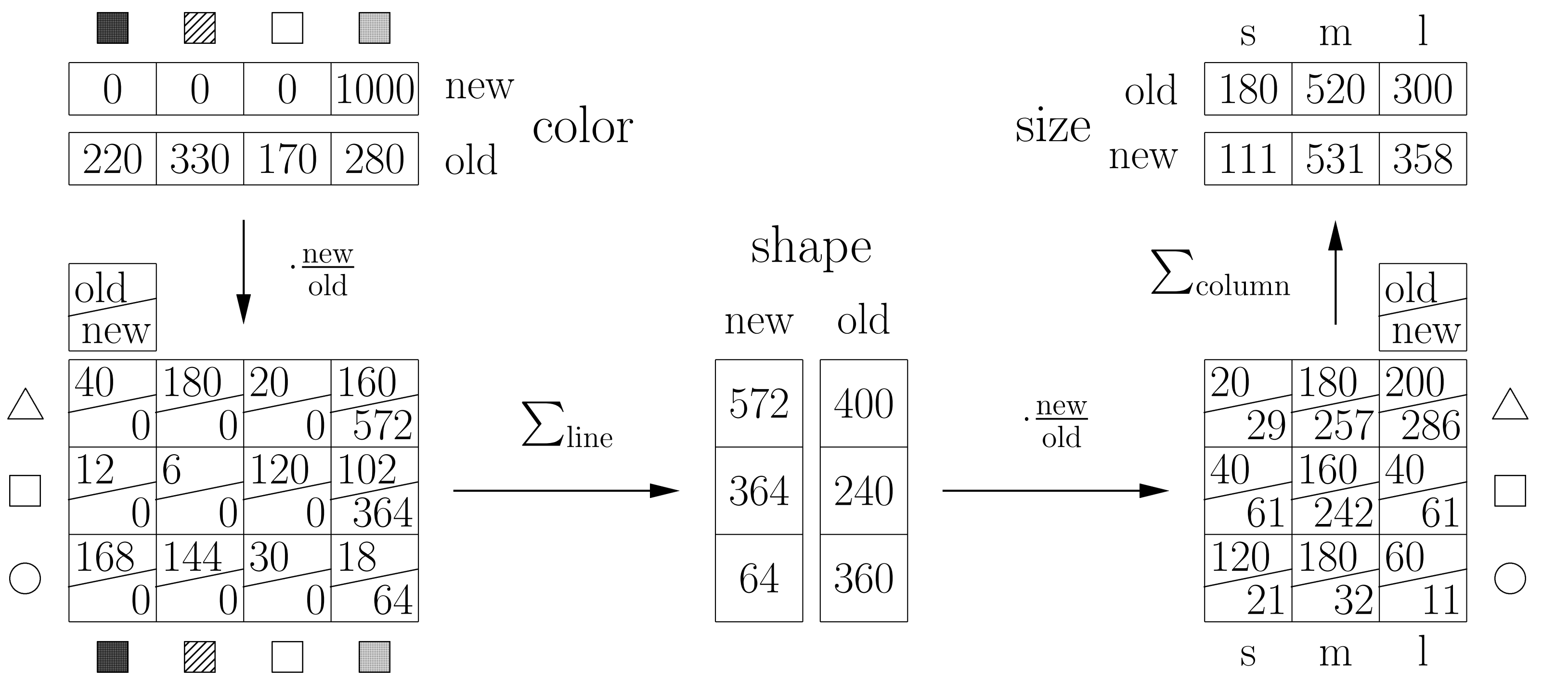
$$= P(\omega_i^{(\text{color})},\omega_j^{(\text{shape})})\cdot \frac{P(\omega_j^{(\text{color})},\omega_k^{(\text{size})})}{P(\omega_j^{(\text{shape})})}$$

- This formula expresses the **conditional independence** of the attributes *color* and *size* given the attribute *shape*, since they only hold if

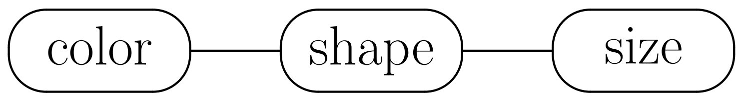
$$\forall i,j,k:\quad P(\omega_k^{(\text{size})} \mid \omega_j^{(\text{shape})}) = P(\omega_k^{(\text{size})} \mid \omega_i^{(\text{color})},\omega_j^{(\text{shape})})$$

Reasoning with Projections

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



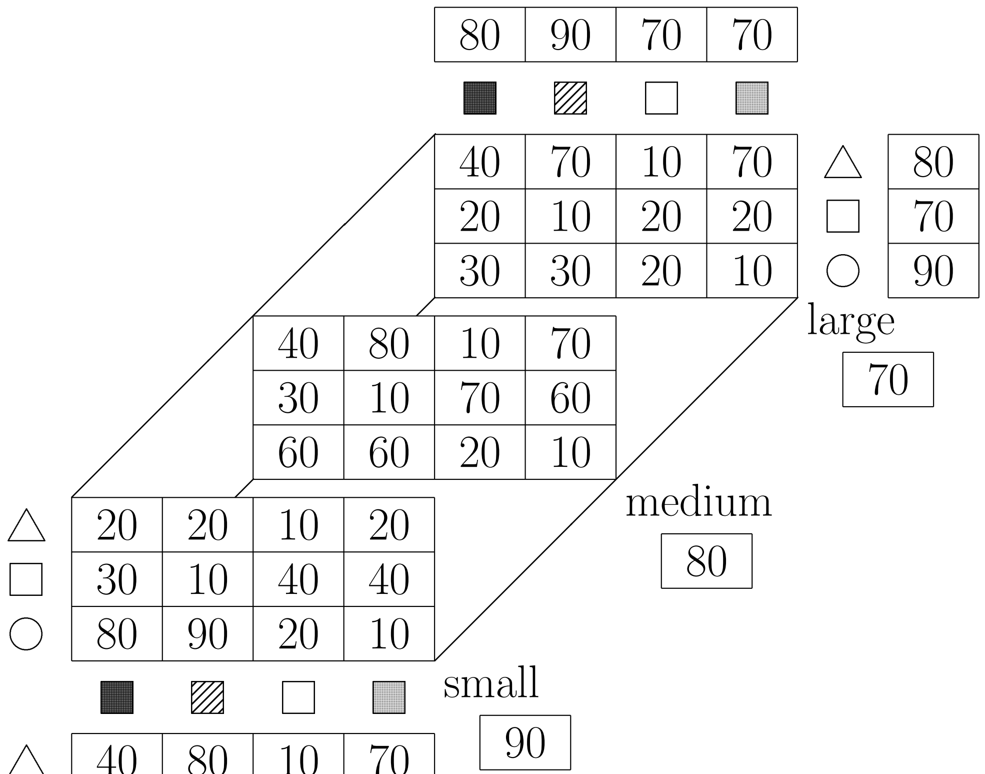
This justifies a network representation:



