Bayesian Networks

Prof. Dr. Rudolf Kruse,
Matthias Steinbrecher

Computational Intelligence Group
Department of Knowledge Processing and Language Engineering
Faculty of Computer Science
`kruse@iws.cs.uni-magdeburg.de`

Organisational

- Lecture
  - Consultation: Wednesday, 11:00 a. m.– noon, G29-008
  - Preferredly reachable by e-mail: `kruse@iws.cs.uni-magdeburg.de`
- Excercises
  - Tutor: Matthias Steinbrecher, at all hours
  - G29-015, `msteinbr@iws.cs.uni-magdeburg.de`
- Updated information on the course:
  - `http://fuzzy.cs.uni-magdeburg.de/`

- **Human Expert**

  A human *expert* is a specialist for a specific differentiated application field who creates solutions to customer problems in this respective field and supports them by applying these solutions.

- **Requirements**
  - Formulate precise problem scenarios from customer inquiries
  - Find correct and complete solution
  - Understandable answers
  - Explanation of solution
  - Support the deployment of solution

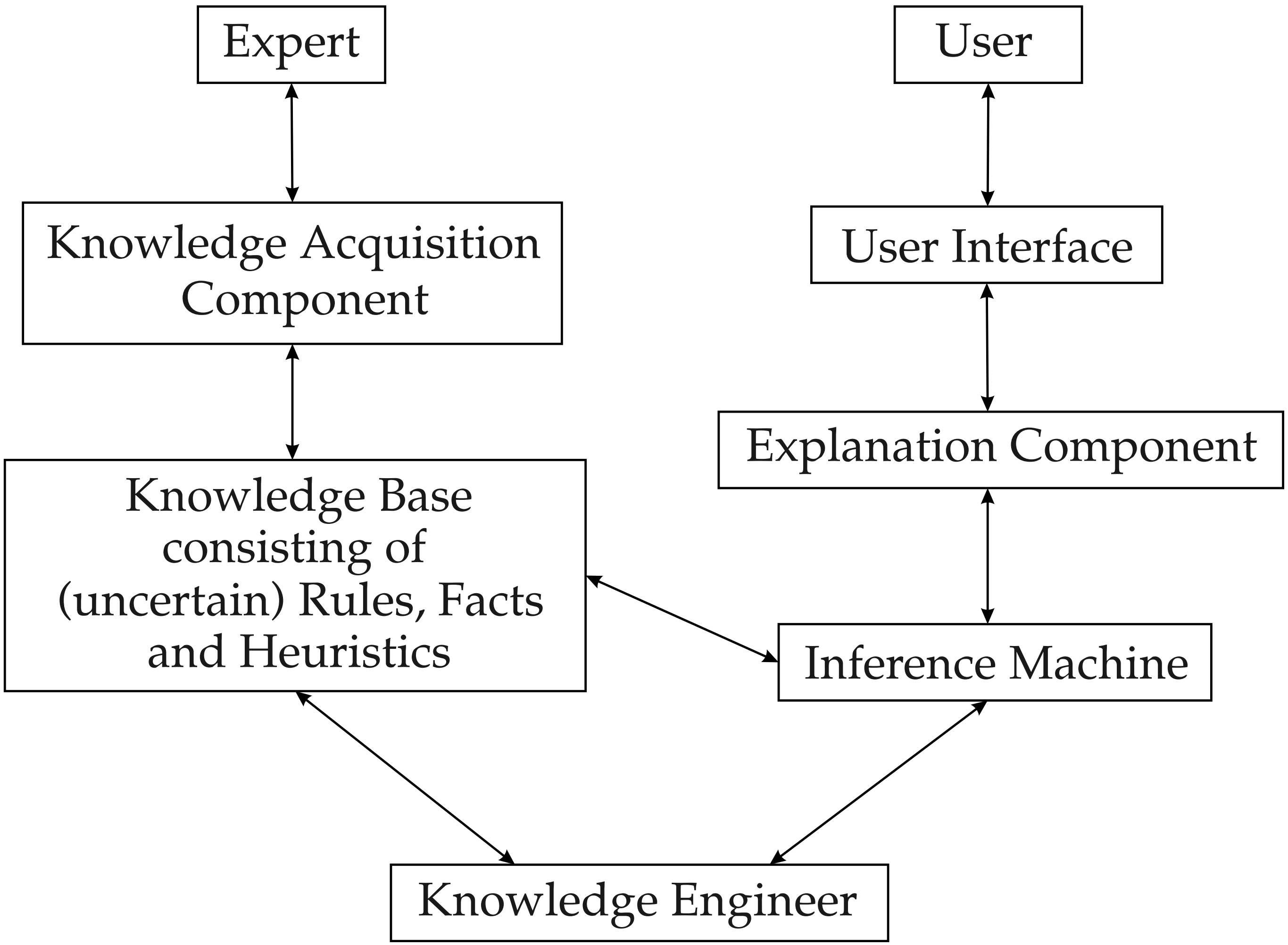
Knowledge Based Systems (2)

- **"Intelligent" System**

  An intelligent system is a program that models the knowledge and inference methods of a human expert of a specific field of application.

- **Requirements for construction:**
  - Knowledge Representation
  - Knowledge Acquisition
  - Knowledge Modification

Expert System Architecture



Qualities of Knowledge

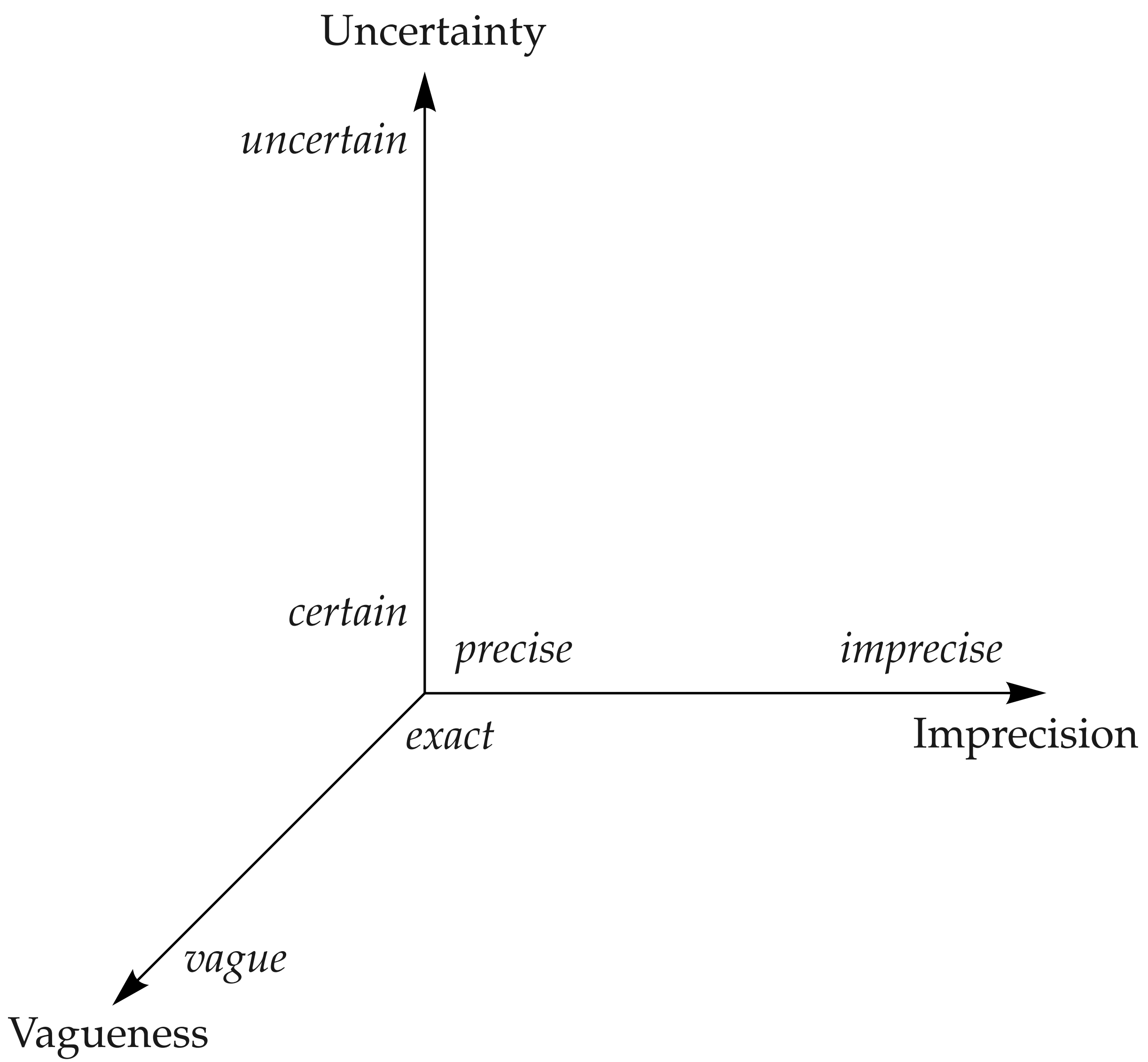
In most cases our knowledge about the present world is

- **imprecise/missing** (knowledge is not comprehensive)
  - e. g. "I don't know the bus departure times for public holidays because I only take the bus on working days."
- **vague/fuzzy** (knowledge is not exact)
  - e. g. "The bus departs roughly every full hour."
- **uncertain** (knowledge is unreliable)
  - e. g. "The bus departs probably at 12 o'clock."

We have to decide nonetheless!

- Reasoning under Vagueness
- Reasoning with Probabilities
- ...and Cost/Benefit

Knowledge Characteristics



Example

Objective: *Be at the university at 9:15 to attend a lecture.*

- There are several plans to reach this goal:
  - P_1: Get up at 8:00, leave at 8:55, take the bus at 9:00 ...
  - P_2: Get up at 7:30, leave at 8:25, take the bus at 8:30 ...
  - ...
- All plans are *correct*, but
  - they imply different *costs* and different *probabilities* to *actually* reach that goal.
  - P_2 would be the plan of choice as the lecture is important and the success rate of P_1 is only about 80–95%.
- Question: *Is a computer capable of solving these problems involving uncertainty?*

Uncertainty and Rules (1)

- Example: We are given a simple expert system for dentists that may contain the following rule:

$$\forall p : [\text{Symptom}(p, \mathsf{toothache}) \Rightarrow \text{Disease}(p, \mathsf{cavity})]$$

- This rule is *incorrect*! Better:

$$\forall p : \Big[\text{Symptom}(p, \mathsf{toothache}) \Rightarrow \\ \text{Disease}(p, \mathsf{cavity}) \lor \text{Disease}(p, \mathsf{gumdisease}) \lor \ldots \Big]$$

- Maybe take the *causal* rule?

$$\forall p : \Big[\text{Disease}(p, \mathsf{cavity}) \Rightarrow \text{Symptom}(p, \mathsf{toothache})\Big]$$

- Incorrect, too.

Uncertainty and Rules (2)

Problems with propositional logic:

- We cannot enumerate all possible causes, even though ...
- We do not know the (medical) cause-effect interactions, and even though ...
- Uncertainty about the patient remains:
  - Caries and toothache may co-occurr by chance.
  - Were (exhaustively) all examinations conducted?
    — If yes: correctly?
  - Did the patient answer all questions?
    — If yes: appropriately?
- Without perfect knowledge no correct logical rules!

Uncertainty and Facts

Example:

- We would like to support a robot's localization by fixed landmarks.
  From the presence of a landmark we may infer the location.

Problem:

- Sensors are imprecise!
  - We cannot conclude definitely a location simply because
    there was a landmark detected by the sensors.
  - The same holds true for undetected landmarks.
  - Only probabilities are being increased or decreased.

Degrees of Belief

- We (or other agents) are only believing facts or rules to some extent.
- One possibility to express this *partial belief* is by using *probability theory*.
- "The agent believes the sensor information to 0.9" means:
  In 9 out of 10 cases the agent trusts in the correctness of the sensor output.
- Probabilities gather the "uncertainty" that originates due to ignorance.
- Probabilities $\neq$ Vagueness/Fuzziness!
  - The predicate "large" is fuzzy whereas "This might be Peter's watch." is uncertain.

Rational Decisions under Uncertainty

- Choice of several *actions* or *plans*
- These may lead to different results with different *probabilities*.
- The *actions* cause different (possibly subjective) *costs*.
- The *results* yield different (possibly subjective) *benefits*.
- It would be rational to choose that action that yields the largest total benefit.

$$\text{Decision Theory} = \text{Utility Theory} + \text{Probability Theory}$$

Decision-theoretic Agent

input perception
output action

- $K \leftarrow$ a set of probabilistic beliefs about the state of the world
- calculate updated probabilities for current state based on available evidence including current percept and previous action
- calculate outcome probabilities for actions, given action descriptions and probabilities of current states
- select action A with highest expected utility given probabilities of outcomes and utility information
- **return** A

Decision Theory: An agent is rational if and only if it chooses the action yielding the largest utility averaged over all possible outcomes of all actions.

Rule-based Expert Systems

Rule-based Expert Systems

Modi of usage:

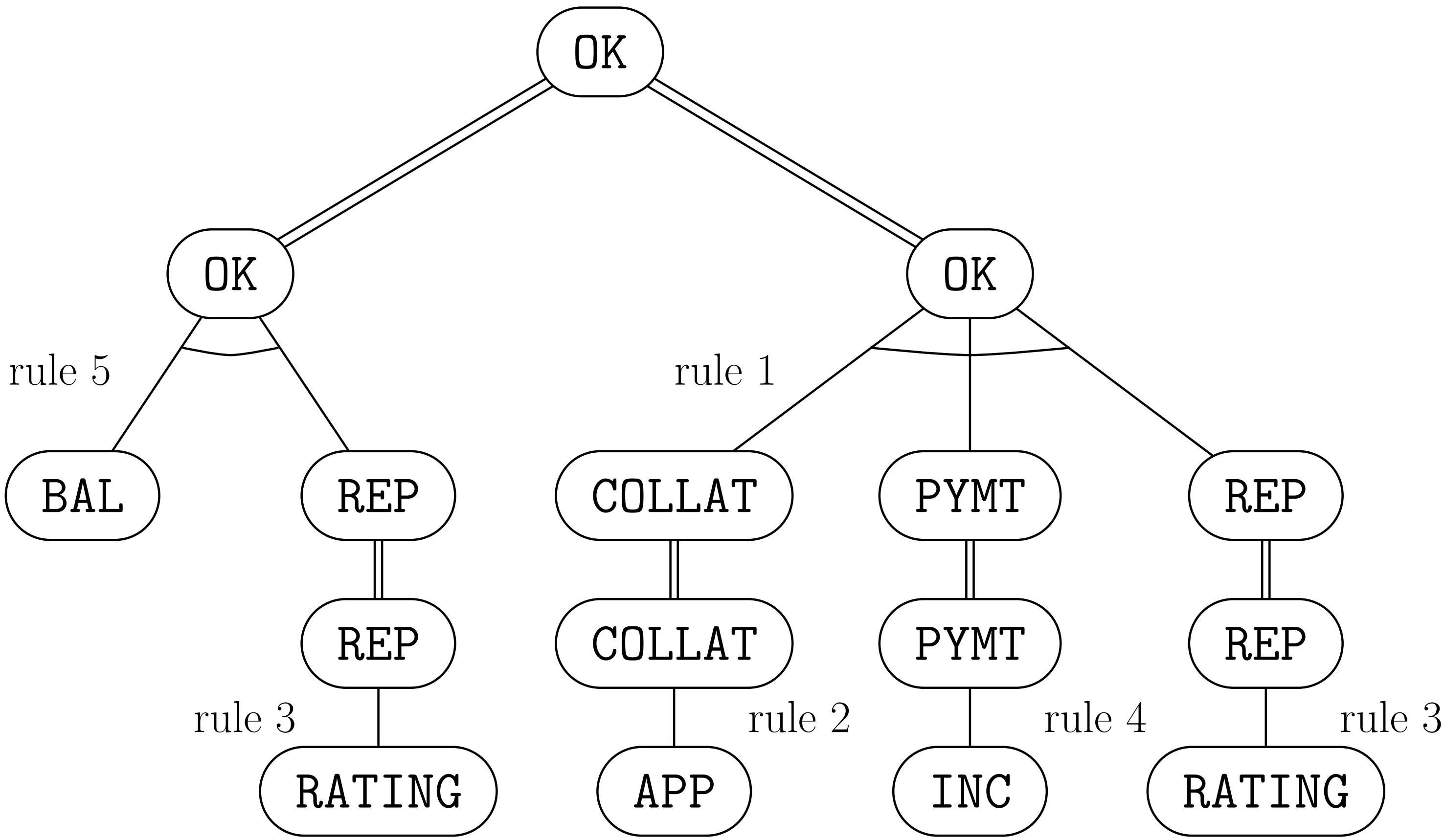
- **Query:** Facts are retrieved from database or user is interrogated
- **Explanation:** System answers questions how a decision was concluded

Example rule base:

1. $\texttt{COLLAT} \wedge \texttt{PYMT} \wedge \texttt{REP} \Rightarrow \texttt{OK}$
2. $\texttt{APP} \Rightarrow \texttt{COLLAT}$
3. $\texttt{RATING} \Rightarrow \texttt{REP}$
4. $\texttt{INC} \Rightarrow \texttt{PYMT}$
5. $\texttt{BAL} \wedge \texttt{REP} \Rightarrow \texttt{OK}$

Decision Rule Tree

Corresponding **and/or**-Tree:



Example Queries

- **User:** Why do you believe that the loan should be approved?
- **System:** Because the collateral for the loan is satisfactory, and the applicant is able to make the loan payments, and the applicant has a good financial reputation.
- **User:** Why did you ask me about whether or not the applicant's income exceeds his/her expenses?
- **System:** Because I was trying to establish whether or not the applicant is able to make the loan payments.
- **User:** Why do you believe that the applicant has a good financial reputation?
- **System:** Because the credit rating service said that the applicant has a good credit rating.

Rules

- A rule in general is a *if-then*-construct consisting of a *condition* and an *action*.

$$\boxed{\text{If } \textit{condition} \text{ then } \textit{conclusion}}$$

- These two parts may be interpreted differently according to the context:
  - **Inference rules**: If *premise* then *conclusion*
  - **Hypotheses:** If *evidence* then *hypothesis*
  - **Productions:** If *condition* then *action*

- Rules are often referred to as *productions* or *production rules*.

Rules

- A rule in the ideal case represents a unit of knowledge.
- A set of rules together with an execution/evaluation strategy comprises a program to find solutions to specific problem classes.
- Prolog program: rule-based system
- Rule-based systems are historically the first types of AI systems and were for a long time considered prototypical expert systems.
- Nowadays, not every expert systems uses rules as its core inference mechanism.
- Rising importance in the field of business process rules.

Rule Evaluation

Forward chaining

- Expansion of knowledge base: as soon as new facts are inserted the system also calculates the conclusions/consequences.
- Data-driven behavior
- Premises-oriented reasoning: the chaining is determined by the left parts of the rules.

Backward chaining

- Answering queries
- Demand-driven behavior
- Conclusion-oriented reasoning: the chaining is determined by the right parts of the rules.

Components of a Rules-based System

Data base

- Set of structured data objects
- Current state of modeled part of world

Rule base

- Set of rules
- Application of a rule will alter the data base

Rule interpreter

- Inference machine
- Controls the program flow of the system

Rule Interpretation

- Main scheme forward chaining
  - Select and apply rules from the set of rules with valid antecedences. This will lead to a modified data base and the possibility to apply further rules.
- Run this cycle as long as possible.
- The process terminates, if
  - there is no rule left with valid antecendence
  - a solution criterion is satisfied
  - a stop criterion is satisfied (e. g. maximum number of steps)
- Following tasks have to be solved:
  - Identify those rules with a valid condition
    $\Rightarrow$ **Instantiation** or **Matching**
  - Select rules to be executed
    $\Rightarrow$ need for **conflict resolution**
    (e. g. via partial or total orderings on the rules)

Certainty Factors

Mycin (1970)

- **Objective:** Development of a system that supports physicians in diagnosing bacterial infections and suggesting antibiotics.
- **Features:** Uncertain knowledge was represented and processed via *uncertainty factors*.
- **Expert Knowlegde:** 500 (uncertain) decision rules as static knowledge base.
- **Case-specific knowledge:**
  - static: patients' data
  - dynamic: intermediate results (facts)
- **Strengths:**
  - diagnosis-oriented interrogation
  - hypotheses generation
  - finding notification
  - therapy recommendation
  - explanation of inference path

Uncertainty Factors

- Uncertainty factor $\quad \mathrm{CF} \in [-1, 1] \quad \approx \quad$ degree of belief.

- Rules:

$$\mathrm{CF}(A \to B) \begin{cases} = 1 & B \text{ is certainly true given } A \\ > 0 & A \text{ supports } B \\ = 0 & A \text{ has no influence on } B \\ < 0 & A \text{ provides evidence against } B \\ = -1 & B \text{ is certainly false given } A \end{cases}$$

A Mycin Rule

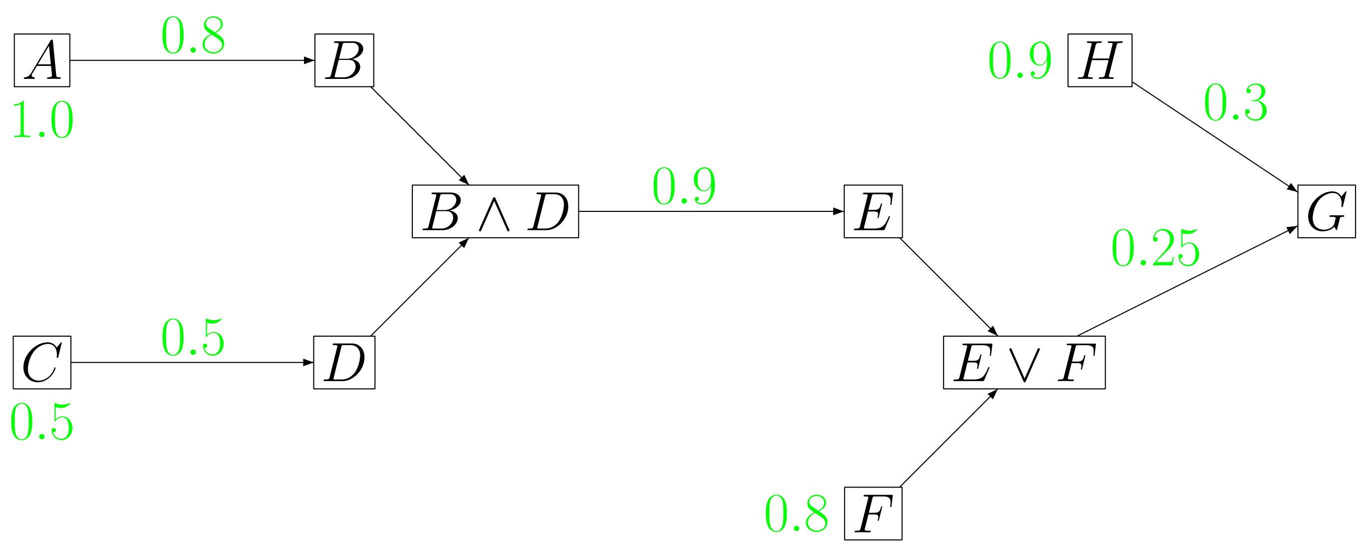
```
RULE035

PREMISE:   ($AND    (SAME CNTXT GRAM GRAMNEG)
                    (SAME CNTXT MORPH ROD)
                    (SAME CNTXT AIR ANAEROBIC))
ACTION:    (CONCL.CNTXT IDENTITY BACTEROIDES TALLY .6)
```

If	1) the *gram stain* of the organism is *gramneg*, and 2) the *morphology* of the organism is *rod*, and 3) the *aerobicity* of the organism is *anaerobic*
then	there is suggestive evidence (0.6) that the *identity* of the organism is *bacteroides*

Example

$$\begin{aligned} A &\rightarrow B \,[0.80] \\ C &\rightarrow D \,[0.50] \\ B \wedge D &\rightarrow E \,[0.90] \\ E \vee F &\rightarrow G \,[0.25] \\ H &\rightarrow G \,[0.30] \end{aligned} \qquad\qquad \begin{aligned} A \,[1.00] \\ C \,[0.50] \\ F \,[0.80] \\ H \,[0.90] \end{aligned}$$



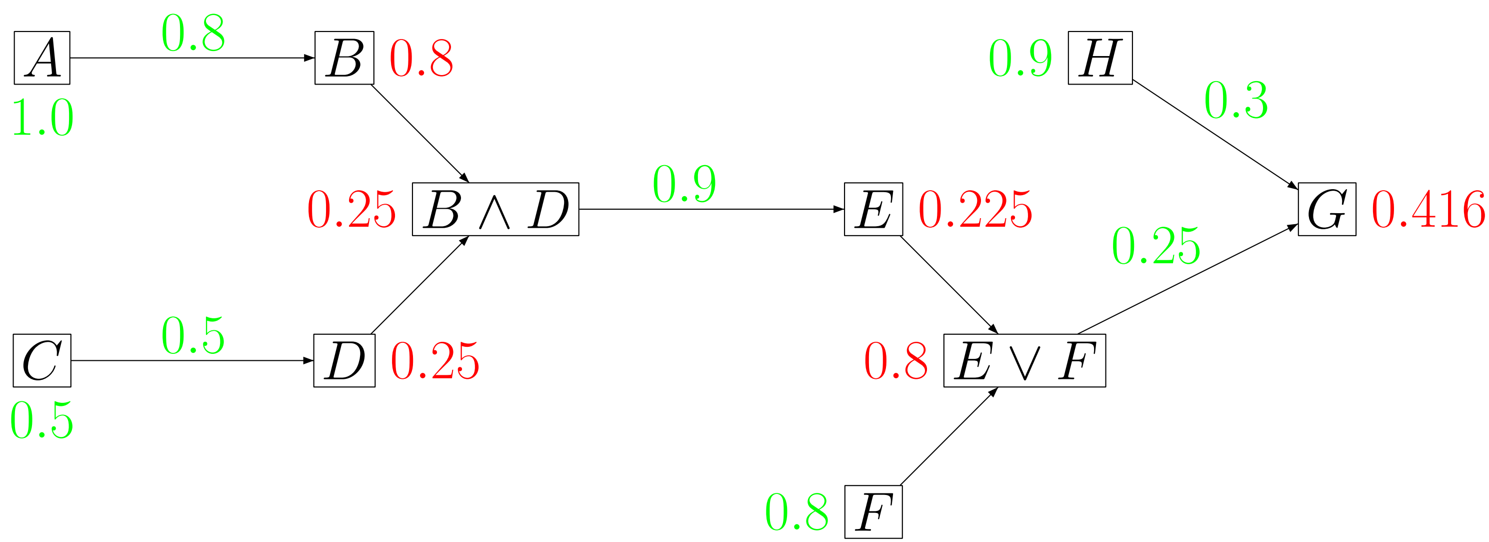
Propagation Rules

- **Conjunction:** $\mathrm{CF}(A \wedge B) = \min\{\mathrm{CF}(A), \mathrm{CF}(B)\}$
- **Disjunction:** $\mathrm{CF}(A \vee B) = \max\{\mathrm{CF}(A), \mathrm{CF}(B)\}$
- **Serial Combination:** $\mathrm{CF}(B, \{A\}) = \mathrm{CF}(A \rightarrow B) \cdot \max\{0, \mathrm{CF}(A)\}$
- **Parallel Combination:** for $n > 1$:
  $$\mathrm{CF}(B, \{A_1, \ldots, A_n\}) = f(\mathrm{CF}(B, \{A_1, \ldots, A_{n-1}\}), \mathrm{CF}(B, \{A_n\}))$$

with

$$f(x, y) = \begin{cases} x + y - xy & \text{if} \quad x, y > 0 \\ x + y + xy & \text{if} \quad x, y < 0 \\ \dfrac{x + y}{1 - \min\{|x|, |y|\}} & \text{otherwise} \end{cases}$$

Example (cont.)



$$f(0.3 \cdot 0.9, 0.25 \cdot 0.8) \;=\; 0.27 + 0.2 - 0.27 \cdot 0.2 = 0.416$$

Was Mycin a failure?

It can be shown that the rule combination scheme is inconsistent in general.
It worked in the Mycin case because the rules had tree-like structure.

Mycin was never used for its intented purpose, because

- physicians were distrustful and not willing to accept Mycin's recommendations.
- Mycin was too good.

However,

- Mycin was a milestone for the development of expert systems.
- it gave rise to impulses for expert system development in general.

Probabilistic Rules

How to assign probabilities to rules (implications)?

$$\boxed{P(B \mid A) \;\leq\; P(A \rightarrow B) = P(\neg A \vee B)}$$

A	B	$P(\cdot)$
0	0	0.04
0	1	0.95
1	0	0.01
1	1	0

$P(B \mid A) = 0$, but $P(A \rightarrow B) = 0.99$!

In the following, probabilistic rules are evaluated with conditional probabilities.

Elements of Graph Theory

Simple Graph

A simple graph (or just: graph) is a tuple $\mathcal{G} = (V, E)$ where

$$V = \{A_1, \ldots, A_n\}$$

represents a finite set of **vertices** (or **nodes**) and

$$E \subseteq (V \times V) \setminus \{(A, A) \mid A \in V\}$$

denotes the set of **edges**.
It is called simple since there are no self-loops and no multiple edges.