Probabilistic Evidence Propagation, Step 1

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

A Possibility Distribution



△	40	80	10	70
□	30	10	70	60
○	80	90	20	10

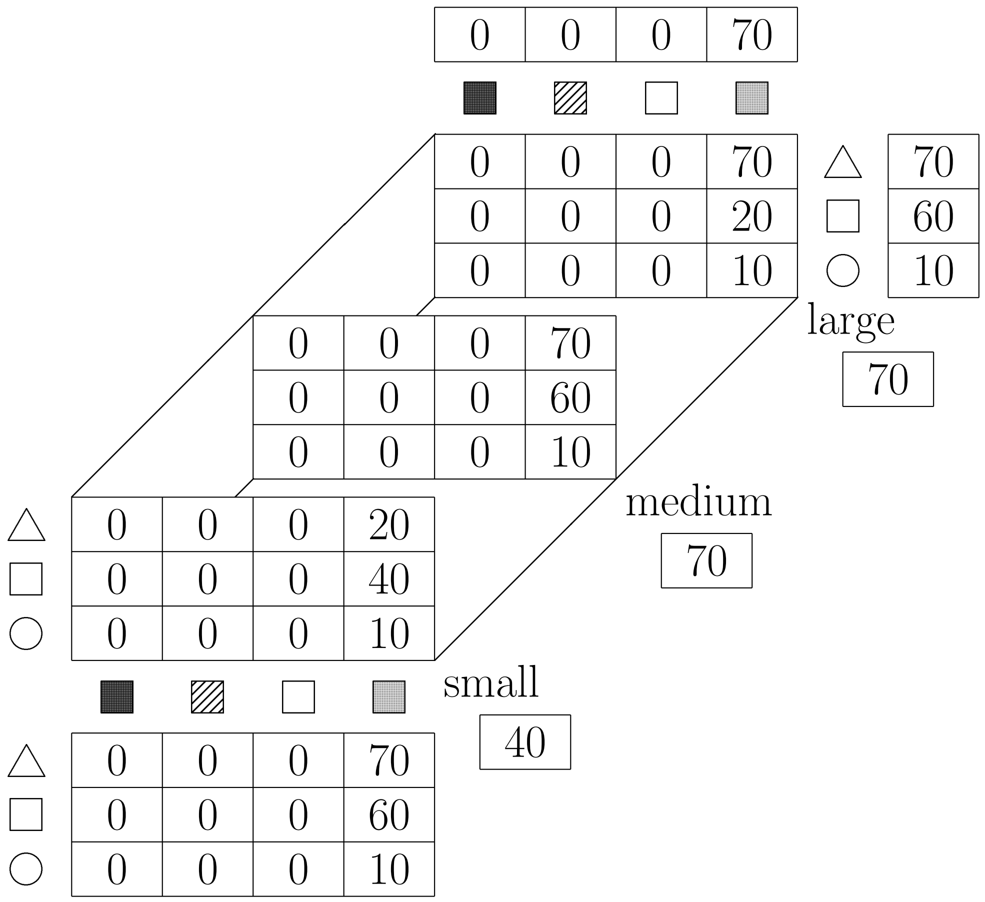
all numbers in parts per 1000

	s	m	l
△	20	80	70
□	40	70	20
○	90	60	30

large	40	70	20	70
medium	60	80	70	70
small	80	90	40	40

- The numbers state the degrees of possibility of the corresp. value combination.

Reasoning



all numbers in
parts per 1000

	s	m	l
△	20	70	70
□	40	60	20
○	10	10	10

large	0	0	0	70
medium	0	0	0	70
small	0	0	0	40

- Using the information that the given object is green.

Possibilistic Decomposition

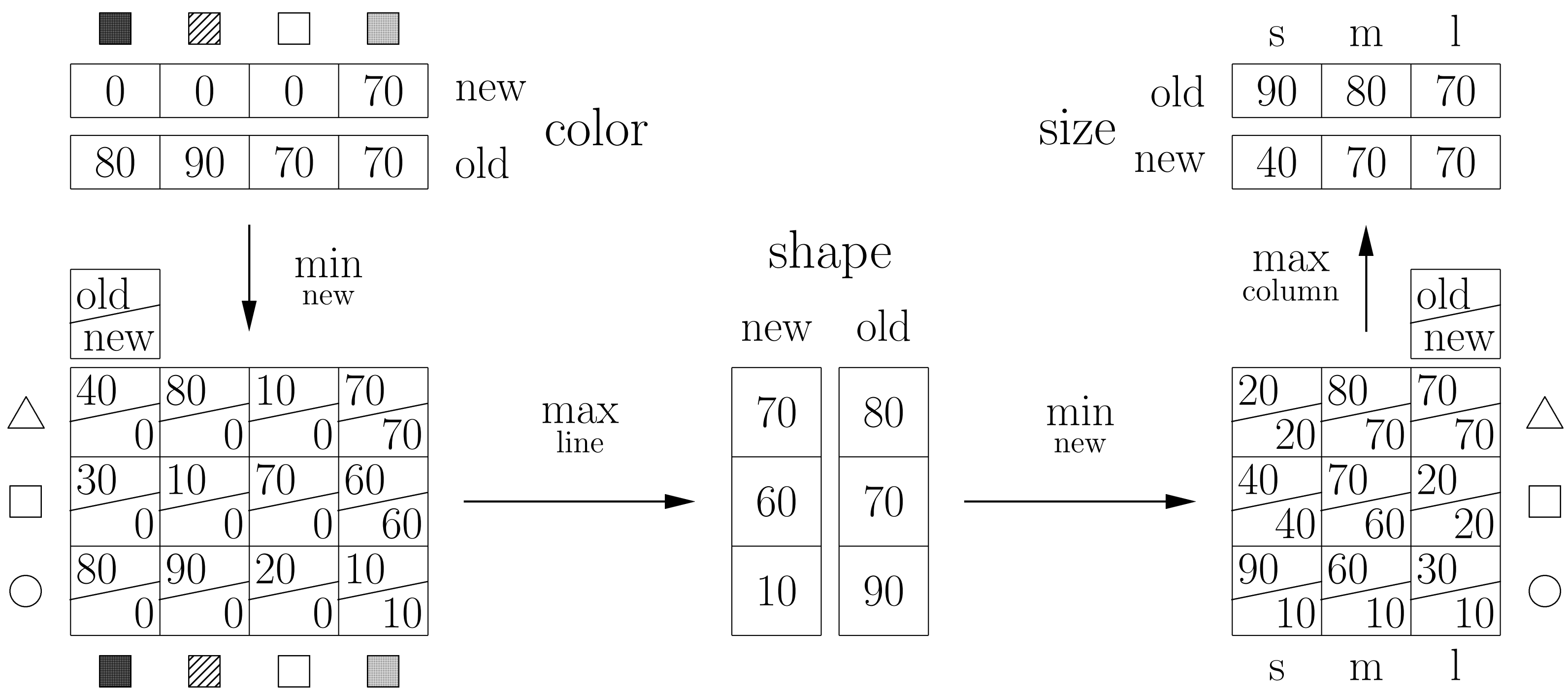
- As for relational and probabilistic networks, the three-dimensional possibility distribution can be decomposed into projections to subspaces, namely:
  – the maximum projection to the subspace color × shape and
  – the maximum projection to the subspace shape × size.