Edge Types

Let $\mathcal{G} = (V, E)$ be a graph. An edge $e = (A, B)$ is called

- **directed** if $(A, B) \in E \;\Rightarrow\; (B, A) \notin E$
  Notation: $A \rightarrow B$
- **undirected** if $(A, B) \in E \;\Rightarrow\; (B, A) \in E$
  Notation: $A - B$ or $B - A$

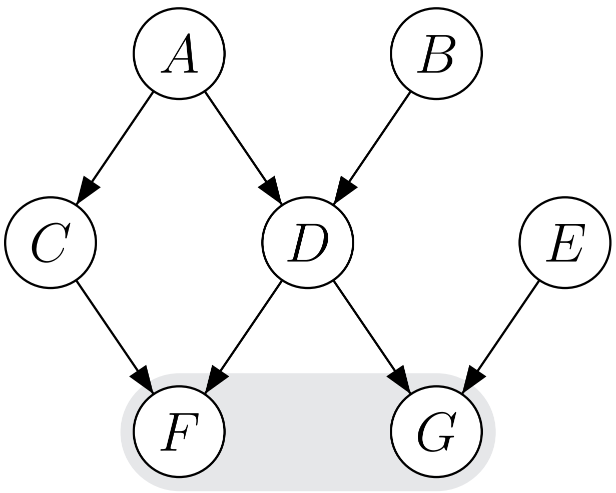
(Un)directed Graph

A graph with only (un)directed edges is called an (un)directed graph.

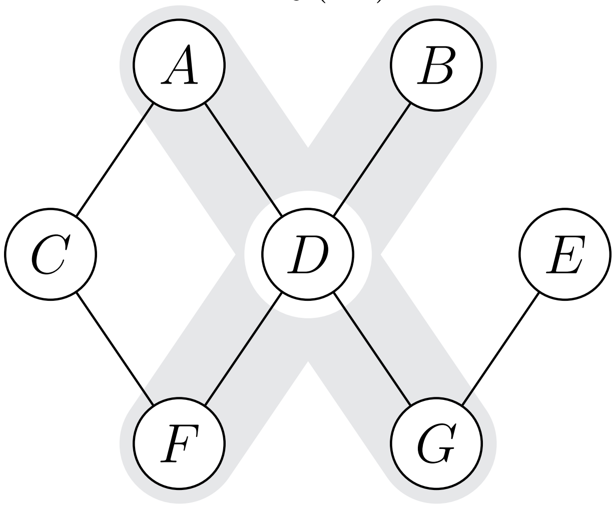
Adjacency Set

Let $\mathcal{G} = (V, E)$ be a graph. The set of nodes that is accessible via a given node $A \in V$ is called the **adjacency set** of A:

$$\mathrm{adj}(A) \;=\; \{B \in V \mid (A, B) \in E\}$$



adj(D)



Paths

Let $\mathcal{G} = (V, E)$ be a graph. A series ρ of r pairwise different nodes

$$\rho = \left\langle A_{i_1}, \ldots, A_{i_r} \right\rangle$$

is called a **path** from A_i to A_j if

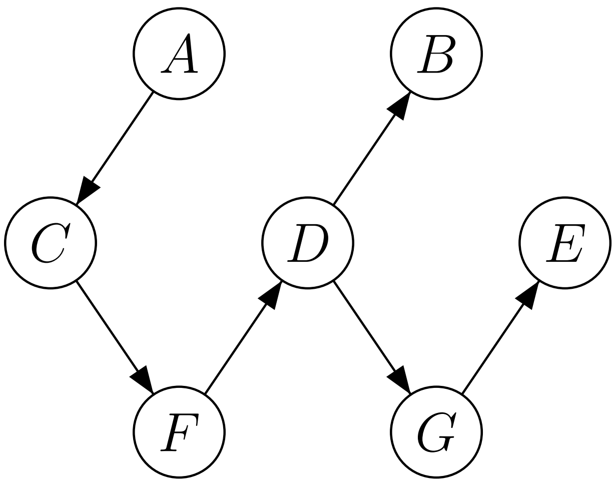
- $A_{i_1} = A_i, \quad A_{i_r} = A_j$
- $A_{i_{k+1}} \in \mathrm{adj}(A_{i_k}), \quad 1 \leq k < r$

A path with only undirected edges is called an **undirected path**

$$\rho = A_{i_1} - \cdots - A_{i_r}$$

whereas a path with only directed edges is referred to as a **directed path**

$$\rho = A_{i_1} \to \cdots \to A_{i_r}$$



If there is a directed path ρ from node A to node B in a directed graph $\mathcal{G}$ we write

$$A \overset{\rho}{\underset{\mathcal{G}}{\rightsquigarrow}} B.$$

If the path ρ is undirected we denote this with

$$A \overset{\rho}{\underset{\mathcal{G}}{\leftrightsquigarrow}} B.$$

Graph Types

Loop

Let $\mathcal{G} = (V, E)$ be an undirected graph. A path

$$\rho = X_1 - \cdots - X_k$$

with $X_k - X_1 \in E$ is called a loop.

Cycle

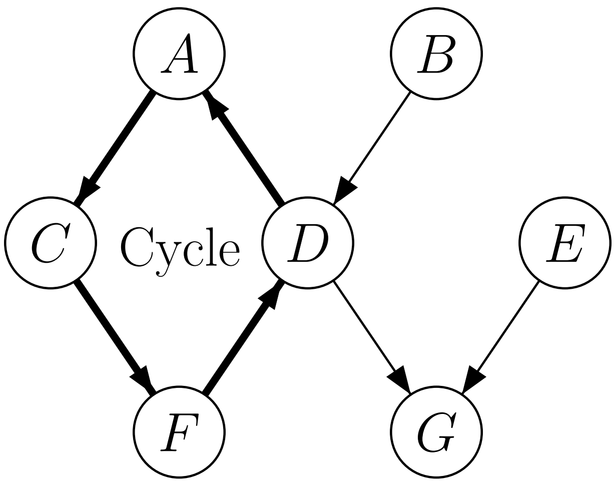
Let $\mathcal{G} = (V, E)$ be a directed graph. A path

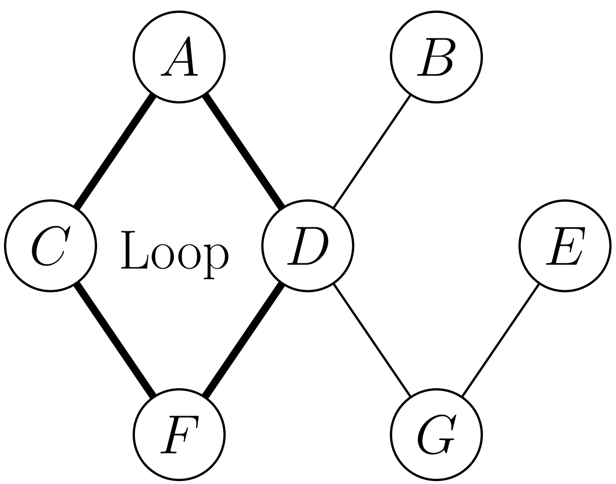
$$\rho = X_1 \to \cdots \to X_k$$

with $X_k \to X_1 \in E$ is called a cycle.

Directed Acyclic Graph (DAG)

A directed graph $\mathcal{G} = (V, E)$ is called **acyclic** if for every path $X_1 \to \cdots \to X_k$ in $\mathcal{G}$ the condition $X_k \to X_1 \notin E$ is satisfied, i. e. it contains no cycle.





Parents, Children and Families

Let $\mathcal{G} = (V, E)$ be a directed graph. For every node $A \in V$ we define the following sets:

- **Parents:**

  $$\mathrm{parents}_{\mathcal{G}}(A) = \{B \in V \mid B \to A \in E\}$$

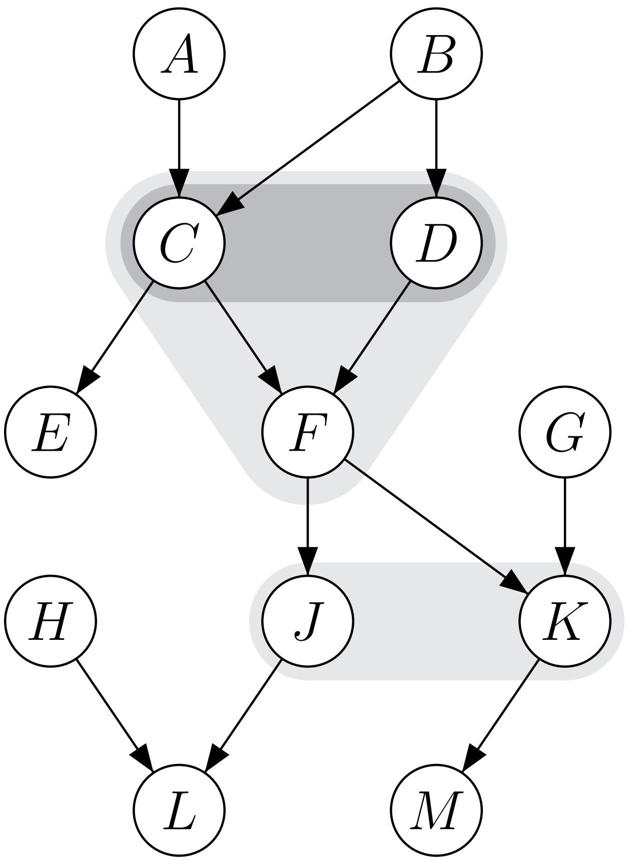
- **Children:**

  $$\mathrm{children}_{\mathcal{G}}(A) = \{B \in V \mid A \to B \in E\}$$

- **Family:**

  $$\mathrm{family}_{\mathcal{G}}(A) = \{A\} \cup \mathrm{parents}_{\mathcal{G}}(A)$$

If the respective graph is clear from the context, the index $\mathcal{G}$ is omitted.



$$\begin{aligned}
\mathrm{parents}(F) &= \{C, D\} \\
\mathrm{children}(F) &= \{J, K\} \\
\mathrm{family}(F) &= \{C, D, F\}
\end{aligned}$$

Ancestors, Descendants, Non-Descendants

Let $\mathcal{G} = (V, E)$ be a DAG. For every node $A \in V$ we define the following sets:

- **Ancestors:**

  $$\mathrm{ancs}_{\mathcal{G}}(A) \;=\; \{B \in V \mid \exists \rho : B \overset{\rho}{\underset{\mathcal{G}}{\rightsquigarrow}} A\}$$

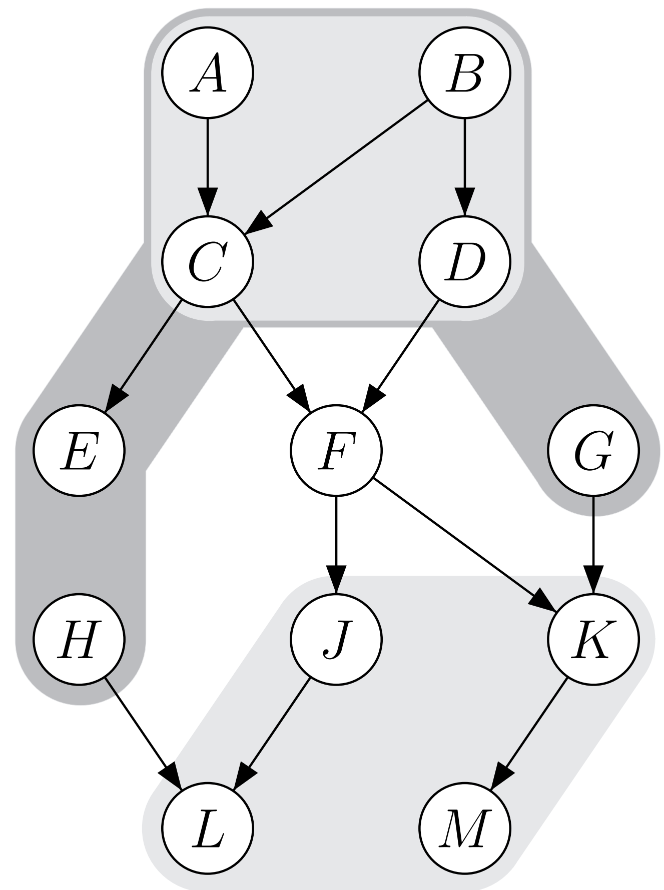
- **Descendants:**

  $$\mathrm{descs}_{\mathcal{G}}(A) \;=\; \{B \in V \mid \exists \rho : A \overset{\rho}{\underset{\mathcal{G}}{\rightsquigarrow}} B\}$$

- **Non-Descendants:**

  $$\text{non-descs}_{\mathcal{G}}(A) \;=\; V \setminus \{A\} \setminus \mathrm{descs}_{\mathcal{G}}(A)$$

If the respective graph is clear from the context, the index $\mathcal{G}$ is omitted.



$$\mathrm{ancs}(F) = \{A, B, C, D\}$$
$$\mathrm{descs}(F) = \{J, K, L, M\}$$
$$\text{non-descs}(F) = \{A, B, C, D, E, G, H\}$$

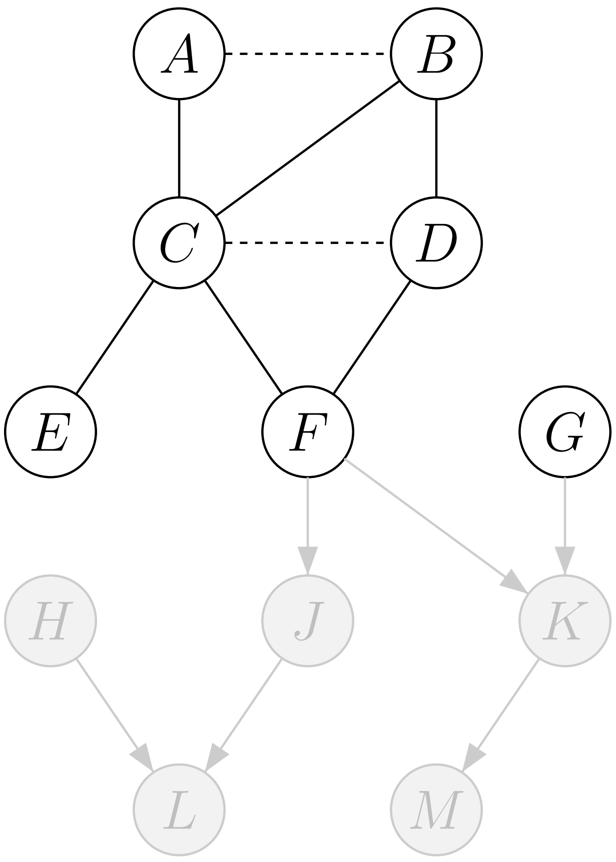
Operations on Graphs

Let $\mathcal{G} = (V, E)$ be a DAG.

The **Minimal Ancestral Subgraph** of $\mathcal{G}$ given a set $M \subseteq V$ of nodes is the smallest subgraph that contains all ancestors of all nodes in M.

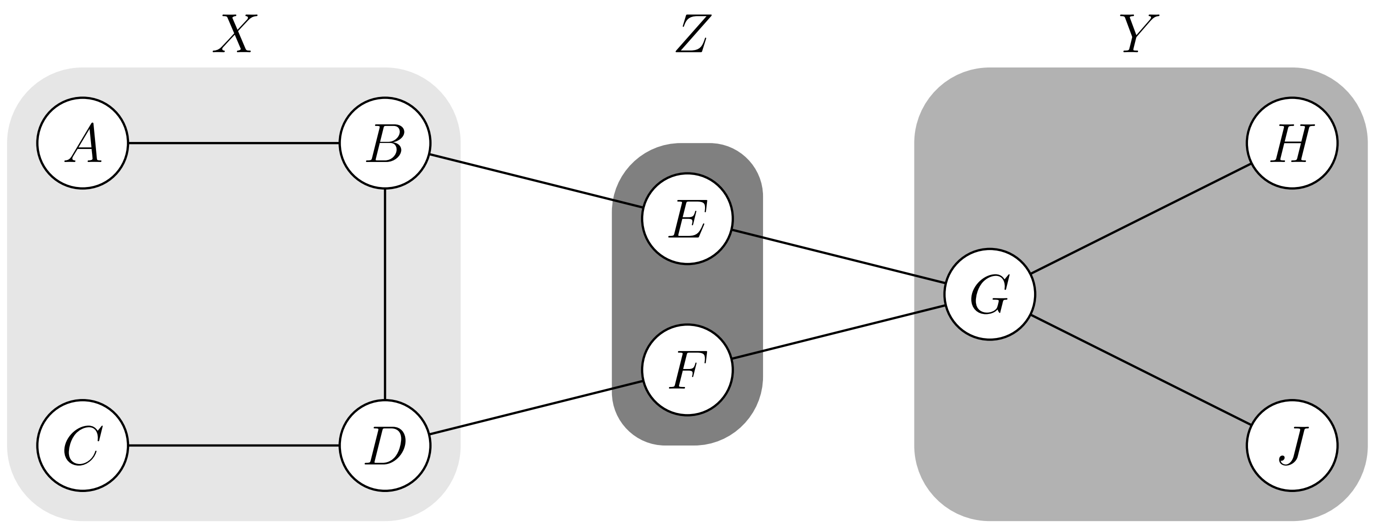
The **Moral Graph** of $\mathcal{G}$ is the undirected graph that is obtained by

1. connecting nodes that share a common child with an arbitrarily directed edge and,
2. converting all directed edges into undirected ones by dropping the arrow heads.



Moral graph of ancestral graph induced by the set $\{E, F, G\}$.

u-Separation



Let $\mathcal{G} = (V, E)$ be an undirected graph and $X, Y, Z \subseteq V$ three disjoint subsets of nodes. We agree on the following separation criteria:

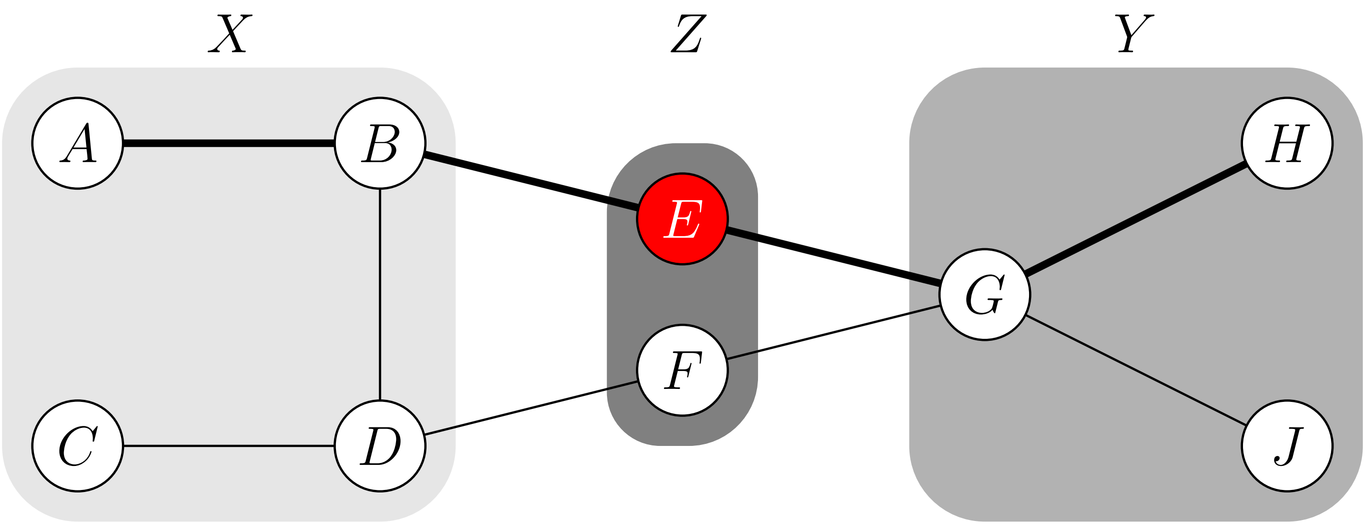
1. Z u-separates X from Y — written as

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z,$$

   if every possible path from a node in X to a node in Y is blocked.

2. A path is blocked if it contains one (or more) **blocking nodes**.
3. A node is a blocking node if it lies in Z.

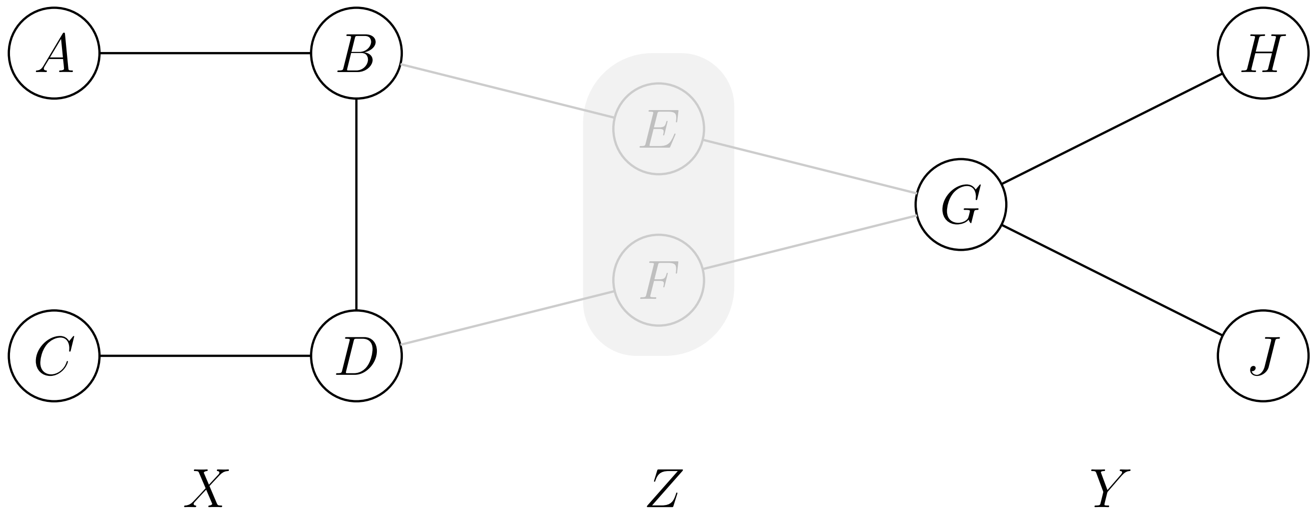
u-Separation



E. g. path $A - B - E - G - H$ is blocked by $E \in Z$. It can be easily verified, that every path from X to Y is blocked by Z. Hence we have:

$$\{A, B, C, D\} \perp\!\!\!\perp_{\mathcal{G}} \{G, H, J\} \mid \{E, F\}$$

u-Separation



Another way to check for u-separation: Remove the nodes in Z from the graph (and all the edges adjacent to these nodes). X and Y are u-separated by Z if the remaining graph is disconnected with X and Y in separate subgraphs.

Now: Separation criterion for directed graphs.

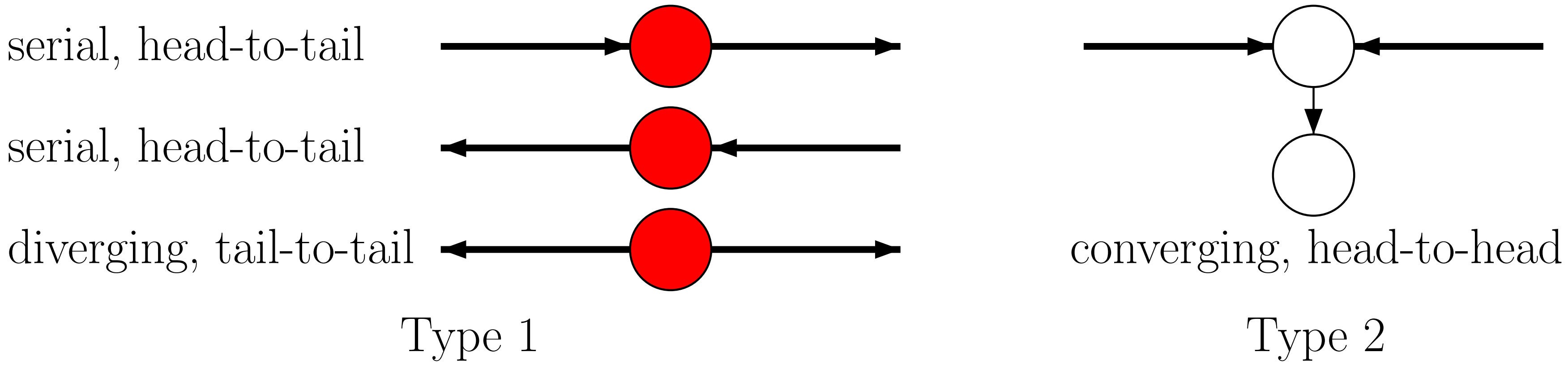
We use the same priciples as for u-separation. Two modifications are necessary:

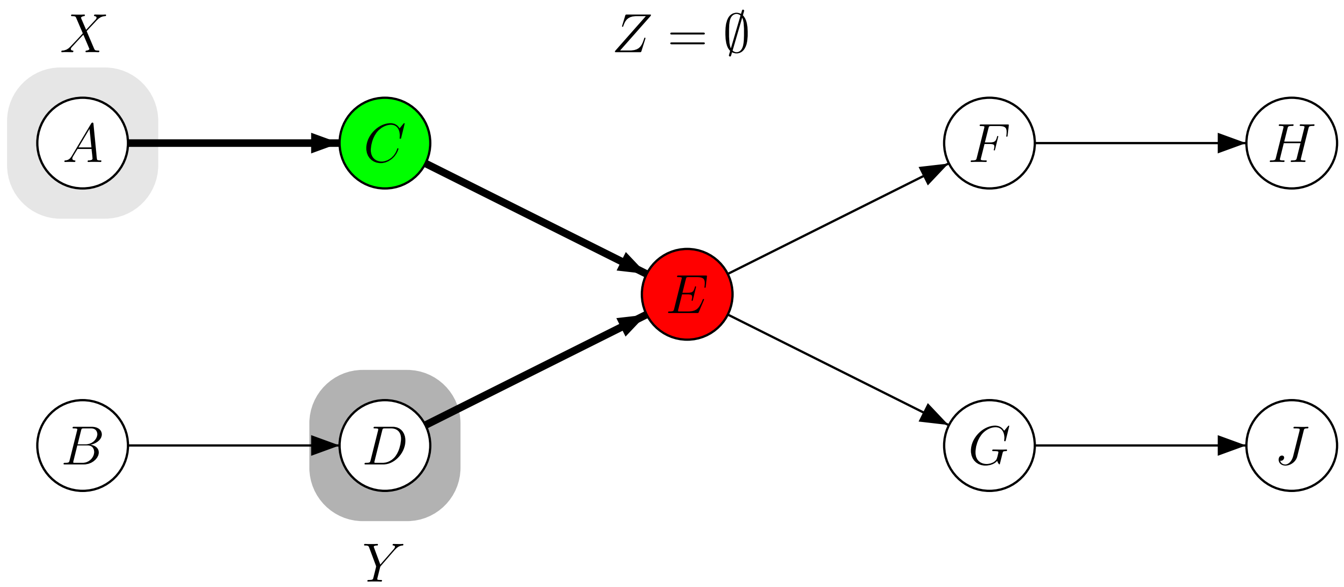
- Directed paths may lead also in reverse to the arrows.
- The blocking node condition is more sophisticated.

Blocking Node (in a directed path)

A node A is blocked if its edge directions **along the path**

- are of type 1 and $A \in Z$, or
- are of type 2 and neither A nor one of its descendants is in Z.



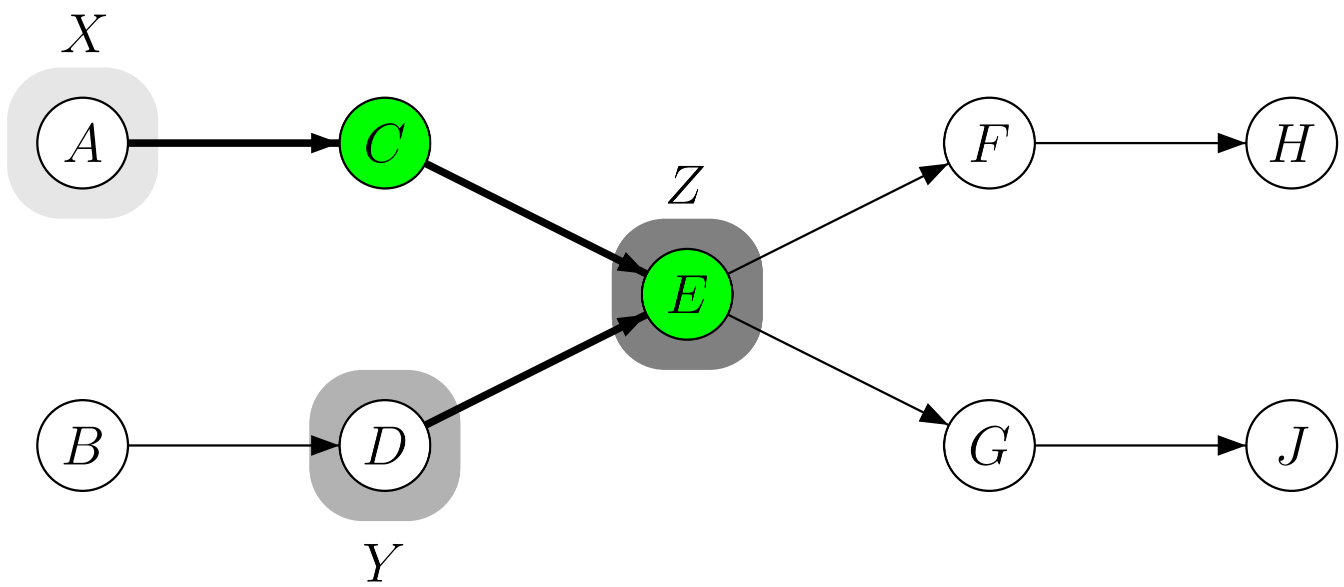


Checking path $A \to C \to E \leftarrow D$:

- C is **serial** and not in Z: non-blocking
- E is **converging** and not in Z, neither is F, G, H or J: **blocking**