- It can be reconstructed using the following formula:

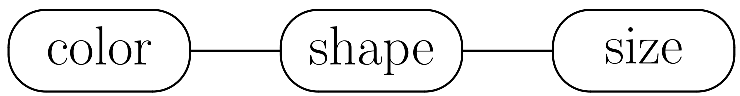
$$\begin{aligned}
\forall i,j,k:\ & \pi(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}, \omega_k^{(\text{size})}) \\
&= \min\left\{\pi(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}), \pi(\omega_j^{(\text{shape})}, \omega_k^{(\text{size})})\right\} \\
&= \min\Big\{ \max_k \pi(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}, \omega_k^{(\text{size})}), \\
&\qquad\quad \max_i \pi(\omega_i^{(\text{color})}, \omega_j^{(\text{shape})}, \omega_k^{(\text{size})})\Big\}
\end{aligned}$$

Reasoning with Projections

Again the same result can be obtained using only projections to subspaces (maximal degrees of possibility):



This justifies a network representation:

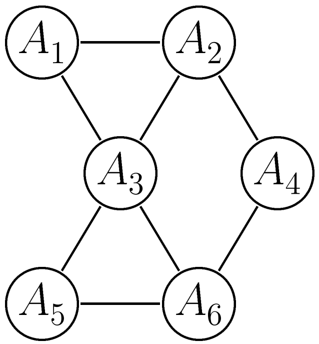


Possibilistic Evidence Propagation, Step 1

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

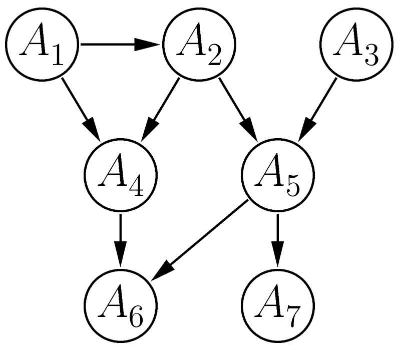
Graphs and Decompositions

Undirected Graphs



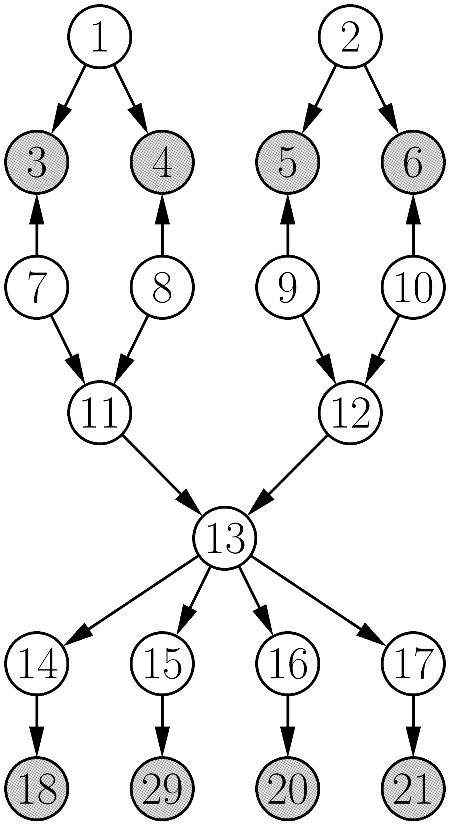
$$\begin{aligned}
&\pi_U(A_1 = a_1, \ldots, A_6 = a_6) \\
&\quad = \min\{ \quad \pi_{A_1A_2A_3}(A_1 = a_1, A_2 = a_2, A_3 = a_3), \\
&\qquad\qquad\quad \pi_{A_3A_5A_6}(A_3 = a_3, A_5 = a_5, A_6 = a_6), \\
&\qquad\qquad\quad \pi_{A_2A_4}(A_2 = a_2, A_4 = a_4), \\
&\qquad\qquad\quad \pi_{A_4A_6}(A_4 = a_4, A_6 = a_6) \quad \}
\end{aligned}$$

Directed Graphs



$$\begin{aligned}
&\pi_U(A_1 = a_1, \ldots, A_7 = a_7) \\
&\quad = \min\{ \quad \pi(A_1 = a_1), \pi(A_2 = a_2 \mid A_1 = a_1), \pi(A_3 = a_3), \\
&\qquad\qquad\quad \pi(A_4 = a_4 \mid A_1 = a_1, A_2 = a_2), \\
&\qquad\qquad\quad \pi(A_5 = a_5 \mid A_2 = a_2, A_3 = a_3), \\
&\qquad\qquad\quad \pi(A_6 = a_6 \mid A_4 = a_4, A_5 = a_5), \\
&\qquad\qquad\quad \pi(A_7 = a_7 \mid A_5 = a_5) \quad \}
\end{aligned}$$

Example: Danish Jersey Cattle Blood Type Determination

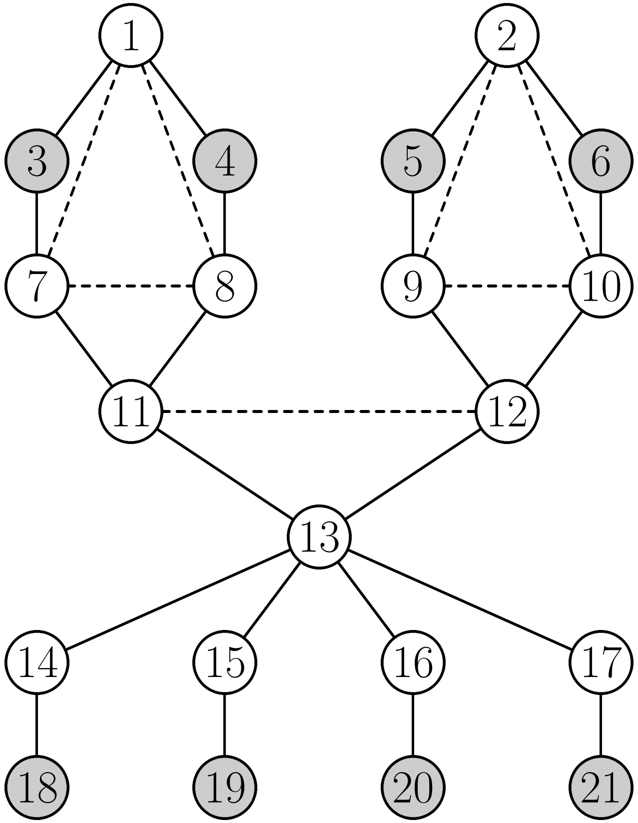


21 attributes:

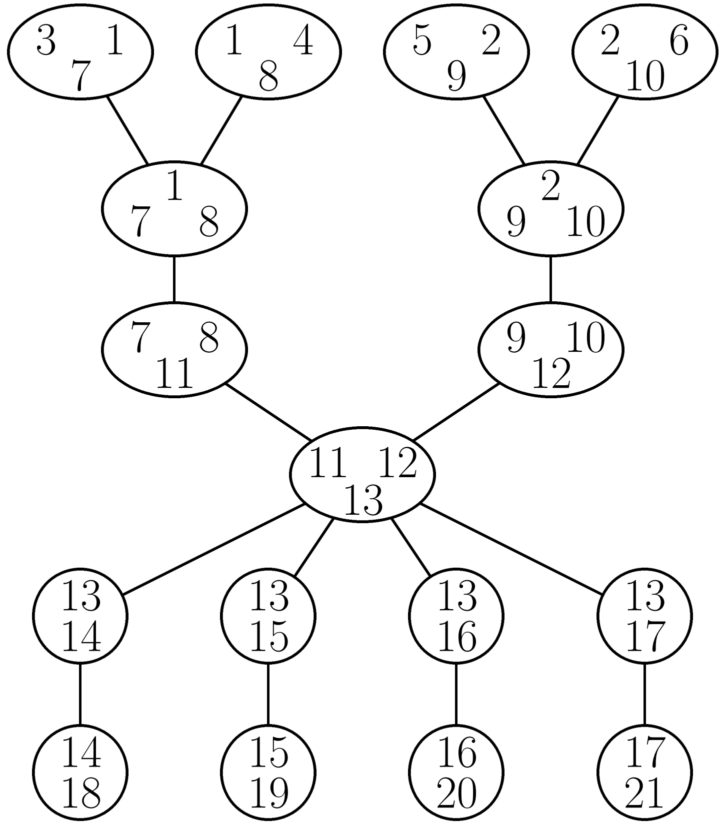
1 – dam correct?
2 – sire correct?
3 – stated dam ph.gr. 1
4 – stated dam ph.gr. 2
5 – stated sire ph.gr. 1
6 – stated sire ph.gr. 2
7 – true dam ph.gr. 1
8 – true dam ph.gr. 2
9 – true sire ph.gr. 1
10 – true sire ph.gr. 2
11 – offspring ph.gr. 1
12 – offspring ph.gr. 2
13 – offspring genotype
14 – 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.

Example: Danish Jersey Cattle Blood Type Determination



Moral Graph



Join Tree

Learning Possibilistic Graphical Models from Data

Quantitative or Parameter Learning

- Determine the parameters of the (marginal or conditional) distributions indicated by a given graph from a database of sample cases.
  - Trivial in the relational and the probabilistic case.
  - In the possibilistic case, however, this poses a problem.

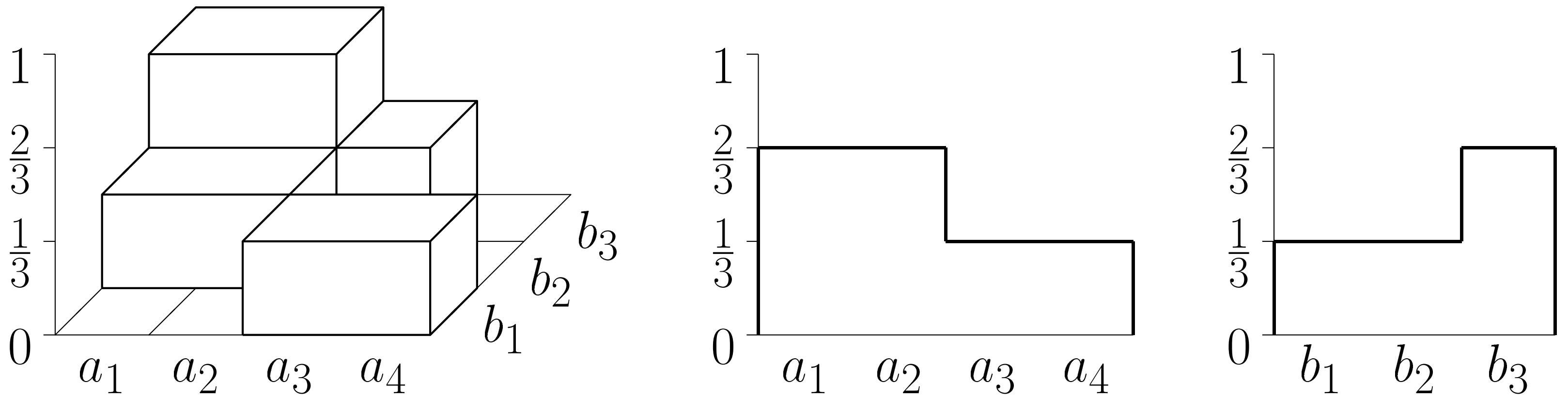
Qualitative or Structural Learning

- Find a graph that describes (a good approximation of) a decomposition of the distribution underlying a database of sample cases.
  - Has been a popular area of research in recent years.
  - Several good algorithms exit for the probabilistic case.
  - Most ideas can easily be transferred to the possibilistic case.

Why is Computing Maximum Projections a Problem?

Database: $(\{a_1, a_2, a_3\}, \{b_3\}) : {}^1\!/_3$
$(\{a_1, a_2\}, \{b_2, b_3\}) : {}^1\!/_3$
$(\{a_3, a_4\}, \{b_1\}) : {}^1\!/_3$

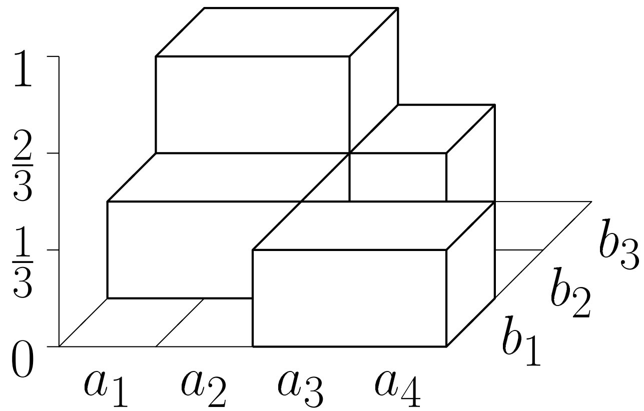
There are 3 tuples (contexts), hence the weight of each is ${}^1\!/_3$.



- Taking the maximum over all tuples containing a_1 to compute $\pi(A = a_1)$ yields a possibility degree of ${}^1\!/_3$, but actually it is ${}^2\!/_3$.
- Taking the sum over all tuples containing a_3 to compute $\pi(A = a_3)$ yields a possibility degree of ${}^2\!/_3$, but actually it is ${}^1\!/_3$.

Computation via Support and Closure

Database	Support		Closure
$(\{a_1, a_2, a_3\}, \{b_3\}) : {}^1/_3$ $(\{a_1, a_2\}, \{b_2, b_3\}) : {}^1/_3$ $(\{a_3, a_4\}, \{b_1\}) : {}^1/_3$	$(a_1, b_2) : {}^1/_3$ $(a_1, b_3) : {}^2/_3$ $(a_2, b_2) : {}^1/_3$ $(a_2, b_3) : {}^2/_3$	$(a_3, b_1) : {}^1/_3$ $(a_3, b_3) : {}^1/_3$ $(a_4, b_1) : {}^1/_3$	$(\{a_1, a_2, a_3\}, \{b_3\}) : {}^1/_3$ $(\{a_1, a_2\}, \{b_2, b_3\}) : {}^1/_3$ $(\{a_3, a_4\}, \{b_1\}) : {}^1/_3$ $(\{a_1, a_2\}, \{b_3\}) : {}^2/_3$
3 tuples	7 tuples		4 tuples



Taking the maximum over compatible tuples in the support yields the same result as taking the maximum over compatible tuples in the closure [Borgelt and Kruse 1998].

Experimental Results

dataset	number of cases	tuples in R	tuples in support(R)	tuples in closure(R)
Danish Jersey Cattle	500	283	712818	291
Soybean Diseases	683	631	n.a.	631
Congress Voting Data	435	342	98753	400

- The relation R results from the dataset by removing duplicate tuples.
- The frequency information is kept in a counter associated with each tuple.
- None of these databases is a true "imprecise" database, the only imprecision results from unknown values.
- An unknown value for an attribute A is interpreted as the set dom(A).
- "n.a." (not available) means that the relation is too large to be computed.

Naive Bayes Classifiers

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

Bayes' Rule:

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

Chain Rule of Probability:

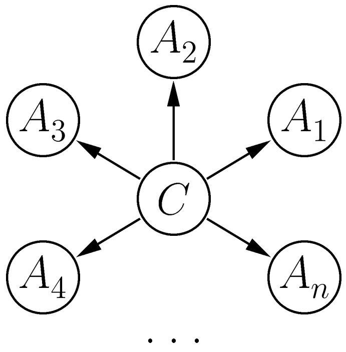
$$P(C = c_i \mid \mathbf{e}) = \frac{P(C = c_i)}{p_0} \cdot \prod_{j=1}^{n} P(A_j = a_j \mid A_1 = a_1, \ldots, A_{j-1} = a_{j-1}, C = c_i)$$

Conditional Independence Assumptions:

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

Star-like Probabilistic Networks

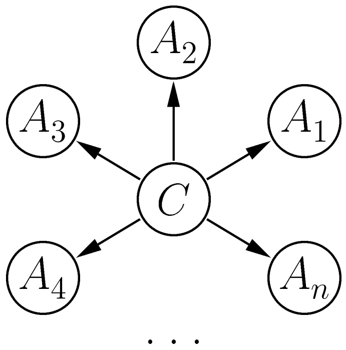
- A naive Bayes classifier is a probabilistic network with a star-like structure.
- Class attribute is the only unconditioned attribute.
- All other attributes are conditioned on the class only.