$\Rightarrow$ Path is blocked

$$A \perp\!\!\!\perp D \mid \emptyset$$



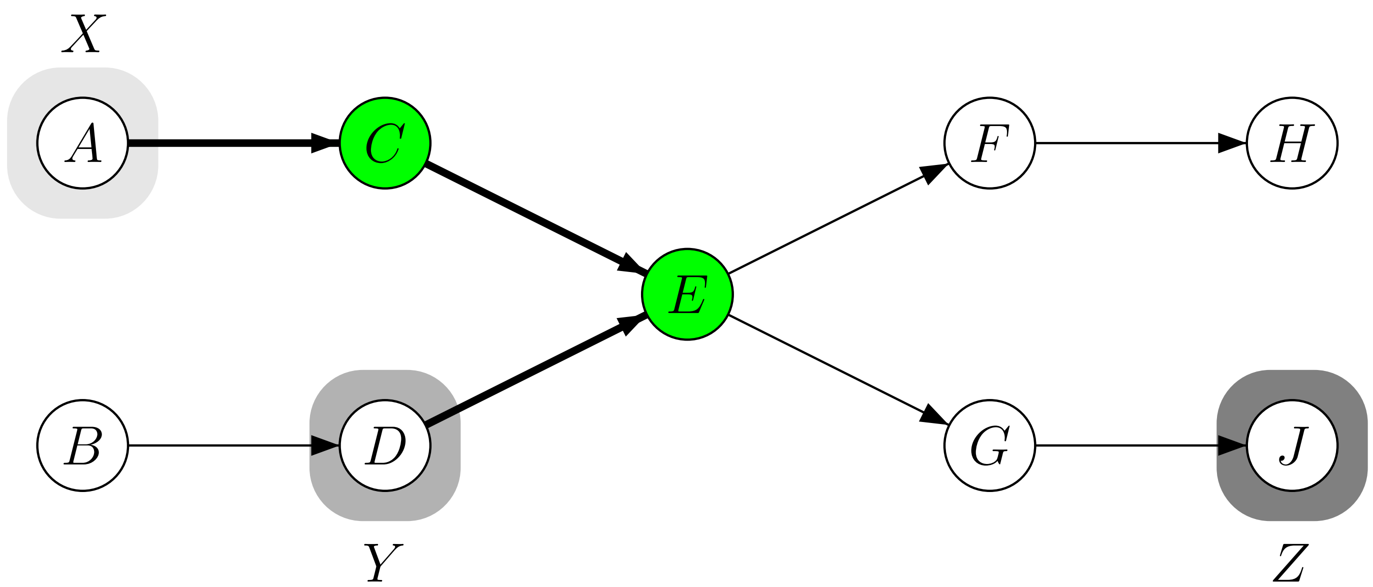
Checking path $A \to C \to E \leftarrow D$:

- C is **serial** and not in Z: non-blocking
- E is **converging** and in Z: non-blocking

$\Rightarrow$ Path is not blocked

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

d-Separation



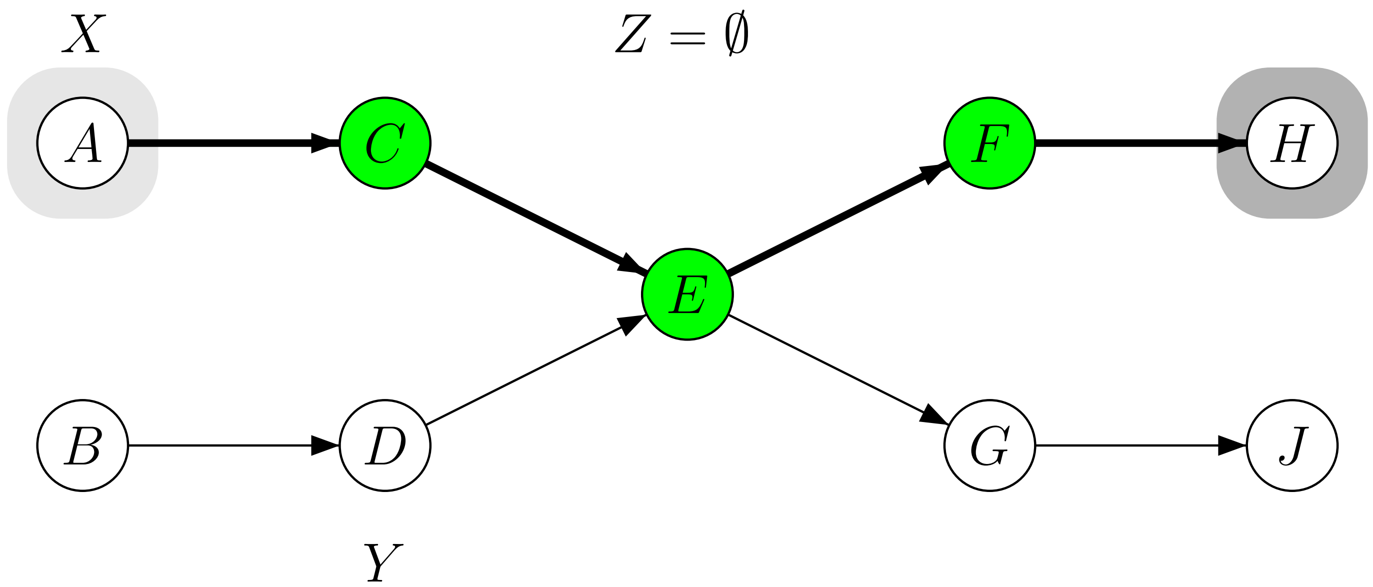
Checking path $A \to C \to E \leftarrow D$:

- C is **serial** and not in Z: non-blocking
- E is **converging** and not in Z but one of its descendants (J) is in Z: non-blocking

$\Rightarrow$ Path is not blocked

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

d-Separation



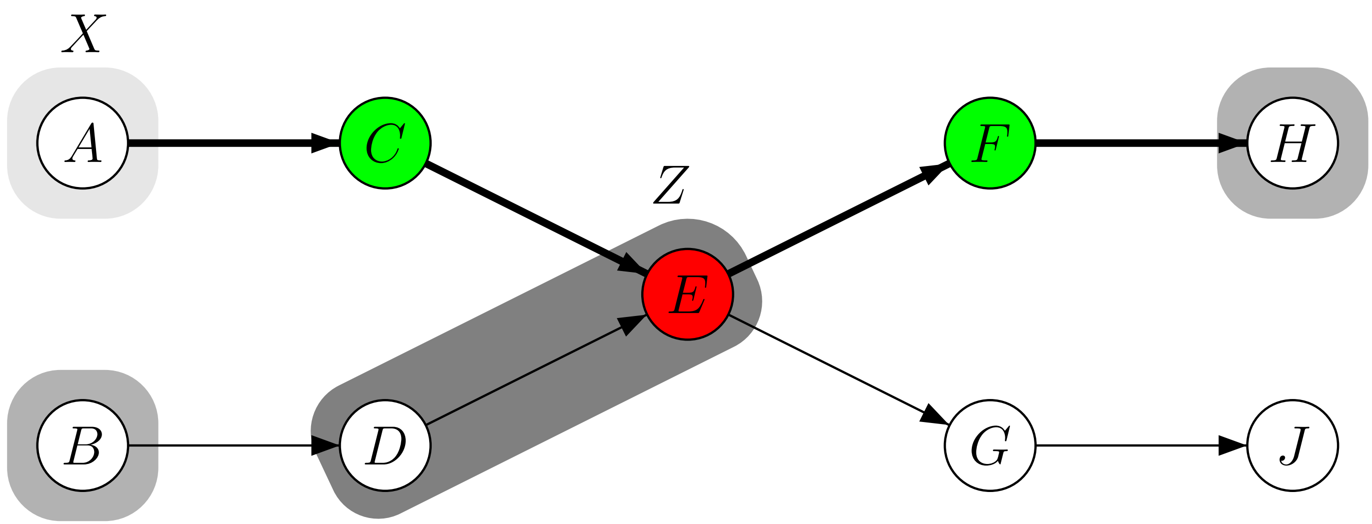
Checking path $A \to C \to E \to F \to H$:

- C is **serial** and not in Z: non-blocking
- E is **serial** and not in Z: non-blocking
- F is **serial** and not in Z: non-blocking

$\Rightarrow$ Path is not blocked

$$A \not\perp\!\!\!\perp H \mid \emptyset$$

d-Separation



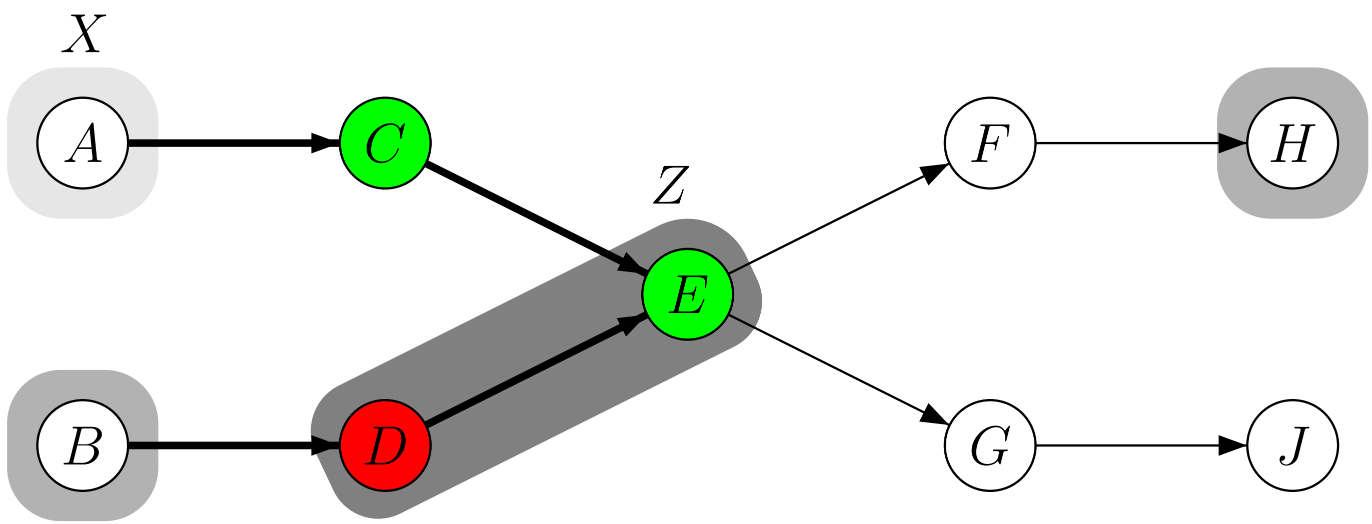
$Y = \{B, H\}$

Checking path $A \to C \to E \to F \to H$:

- C is **serial** and not in Z: non-blocking
- E is **serial** and in Z: **blocking**
- F is **serial** and not in Z: non-blocking

$\Rightarrow$ Path is blocked

d-Separation



$$Y = \{B, H\}$$

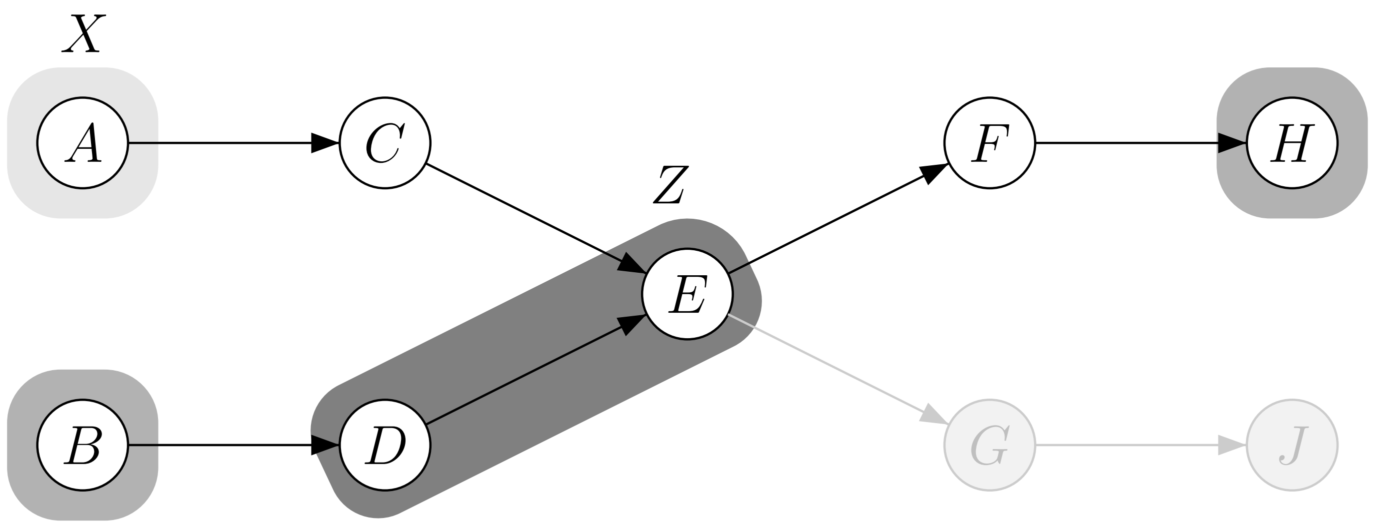
Checking path $A \to C \to E \leftarrow D \to B$:

- C is **serial** and not in Z: non-blocking
- E is **converging** and in Z: non-blocking
- D is **serial** and in Z: **blocking**

$\Rightarrow$ Path is blocked

$$A \perp\!\!\!\perp H, B \mid D, E$$

d-Separation: Alternative Way for Checking

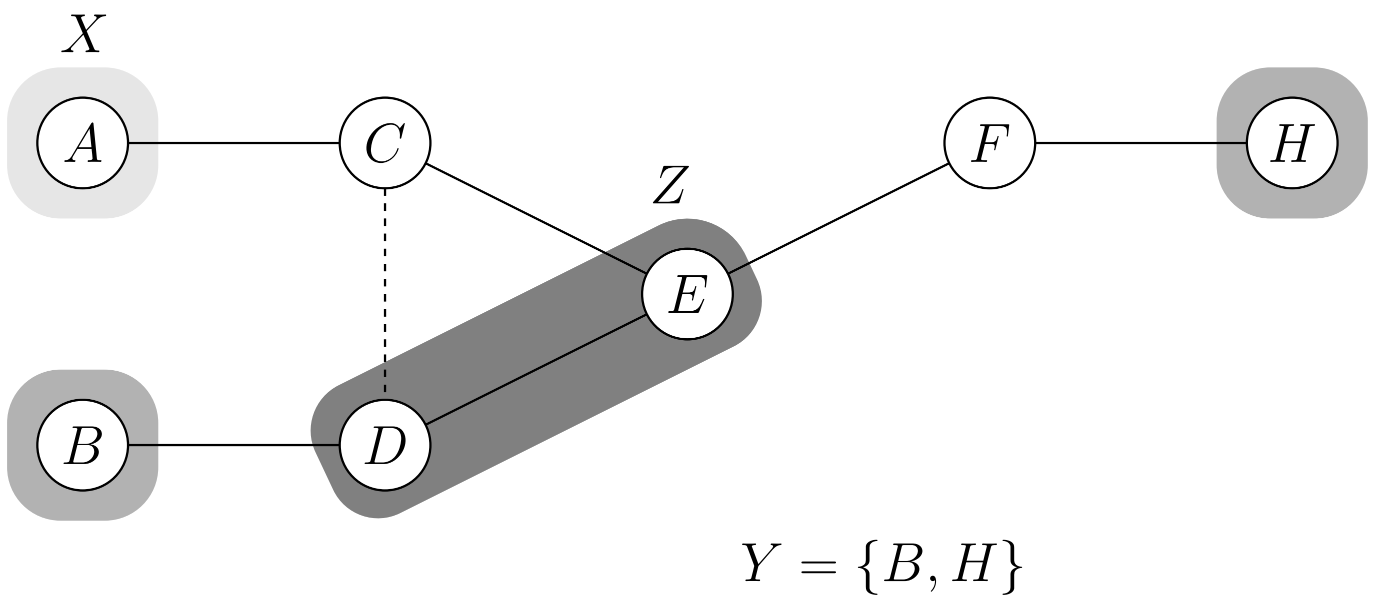


$Y = \{B, H\}$

Steps

- Create the minimal ancestral subgraph induced by $X \cup Y \cup Z$.