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

A Naive Possibilistic Classifier

- Idea: Possibilistic network with a star-like structure. [Borgelt and Gebhardt 1999].
- Class attribute is the only unconditioned attribute.
- All other attributes are conditioned on the class only.



$$\pi(C = c_i, \mathbf{e}) = \pi(C = c_i \mid \mathbf{e}) = \min_{j=1}^{n} \pi(A_j = a_j \mid C = c_i)$$

Naive Possibilistic Classifiers

- Try to compute $\pi(C = c_i \mid \mathbf{e}) = \pi(C = c_i \mid A_1 = a_1, \ldots, A_n = a_n)$.
- Predict the class with the highest conditional degree of possibility.

Analog of Bayes' Rule:

$$\pi(C = c_i \mid \mathbf{e}) = \pi(A_1 = a_1, \ldots, A_n = a_n \mid C = c_i)$$

Chain Rule of Possibility:

$$\pi(C = c_i \mid \mathbf{e}) = \min\nolimits_{j=1}^{n} \pi(A_j = a_j \mid A_1 = a_1, \ldots, A_{j-1} = a_{j-1}, C = c_i)$$

Conditional Independence Assumptions:

$$\pi(C = c_i \mid \mathbf{e}) = \min\nolimits_{j=1}^{n} \pi(A_j = a_j \mid C = c_i)$$

Experimental Results

dataset		num. of tuples	possibilistic classifier add. att.	possibilistic classifier rem. att.	naive Bayes classifier add. att.	naive Bayes classifier rem. att.	decision tree unpruned	decision tree pruned
audio	train	113	7(6.2%)	2(1.8%)	12(10.6%)	16(14.2%)	13(11.5%)	16(14.2%)
	test	113	33(29.2%)	36(31.9%)	35(31.0%)	31(27.4%)	25(22.1%)	25(22.1%)
69 atts.	selected		15	21	9	42	14	12
bridges	train	54	8(14.8%)	8(14.8%)	10(18.5%)	7(13.0%)	9(16.7%)	9(16.7%)
	test	54	23(42.6%)	23(42.6%)	24(44.4%)	19(35.2%)	24(44.4%)	24(44.4%)
10 atts.	selected		6	6	5	8	8	6
soybean	train	342	18(5.3%)	20(5.9%)	17(5.0%)	14(4.1%)	16(4.7%)	22(6.4%)
	test	341	59(17.3%)	57(16.7%)	48(14.1%)	45(13.2%)	47(13.8%)	39(11.4%)
36 atts.	selected		15	17	14	14	19	16
vote	train	300	9(3.0%)	8(2.7%)	9(3.0%)	8(2.7%)	6(2.0%)	7(2.3%)
	test	135	11(8.2%)	10(7.4%)	11(8.2%)	8(5.9%)	11(8.2%)	8(5.9%)
16 atts.	selected		2	3	2	4	6	4

- Possibilistic classifier performs equally well or only slightly worse.
- Datasets are not well suited to show the strengths of a possibilistic approach.

Learning the Structure of Graphical Models

- **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.

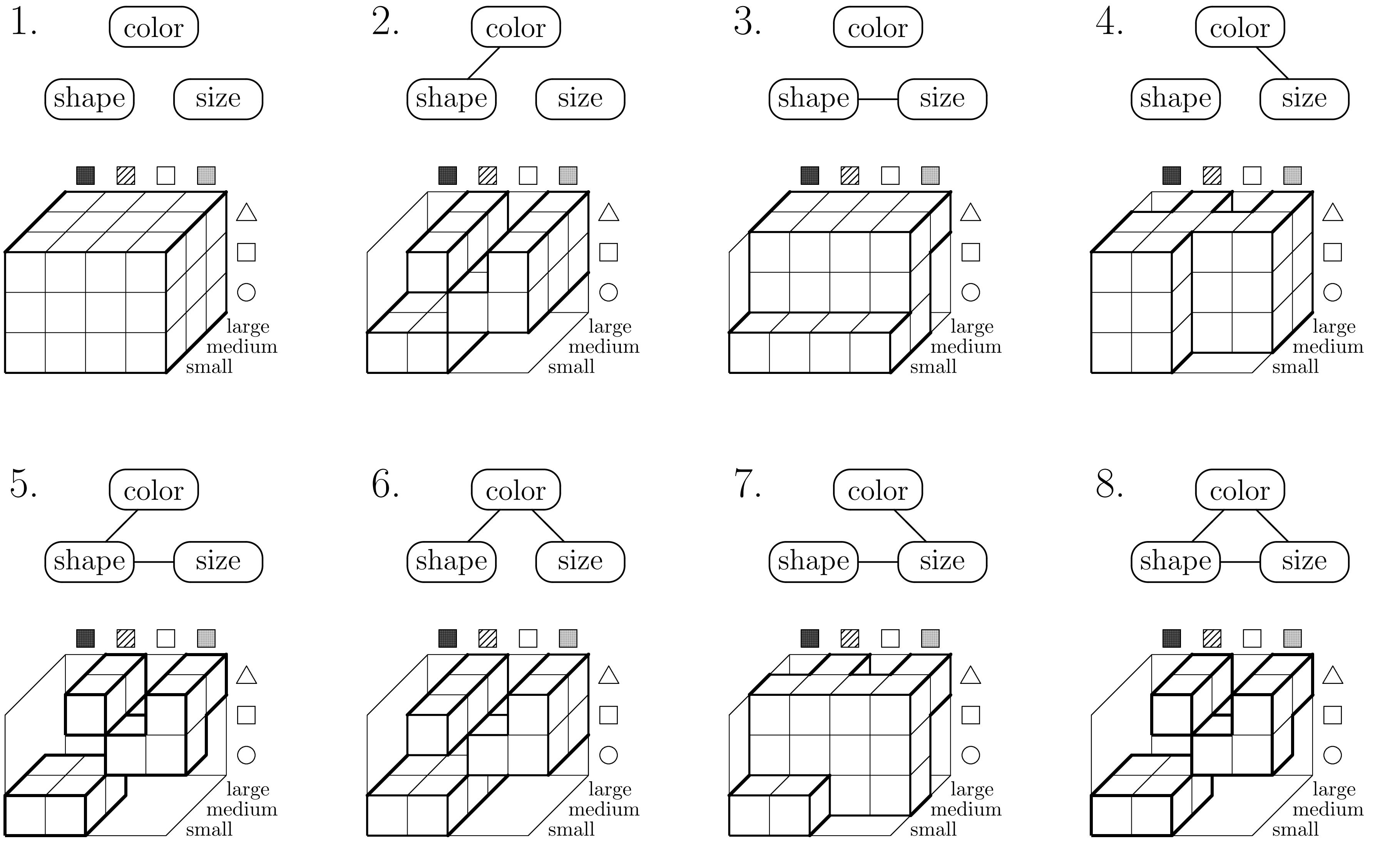
- **Find an independence map by conditional independence tests.**

  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 distribution that is decomposable w.r.t. a graph an attribute is more strongly dependent on adjacent attributes than on attributes that are not directly connected to them.

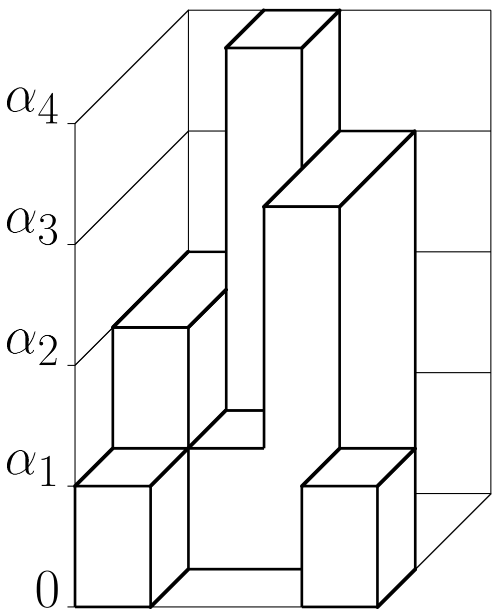
Learning Graphical Models from Data

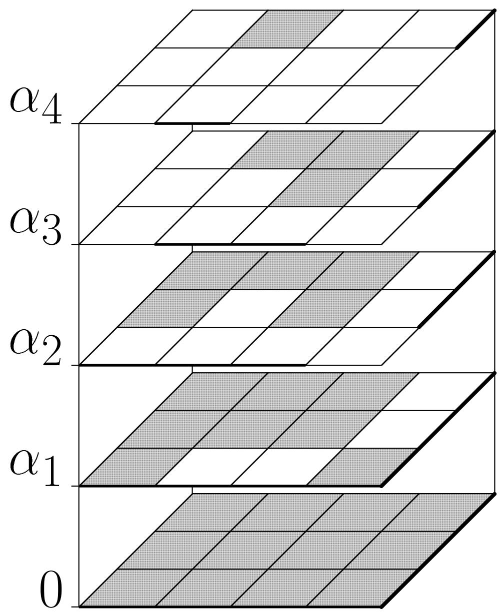


α-Cut View of Possibility Distributions

Definition: Let Π be a possibility measure on a sample space Ω.
The α**-cut** of Π, written $[\Pi]_\alpha$, is the function

$$[\Pi]_\alpha : 2^\Omega \to \{0,1\}, \qquad E \mapsto \begin{cases} 1, & \text{if } \Pi(E) \geq \alpha, \\ 0, & \text{otherwise.} \end{cases}$$





Evaluating Approximations of Possibility Distributions

The α-cut view of possibility distributions suggests the following measure for the "closeness" of an approximate decomposition to the original distribution:

$$\mathrm{diff}(\pi_1, \pi_2) = \int_0^1 \Big(\sum_{E \in \mathcal{E}} [\pi_2]_\alpha(E) - \sum_{E \in \mathcal{E}} [\pi_1]_\alpha(E) \Big) \, \mathrm{d}\alpha,$$

where π_1 is the original distribution, π_2 is the approximation, and $\mathcal{E}$ is their domain of definition.

- This measure is zero if the two distributions coincide and it is the larger, the more they differ.
- This measure presupposes that $\forall \alpha \in [0,1] : \forall E \in \mathcal{E} : [\pi_2]_\alpha(E) \geq [\pi_1]_\alpha(E)$

Specificity Divergence

Definition: Let π be a possibility distribution on a set $\mathcal{E}$ of events. Then

$$\text{nonspec}(\pi) = \int_0^{\sup_{E \in \mathcal{E}} \pi(E)} \log_2 \Big(\sum_{E \in \mathcal{E}} [\pi]_\alpha(E) \Big) \, \mathrm{d}\alpha$$

is called the **nonspecificity** of the possibility distribution π.

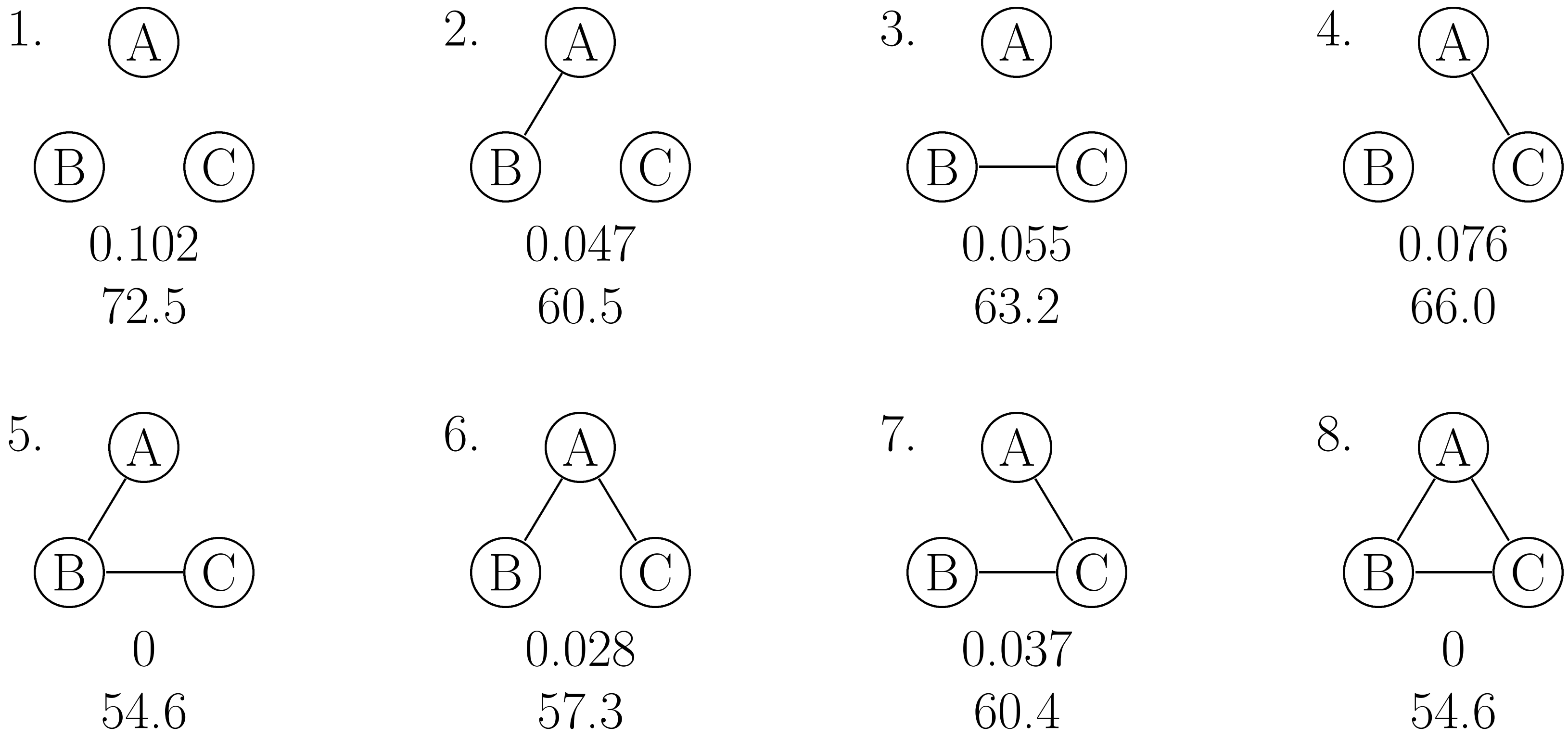
- U-uncertainty measure of nonspecificity [Higashi and Klir 1982].
- Generalization of Hartley information [Hartley 1928].