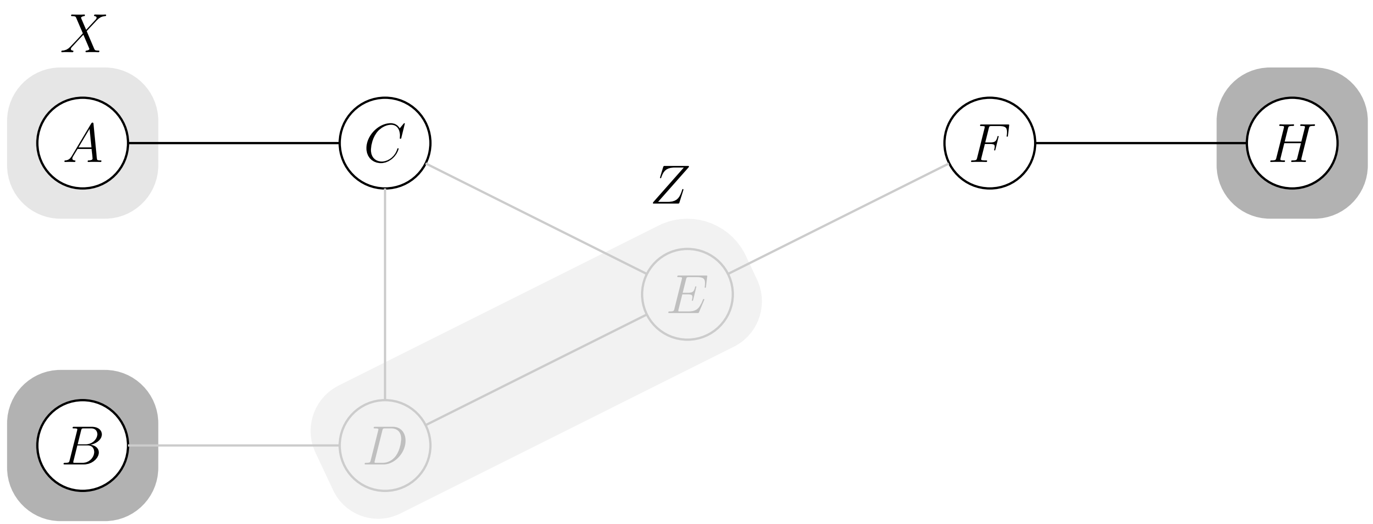
d-Separation: Alternative Way for Checking



Steps

- Create the minimal ancestral subgraph induced by $X \cup Y \cup Z$.
- Moralize that subgraph.

d-Separation: Alternative Way for Checking



$$Y = \{B, H\}$$

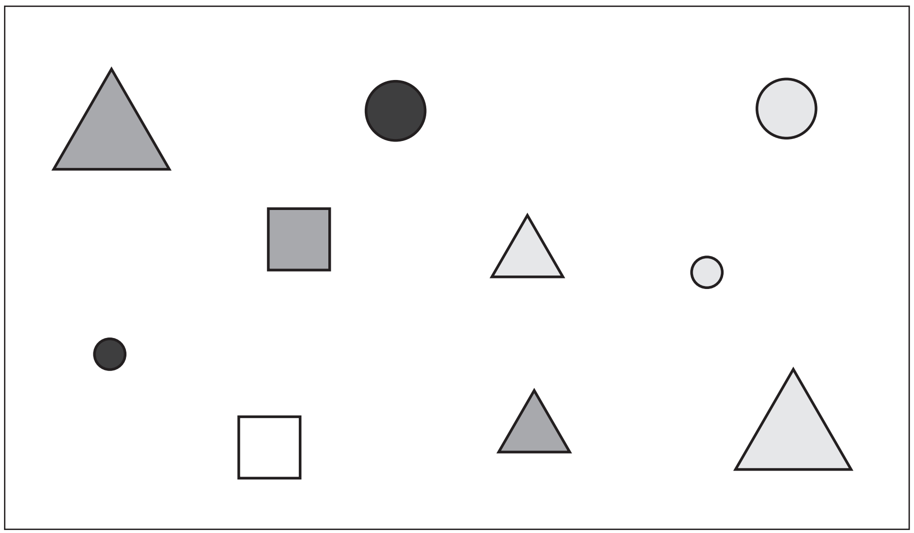
Steps:

- Create the minimal ancestral subgraph induced by $X \cup Y \cup Z$.
- Moralize that subgraph.
- Check for u-Separation in that undirected graph.

$$A \perp\!\!\!\perp H, B \mid D, E$$

Decomposition

Example World



- 10 simple geometric objects
- 3 attributes

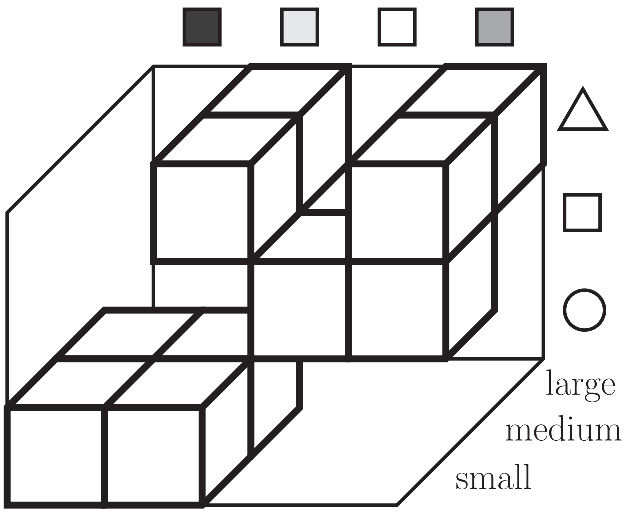
Relation

color	shape	size
■	○	small
■	○	medium
▢	○	small
▢	○	medium
▢	△	medium
▢	△	large
□	□	medium
▦	□	medium
▦	△	medium
▦	△	large

Relation

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

Geometric Representation



Object Representation

- **Universe of Discourse**: Ω
- $\omega \in \Omega$ represents a single abstract object.
- A subset $E \subseteq \Omega$ is called an **event**.
- For every event we use the function R to determine whether E is possible or not.

$$R: \ 2^{\Omega} \to \{0, 1\}$$

- We claim the following properties of R:
  1. $R(\emptyset) = 0$
  2. $\forall E_1, E_2 \subseteq \Omega: \ R(E_1 \cup E_2) \ = \ \max\{R(E_1), R(E_2)\}$
- For example:

$$R(E) \ = \ \begin{cases} 0 & \text{if } \ E = \emptyset \\ 1 & \text{otherwise} \end{cases}$$

Object Representation

- Attributes or Properties of these objects are introduced by functions: (later referred to as **random variables**)

$$A : \ \Omega \to \mathrm{dom}(A)$$

where $\mathrm{dom}(A)$ is the domain (i. e., set of all possible values) of A.

- A set of attibutes $U = \{A_1, \ldots, A_n\}$ is called an **attribute schema**.

- The **preimage** of an attribute defines an **event**:

$$\forall a \in \mathrm{dom}(A) : \ A^{-1}(a) = \{\omega \in \Omega \mid A(\omega) = a\} \subseteq \Omega$$

- Abbreviation: $A^{-1}(a) \quad = \quad \{\omega \in \Omega \mid A(\omega) = a\} \quad = \quad \{A = a\}$

- We will index the function R to stress on which events it is defined. R_{AB} will be short for $R_{\{A,B\}}$.

$$R_{AB} : \bigcup_{a \in \mathrm{dom}(A)} \bigcup_{b \in \mathrm{dom}(B)} \Big\{\{A = a, B = b\}\Big\} \to \{0, 1\}$$

Formal Representation

A = color	B = shape	C = size
$a_1 = \blacksquare$	$b_1 = \bigcirc$	c_1 = small
$a_1 = \blacksquare$	$b_1 = \bigcirc$	c_2 = medium
$a_2 = \square$	$b_1 = \bigcirc$	c_1 = small
$a_2 = \square$	$b_1 = \bigcirc$	c_2 = medium
$a_2 = \square$	$b_3 = \triangle$	c_2 = medium
$a_2 = \square$	$b_3 = \triangle$	c_3 = large
$a_3 = \square$	$b_2 = \square$	c_2 = medium
$a_4 = \square$	$b_2 = \square$	c_2 = medium
$a_4 = \square$	$b_3 = \triangle$	c_2 = medium
$a_4 = \square$	$b_3 = \triangle$	c_3 = large

$$
\begin{aligned}
R_{ABC}(A = a, B = b, C = c) \\
&= R_{ABC}(\{A = a, B = b, C = c\}) \\
&= R_{ABC}(\{\omega \in \Omega \mid A(\omega) = a \wedge \\
&\qquad\qquad B(\omega) = b \wedge \\
&\qquad\qquad C(\omega) = c)\} \\
&= \begin{cases} 0 & \text{if there is no tuple } (a, b, c) \\ 1 & \text{else} \end{cases}
\end{aligned}
$$

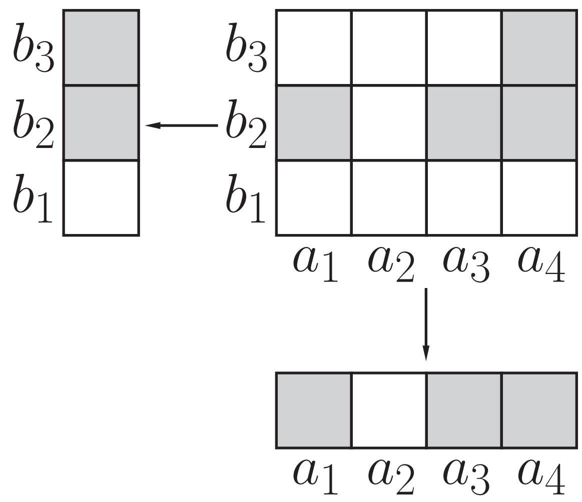
R serves as an indicator function.

Projection / Marginalization

Let R_{AB} be a relation over two attributes A and B. The projection (or marginalization) from schema $\{A, B\}$ to schema $\{A\}$ is defined as:

$$\forall a \in \mathrm{dom}(A): \; R_A(A = a) \;=\; \max_{\forall b \in \mathrm{dom}(B)} \{R_{AB}(A = a, B = b)\}$$

This principle is easily generalized to sets of attributes.

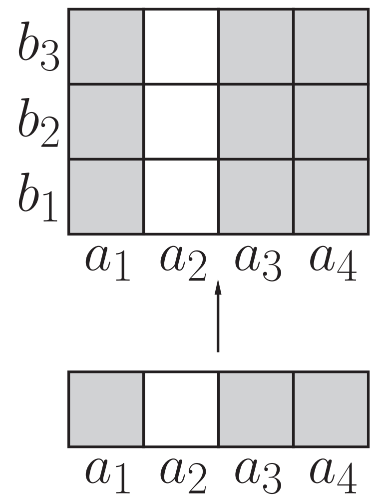


Cylindrical Extention

Let R_A be a relation over an attribute A. The cylindrical extention R_{AB} from $\{A\}$ to $\{A, B\}$ is defined as:

$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : \; R_{AB}(A = a, B = b) \; = \; R_A(A = a)$$

This principle is easily generalized to sets of attributes.

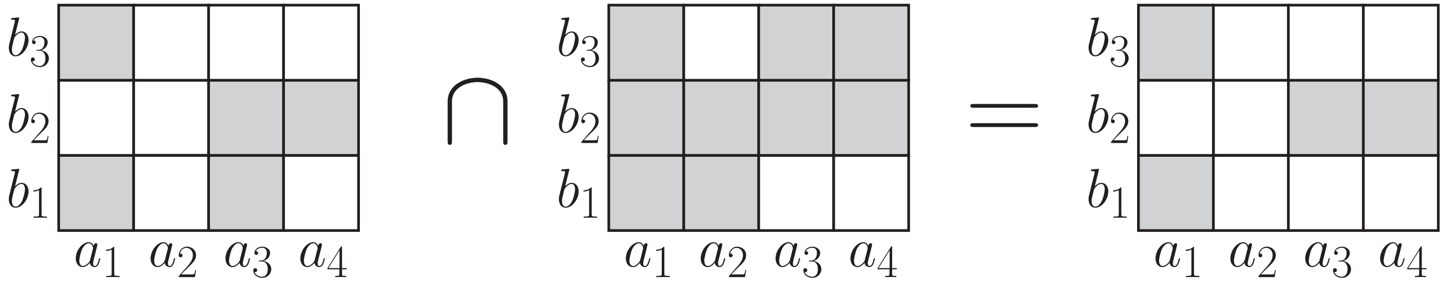


Intersection

Let $R^{(1)}_{AB}$ and $R^{(2)}_{AB}$ be two relations with attribute schema $\{A, B\}$. The intersection R_{AB} of both is defined in the natural way:

$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) :$$
$$R_{AB}(A = a, B = b) \;=\; \min\{R^{(1)}_{AB}(A = a, B = b), R^{(2)}_{AB}(A = a, B = b)\}$$

This principle is easily generalized to sets of attributes.