Definition: Let π_1 and π_2 be two possibility distributions on the same set $\mathcal{E}$ of events with $\forall E \in \mathcal{E} : \pi_2(E) \geq \pi_1(E)$. Then

$$S_{\text{div}}(\pi_1, \pi_2) = \int_0^{\sup_{E \in \mathcal{E}} \pi_1(E)} \log_2 \Big(\sum_{E \in \mathcal{E}} [\pi_2]_\alpha(E) \Big) - \log_2 \Big(\sum_{E \in \mathcal{E}} [\pi_1]_\alpha(E) \Big) \, \mathrm{d}\alpha$$

is called the **specificity divergence** of π_1 and π_2.

Direct Test for Decomposability (continued)



Upper numbers: Specificity divergence of the original distribution and its approximation.

Lower numbers: Sum of possibility degrees for an example database that induces the possibility distribution.

Evaluation w.r.t. a Database of Sample Cases

Transformation of the difference of two possibility distributions:

$$\begin{aligned}
\mathrm{diff}(\pi_1, \pi_2) &= \int_0^1 \Big(\sum_{E \in \mathcal{E}} [\pi_2]_\alpha(E) - \sum_{E \in \mathcal{E}} [\pi_1]_\alpha(E) \Big) \, \mathrm{d}\alpha \\
&= \sum_{E \in \mathcal{E}} \int_0^1 [\pi_2]_\alpha(E) \, \mathrm{d}\alpha - \sum_{E \in \mathcal{E}} \int_0^1 [\pi_1]_\alpha(E) \, \mathrm{d}\alpha \\
&= \sum_{E \in \mathcal{E}} \pi_2(E) - \sum_{E \in \mathcal{E}} \pi_1(E).
\end{aligned}$$

- $\sum_{E \in \mathcal{E}} \pi_1(E)$ can be neglected, since it is the same for all decompositions.
- Restriction to the sample cases in a given database $D = (R, w_R)$.
  ($w_R(t)$ is the *weight*, i.e., the number of occurrences, of a tuple $t \in R$.)

$$Q(G) = \sum_{t \in R} w_R(t) \cdot \pi_G(t)$$

Direct Test for Decomposability (continued)

- Problem: **Vast Search Space** (huge number of possible graphs)
  - $2^{\binom{n}{2}}$ possible undirected graphs for n attributes.
  - Between $2^{\binom{n}{2}}$ and $3^{\binom{n}{2}}$ possible directed acyclic graphs.

    Exact formula: $f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i)$.

- **Restriction of the Search Space**
  - Fix topological order (for directed graphs)
  - Declarative bias (idea from inductive logic programming)

- **Heuristic Search Methods**
  - Greedy Search
  - Simulated Annealing
  - Genetic Algorithms

A Simulated Annealing Approach

Definition: Let $G = (U, E)$ be a graph, $\mathcal{M}$ the family of node sets that induce the maximal cliques of G and $m = |\mathcal{M}|$. G is said to have **hypertree structure** iff all pairs of nodes are connected in G and there is an ordering $M_1, \ldots, M_m$ of the sets in $\mathcal{M}$, such that

$$\forall i \in \{2, \ldots, m\} : \exists k \in \{1, \ldots, i-1\} : \quad M_i \cap \Big(\bigcup_{1 \leq j < i} M_j \Big) \subseteq M_k.$$

Random construction/modification of a graph with hypertree structure by adding cliques randomly according to the following rules [Borgelt 2000]:

1. M_i must contain at least one pair of nodes that are not connected in the graph represented by $\{M_1, \ldots, M_{i-1}\}$.
2. For each maximal subset S of nodes of M_i that are connected to each other in the graph represented by $\{M_1, \ldots, M_{i-1}\}$ there must be a set M_k, $1 \leq k < i$, so that $S \subset M_k$.

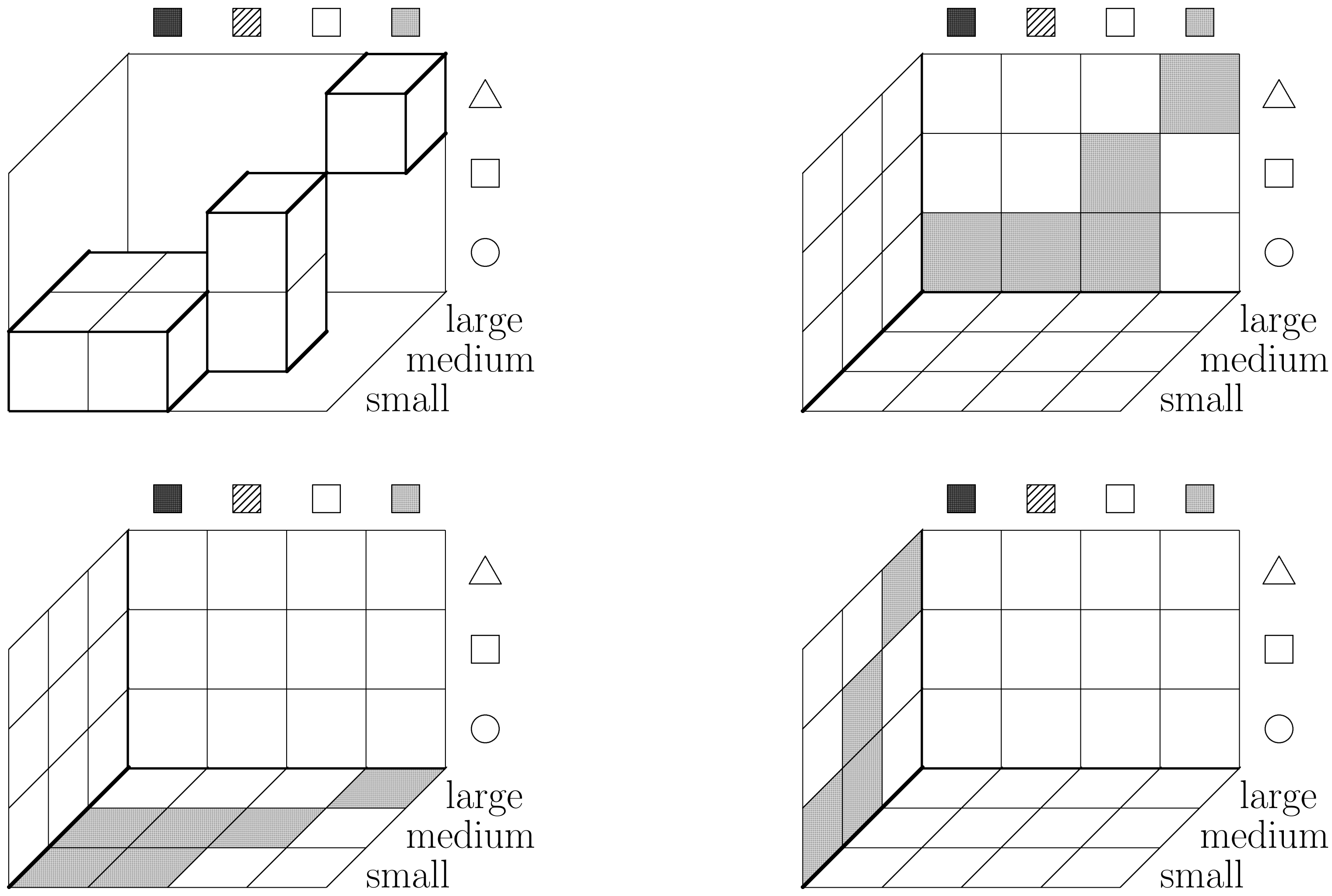
Measuring the Strengths of Marginal Dependences

- **Relational networks:** Find a set of subspaces, for which the intersection of the cylindrical extensions of the projections to these subspaces contains as few additional states as possible.
- The size of the intersection depends on the sizes of the cylindrical extensions, which in turn depend on the sizes of the projections.
- Therefore it is plausible to use the relative number of occurring value combinations to assess the quality of a subspace.