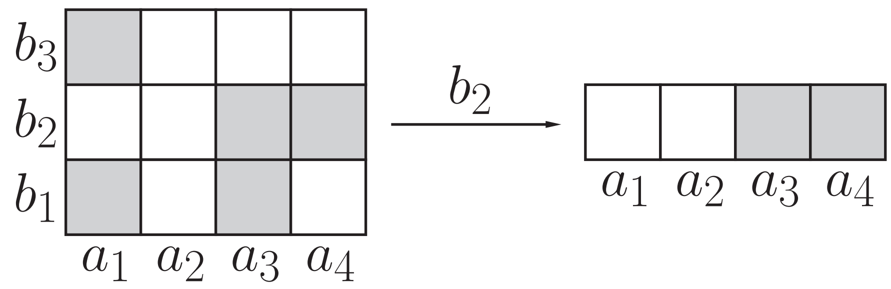


Conditional Relation

Let R_{AB} be a relation over the attribute schema $\{A, B\}$. The conditional relation of A given B is defined as follows:

$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : R_A(A = a \mid B = b) \;=\; R_{AB}(A = a, B = b)$$

This principle is easily generalized to sets of attributes.



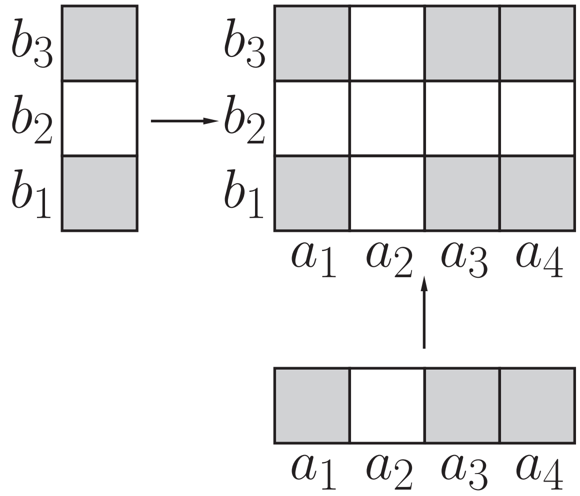
Object Representation

(Unconditional) Independence

Let R_{AB} be a relation over the attribute schema $\{A, B\}$. We call A and B relationally independent (w. r. t. R_{AB}) if the following condition holds:

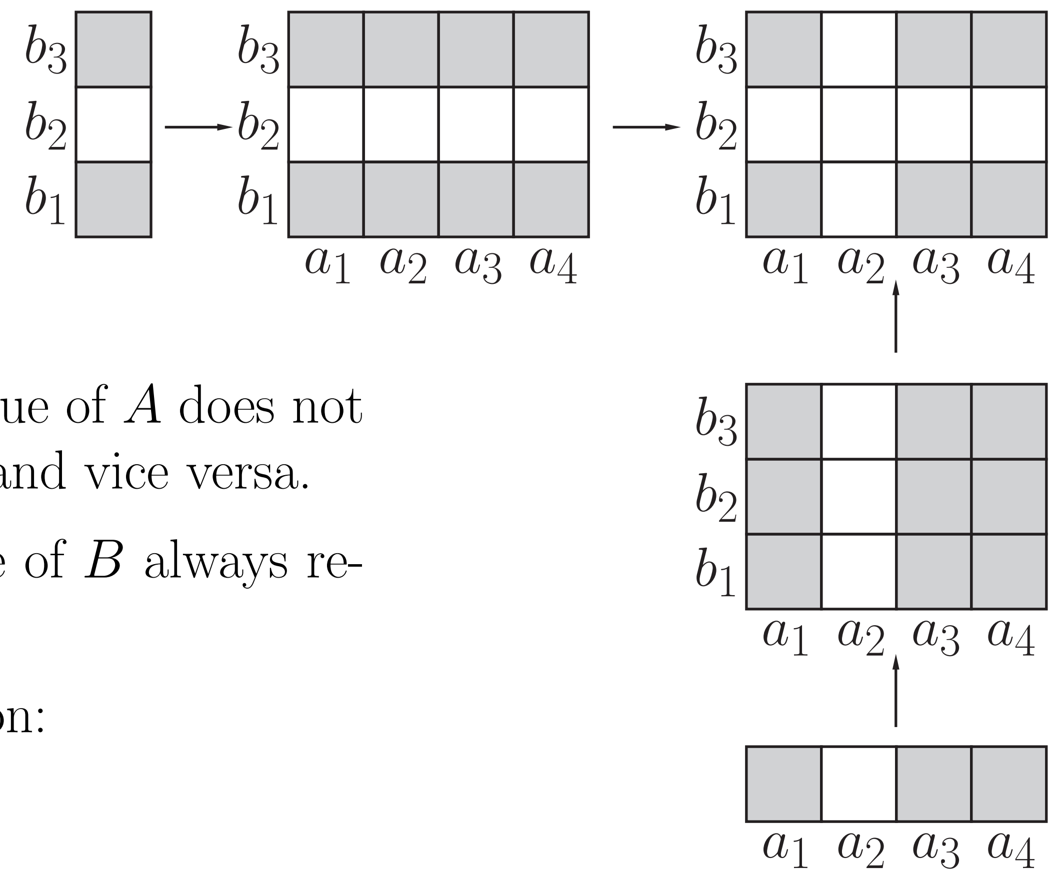
$$\forall a \in \mathrm{dom}(A) : \forall b \in \mathrm{dom}(B) : R_{AB}(A = a, B = b) \;=\; \min\{R_A(A = a), R_B(B = b)\}$$

This principle is easily generalized to sets of attributes.



Object Representation

(Unconditional) Independence



Intuition: Fixing one (possible) value of A does not restrict the (possible) values of B and vice versa.

Conditioning on any possible value of B always results in the same relation R_A.

Alternative independence expression:

$$\forall b \in \mathrm{dom}(B): \ R_B(B = b) = 1 : \quad R_A(A = a \mid B = b) \ = \ R_A(A = a)$$

Decomposition

- Obviously, the original two-dimensional relation can be reconstructed from the two one-dimensional ones, if we have (unconditional) independence.

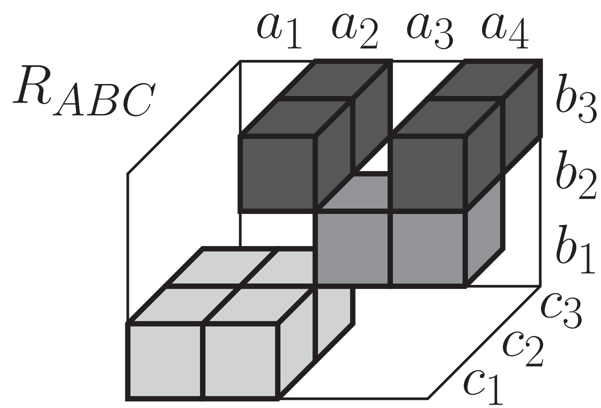
- The definition for (unconditional) independence already told us how to do so:

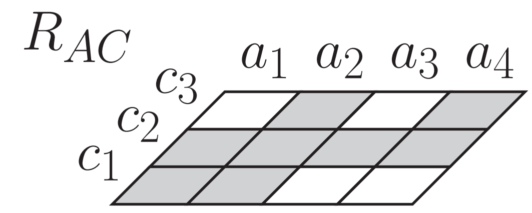
$$R_{AB}(A = a, B = b) \;=\; \min\{R_A(A = a), R_B(B = b)\}$$

- Storing R_A and R_B is sufficient to represent the information of R_{AB}.

- **Question:** The (unconditional) independence is a rather strong restriction. Are there other types of independence that allow for a decomposition as well?

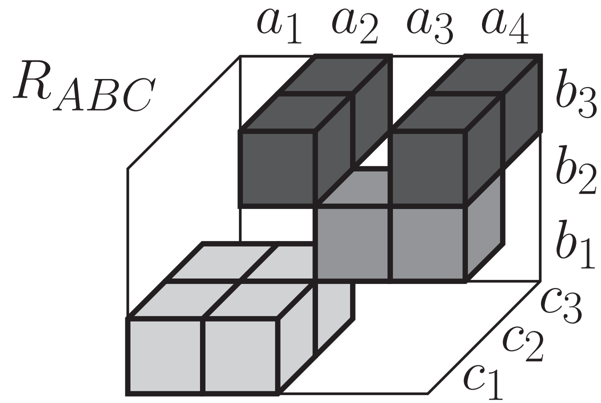
Conditional Relational Independence

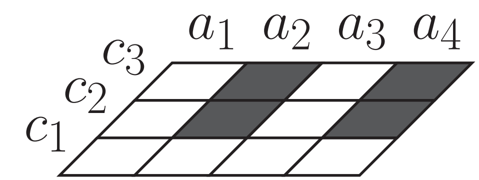




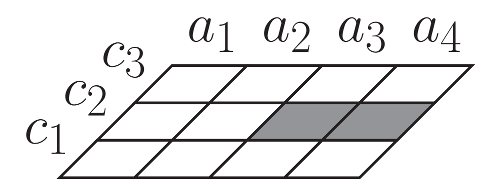
Clearly, A and C are unconditionally dependent, i. e. the relation R_{AC} cannot be reconstructed from R_A and R_C.

Conditional Relational Independence

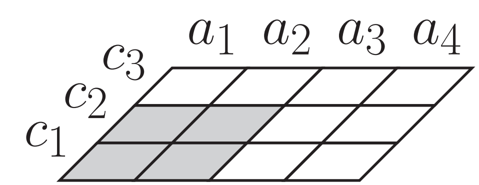




$R_{AC}(\,\cdot,\cdot \mid B = b_3)$



$R_{AC}(\,\cdot,\cdot \mid B = b_2)$



$R_{AC}(\,\cdot,\cdot \mid B = b_1)$

However, given all possible values of B, all respective conditional relations R_{AC} show the independence of A and C.

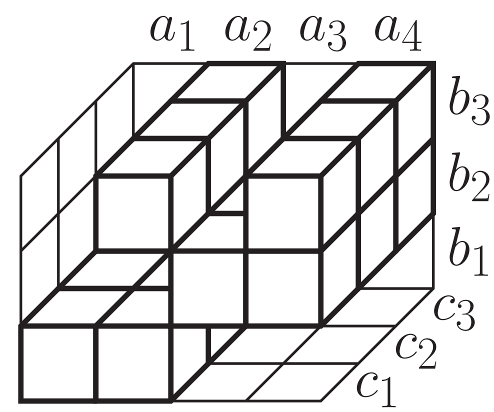
$$R_{AC}(a, c \mid b) \;=\; \min\{R_A(a \mid b), R_C(c \mid b)\}$$

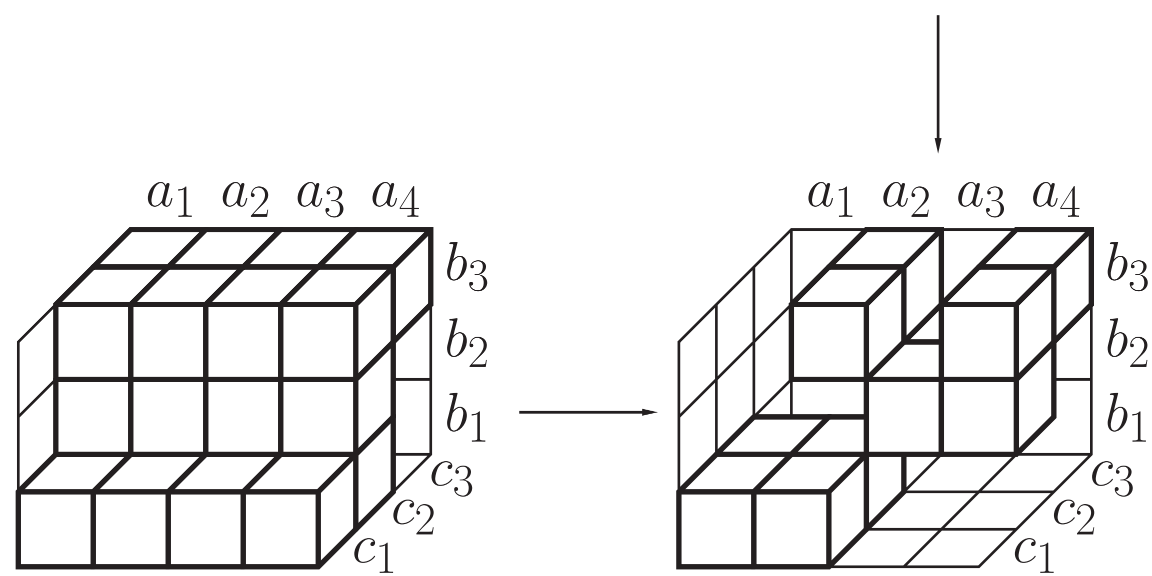
With the definition of a conditional relation, the decomposition description for R_{ABC} reads:

$$R_{ABC}(a, b, c) \;=\; \min\{R_{AB}(a, b), R_{BC}(b, c)\}$$

Again, we reconstruct the initial relation from the cylindrical extentions of the two relations formed by the attributes A, B and B, C.

It is possible since A and C are (relationally) independent given B.





Probability Foundations

Reminder: Probability Theory

- **Goal**: Make statements and/or predictions about results of physical processes.
- Even processes that seem to be simple at first sight may reveal considerable difficulties when trying to predict.
- Describing real-world physical processes always calls for a simplifying mathematical model.
- Although everybody will have some intuitive notion about probability, we have to formally define the underlying mathematical structure.
- Randomness or chance enters as the incapability of precisely modelling a process or the inability of measuring the initial conditions.
  - *Example*: Predicting the trajectory of a billard ball over more than 9 banks requires more detailed measurement of the initial conditions (ball location, applied momentum etc.) than physically possible according to Heisenberg's uncertainty principle.

Formal Approach on the Model Side

- We conduct an experiment that has a set Ω of possible outcomes.
  E. g.:
  - Rolling a die ($\Omega = \{1, 2, 3, 4, 5, 6\}$)
  - Arrivals of phone calls ($\Omega = \mathbb{N}_0$)
  - Bread roll weights ($\Omega = \mathbb{R}_+$)
- Such an outcome is called an **elementary event**.
- All possible elementary events are called the **frame of discernment** Ω (or sometimes **universe of discourse**).
- The set representation stresses the following facts:
  - All possible outcomes are covered by the elements of Ω. (**collectively exhaustive**).
  - Every possible outcome is represented by exactly one element of Ω. (**mutual disjoint**).

Events

- Often, we are interested in *higher-level* events
(e. g. casting an odd number, arrival of at least 5 phone calls or purchasing a bread roll heavier than 80 grams)

- Any subset $A \subseteq \Omega$ is called an **event** which **occurs**, if the outcome $\omega_0 \in \Omega$ of the random experiment lies in A:

$$\text{Event } A \subseteq \Omega \ \text{ occurs} \quad \Leftrightarrow \quad \bigvee_{\omega \in A} (\omega = \omega_0) \ = \ \mathsf{true} \quad \Leftrightarrow \quad \omega_0 \in A$$

- Since events are sets, we can define for two events A and B:
  - $A \cup B$ occurs if A or B occurs; $A \cap B$ occurs if A and B occurs.
  - $\overline{A}$ occurs if A does not occur (i. e., if $\Omega \backslash A$ occurs).
  - A and B are *mutually exclusive*, iff $A \cap B = \emptyset$.