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

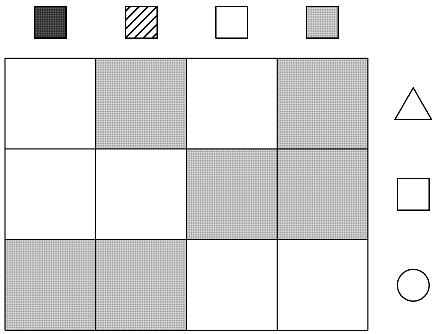


Hartley Information Gain

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

$$I_{\text{gain}}^{(\text{Hartley})}(A, B) = \log_2\Big(\sum_{a\in\text{dom}(A)} R(A = a)\Big) + \log_2\Big(\sum_{b\in\text{dom}(B)} R(B = b)\Big) - \log_2\Big(\sum_{a\in\text{dom}(A)} \sum_{b\in\text{dom}(B)} R(A = a, B = b)\Big),$$

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



Hartley information needed to determine

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

Specificity Gain

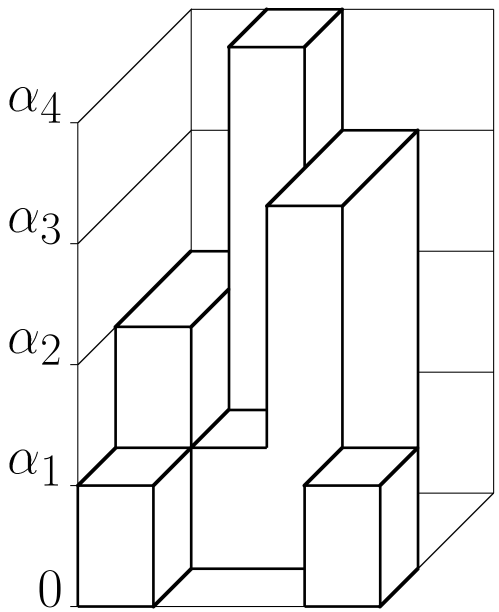
Definition: Let A and B be two attributes and Π a possibility measure.

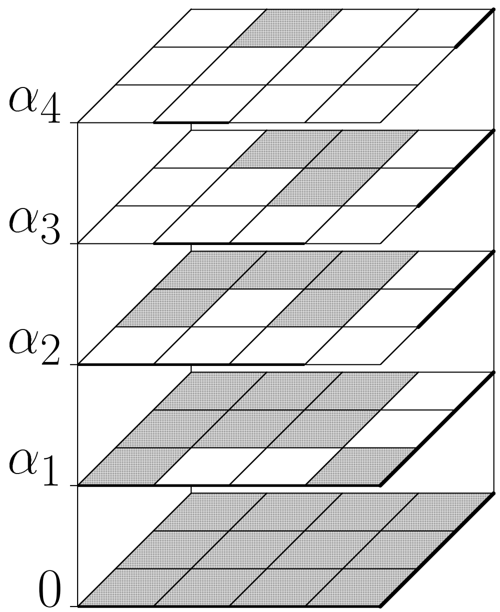
$$\begin{aligned}
S_{\text{gain}}(A, B) \;=\; \int_0^{\sup \Pi} & \log_2 \Big(\sum_{a \in \text{dom}(A)} [\Pi]_\alpha(A = a) \Big) \\
+ \; & \log_2 \Big(\sum_{b \in \text{dom}(B)} [\Pi]_\alpha(B = b) \Big) \\
- \; & \log_2 \Big(\sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} [\Pi]_\alpha(A = a, B = b) \Big) \, \mathrm{d}\alpha
\end{aligned}$$

is called the **specificity gain** of A and B w.r.t. Π.

- Generalization of Hartley information gain on the basis of the α-cut view of possibility distributions.
- Analogous to Shannon information gain.

Idea of Specificity Gain





$$\log_2 1 + \log_2 1 - \log_2 1 = 0$$

$$\log_2 2 + \log_2 2 - \log_2 3 \approx 0.42$$

$$\log_2 3 + \log_2 2 - \log_2 5 \approx 0.26$$

$$\log_2 4 + \log_2 3 - \log_2 8 \approx 0.58$$

$$\log_2 4 + \log_2 3 - \log_2 12 = 0$$

- Exploiting again the α-cut view of possibility distributions:
  Aggregate the Hartley information gain for the different α-cuts.

Specificity Gain in the Example

projection to subspace — minimum of marginals — specificity gain

projection to subspace:

	■	▨	□	▒
△	40	80	10	70
□	30	10	70	60
○	80	90	20	10

minimum of marginals:

	■	▨	□	▒
△	80	80	70	70
□	70	70	70	70
○	80	90	70	70

specificity gain: 0.055 bit

projection to subspace:

	s	m	l
△	20	80	70
□	40	70	20
○	90	60	30

minimum of marginals:

	s	m	l
△	70	70	70
□	80	70	80
○	90	70	80

specificity gain: 0.048 bit

projection to subspace:

	■	▨	□	▒
large	40	70	20	70
medium	60	80	70	70
small	80	90	40	40

minimum of marginals:

	■	▨	□	▒
large	70	70	70	70
medium	80	80	70	70
small	80	90	70	70

specificity gain: 0.027 bit

Evaluation Measures / Scoring Functions

Probabilistic Graphical Models

- Mutual Information / Cross Entropy / Information Gain
- (Symmetric) Information Gain Ratio
- χ^2-Measure
- (Symmetric/Modified) Gini Index
- Bayesian Measures (K2 metric, BDeu metric)
- Measures based on the Minimum Description Length Principle

Possibilistic Graphical Models

- Specificity Gain [Gebhardt and Kruse 1996, Borgelt *et al.* 1996]
- (Symmetric) Specificity Gain Ratio [Borgelt *et al.* 1996]
- Analog of Mutual Information [Borgelt and Kruse 1997]
- Analog of the χ^2-measure [Borgelt and Kruse 1997]

Two Search Methods

- **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.

Experimental Results: Danish Jersey Cattle Data

method	type	edges	params.	min.	avg.	max.
none	independent	0	80	10.064	10.160	11.390
	original	22	308	9.888	9.917	11.318
o.w.s.t.	S_{gain}	20	438	8.878	8.990	10.714
	S_{sgr1}	20	442	8.716	8.916	10.680
	d_{χ^2}	20	472	8.662	8.820	10.334
	d_{mi}	20	404	8.466	8.598	10.386
greedy	S_{gain}	31	1630	8.524	8.621	10.292
	S_{gr}	18	196	9.390	9.553	11.100
	S_{sgr1}	28	496	8.946	9.057	10.740
	d_{χ^2}	35	1486	8.154	8.329	10.200
	d_{mi}	33	774	8.206	8.344	10.416
sim. ann.	w/o penalty	22.6	787.2	8.013	8.291	9.981
	with penalty	20.6	419.1	8.211	8.488	10.133

Summary

- Possibilistic networks can be seen as "fuzzyfications" of relational networks.
- Possibilistic networks are analogous to probabilistic networks:
  - probabilistic networks: sum/product decomposition
  - possibilistic networks: maximum/minimum decomposition
- Reasoning in possibilistic networks aims at finding a full description of the actual state of the world.
- Possibilistic networks can be learned from a database of sample cases.
- Quantitative/parameter learning is more difficult for possibilistic networks.
- Qualitative/structure learning is similar for probabilistic/possibilistic networks:
  - heuristic search methods are necessary
  - learning algorithms consist of a search method and an evaluation measure