Event Algebra

- A family of sets $\mathcal{E} = \{E_1, \dots, E_n\}$ is called an **event algebra**, if the following conditions hold:
  - The **certain event** Ω lies in $\mathcal{E}$.
  - If $E \in \mathcal{E}$, then $\overline{E} = \Omega \backslash E \in \mathcal{E}$.
  - If E_1 and E_2 lie in $\mathcal{E}$, then $E_1 \cup E_2 \in \mathcal{E}$ and $E_1 \cap E_2 \in \mathcal{E}$.
- If Ω is uncountable, we require the additional property:

  For a series of events $E_i \in \mathcal{E}, i \in \mathbb{N}$, the events $\bigcup_{i=1}^{\infty} E_i$ and $\bigcap_{i=1}^{\infty} E_i$ are also in $\mathcal{E}$.

  $\mathcal{E}$ is then called a σ**-algebra**.

Side remarks:

- Smallest event algebra: $\mathcal{E} = \{\emptyset, \Omega\}$
- Largest event algebra (for finite or countable Ω): $\mathcal{E} = 2^\Omega = \{A \subseteq \Omega \mid \mathsf{true}\}$

Probability Function

- Given an event algebra $\mathcal{E}$, we would like to assign every event $E \in \mathcal{E}$ its probability with a **probability function** $P: \ \mathcal{E} \to [0, 1]$.

- We require P to satisfy the so-called **Kolmogorov Axioms**:
    - $\forall E \in \mathcal{E}: \ 0 \ \leq \ P(E) \ \leq \ 1$
    - $P(\Omega) \ = \ 1$
    - If $E_1, E_2 \in \mathcal{E}$ are mutually exclusive, then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

- From these axioms one can conclude the following (incomplete) list of properties:
    - $\forall E \in \mathcal{E}: \ P(\overline{E}) \ = \ 1 - P(E)$
    - $P(\emptyset) \ = \ 0$
    - For pairwise disjoint events $E_1, E_2, \ldots \in \mathcal{E}$ holds:

$$P(\bigcup_{i=1}^{\infty} E_i) \ = \ \sum_{i=1}^{\infty} P(E_i)$$

Note that for $|\Omega| < \infty$ the union and sum are finite also.

Elementary Probabilities and Densities

Question 1: How to calculate P?

Question 2: Are there "default" event algebras?

- Idea for question 1: We have to find a way of distributing (thus the notion *distribution*) the unit mass of probability over all elements $\omega \in \Omega$.
    - If Ω is finite or countable a **probability mass function** p is used:

$$p: \ \Omega \to [0,1] \quad \text{and} \quad \sum_{\omega \in \Omega} p(\omega) = 1$$

    - If Ω is uncountable (i. e., continuous) a **probability density function** f is used:

$$f: \ \Omega \to \mathbb{R} \quad \text{and} \quad \int_{\Omega} f(\omega) \, \mathrm{d}\omega = 1$$

“Default” Event Algebras

- Idea for question 2 (“default” event algebras) we have to distinguish again between the cardinalities of Ω:
  - Ω finite or countable: $\mathcal{E} = 2^{\Omega}$
  - Ω uncountable, e. g. $\Omega = \mathbb{R}$: $\mathcal{E} = \mathcal{B}(\mathbb{R})$
- $\mathcal{B}(\mathbb{R})$ is the **Borel Algebra**, i. e., the smallest σ-algebra that contains all closed intervals $[a, b] \subset \mathbb{R}$ with $a < b$.
- $\mathcal{B}(\mathbb{R})$ also contains all open intervals and single-item sets.
- It is sufficient to note here, that all intervals are contained

$$\{[a, b]\,,]a, b]\,,]a, b[\,, [a, b[\, \subset \mathbb{R} \mid a < b\} \subset \mathcal{B}(\mathbb{R})$$

because the event of a bread roll having a weight between 80 g and 90 g is represented by the interval $[80, 90]$.

Probability Spaces

- For a sample space A, an event algebra B (over A) and a probability function C, we call the triple

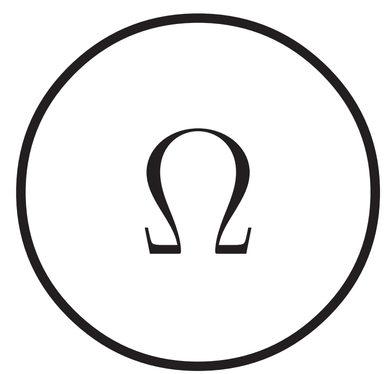
$$(A, B, C)$$

a **probability space**.

Real World

$(\Xi, \mathcal{X}, Q)$

Model



$(\Omega, \mathcal{E}, P)$

Reminder: Preimage of a Function

- Let $f: \ D \to M$ be a function that assigns to every value of D a value in M.
- For every value of $y \in M$ we can ask which values of $x \in D$ are mapped to y:

$$D \supseteq \{x \in D \mid f(x) = y\} \stackrel{\text{Def}}{=} f^{-1}(y)$$

- $f^{-1}(y)$ is called the **preimage** of y under f, denoted also as $\{f = y\}$.
- The notion can be generalized from $y \in M$ to sets $B \subseteq M$:

$$D \supseteq \{x \in D \mid f(x) \in B\} \stackrel{\text{Def}}{=} f^{-1}(B)$$

- If f is bijective then $\forall y \in M : \left|f^{-1}(y)\right| = 1$.
- Examples:
  - $\sin^{-1}(0) = \{k \cdot \pi \mid k \in \mathbb{Z}\}$
  - $\exp^{-1}(1) = \{0\}$
  - $\mathrm{sgn}^{-1}(1) = (0, +\infty) \subset \mathbb{R}$

Random Variable

We still need a means of mapping real-world outcomes in Ξ to our space Ω.

- A function $X : D \to M$ is called a **random variable** iff the preimage of any value of M is an event (in some probability space).

- If X maps Ξ onto Ω, we define

$$P_X(X \in A) = Q(\{\xi \in \Xi \mid X(\xi) \in A\}).$$

- X may also map from Ω to another domain: $X : \Omega \to \mathrm{dom}(X)$. We then define:

$$P_X(X \in A) = P(\{\omega \in \Omega \mid X(\omega) \in A\}).$$

- If X is numeric, we call $F(x)$ with

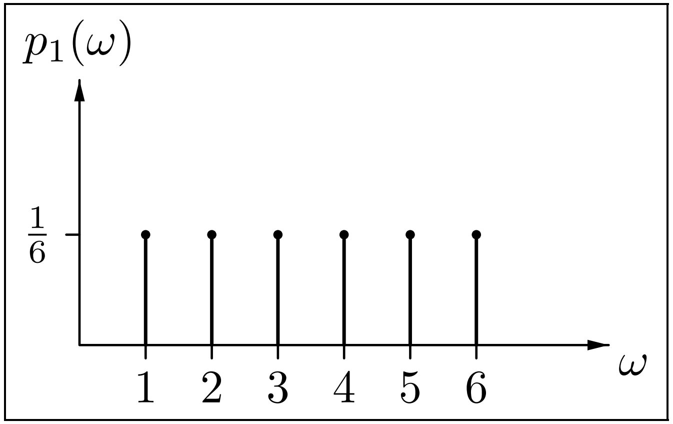
$$F(x) = P(X \leq x)$$

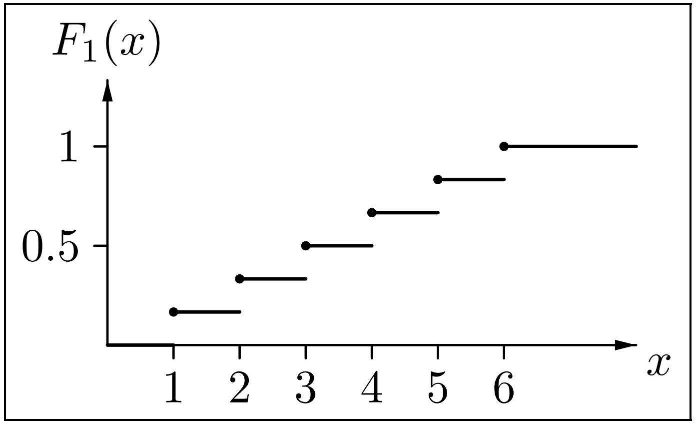
the **distribution function** of X.

Example: Rolling a Die

$\Omega = \{1, 2, 3, 4, 5, 6\} \quad X = \mathrm{id}$

$$p_1(\omega) = \tfrac{1}{6} \qquad\qquad F_1(x) = P(X \le x)$$





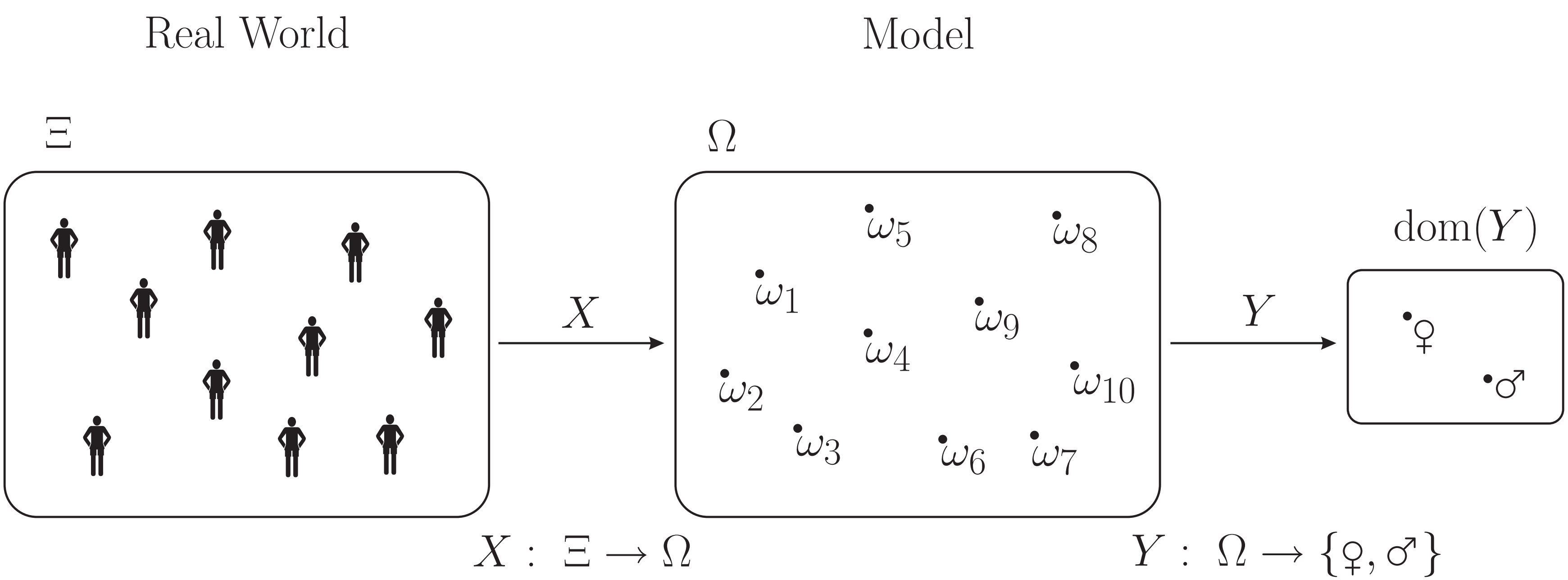
$$\sum_{\omega \in \Omega} p_1(\omega) = \sum_{i=1}^{6} p_1(\omega_i) = \sum_{i=1}^{6} \frac{1}{6} = 1$$

$$P(X \le x) = \sum_{x' \le x} P(X = x')$$

$$P(a < X \le b) = F_1(b) - F_1(a)$$

$$P(X = x) = P(\{X = x\}) = P(X^{-1}(x)) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

The Big Picture



$$Q\Big(\{\xi \in \Xi \mid X(\xi) \in Y^{-1}(\text{♀})\}\Big) \quad = \quad P\Big(\{\omega \in \Omega \mid Y(\omega) = \text{♀}\}\Big) \quad = \quad P\Big(Y = \text{♀}\Big) \quad = \quad P\Big(\text{♀}\Big)$$

Applied Probability Theory

Why (Kolmogorov) Axioms?

- If P models an *objectively* observable probability, these axioms are obviously reasonable.
- However, why should an agent obey formal axioms when modeling degrees of (subjective) belief?
- Objective vs. subjective probabilities
- Axioms constrain the set of beliefs an agent can abide.
- Finetti (1931) gave one of the most plausible arguments why subjective beliefs should respect axioms:

  "When using contradictory beliefs, the agent will eventually fail."

Unconditional Probabilities

- $P(A)$ designates the *unconditioned* or *a priori* probability that $A \subseteq \Omega$ occurs if *no* other additional information is present.

  For example:

  $$P(\mathsf{cavity}) \;=\; 0.1$$

  Note: Here, cavity is a proposition.

- A formally different way to state the same would be via a binary random variable Cavity:

  $$P(\mathsf{Cavity} = \mathsf{true}) \;=\; 0.1$$

- A priori probabilities are derived from statistical surveys or general rules.

Unconditional Probabilities

- In general a random variable can assume more than two values:

$$
\begin{aligned}
P(\ \mathsf{Weather} = \mathsf{sunny}\ ) &= 0.7 \\
P(\ \mathsf{Weather} = \mathsf{rainy}\ ) &= 0.2 \\
P(\ \mathsf{Weather} = \mathsf{cloudy}) &= 0.02 \\
P(\ \mathsf{Weather} = \mathsf{snowy}\ ) &= 0.08 \\
P(\mathsf{Headache} = \mathsf{true}\ ) &= 0.1
\end{aligned}
$$

- $P(X)$ designates the vector of probabilities for the (ordered) domain of the random variable X:

$$
\begin{aligned}
P(\mathsf{Weather}) &= \langle 0.7, 0.2, 0.02, 0.08 \rangle \\
P(\mathsf{Headache}) &= \langle 0.1, 0.9 \rangle
\end{aligned}
$$

- Both vectors define the respective probability distributions of the two random variables.

Conditional Probabilities

- New evidence can alter the probability of an event.
- Example: The probability for cavity increases if information about a toothache arises.
- With additional information present, the a priori knowledge must not be used!
- $P(A \mid B)$ designates the *conditional* or *a posteriori* probability of A *given* the sole observation (*evidence*) B.

$$P(\mathsf{cavity} \mid \mathsf{toothache}) \;=\; 0.8$$

- For random variables X and Y $P(X \mid Y)$ represents the set of conditional distributions for each possible value of Y.

Conditional Probabilities

- $P(\mathsf{Weather} \mid \mathsf{Headache})$ consists of the following table:

	$\mathsf{h} \mathrel{\widehat{=}} \mathsf{Headache} = \mathsf{true}$	$\neg\mathsf{h} \mathrel{\widehat{=}} \mathsf{Headache} = \mathsf{false}$
$\mathsf{Weather} = \mathsf{sunny}$	$P(\mathsf{W} = \mathsf{sunny} \mid \mathsf{h})$	$P(\mathsf{W} = \mathsf{sunny} \mid \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{rainy}$	$P(\mathsf{W} = \mathsf{rainy} \mid \mathsf{h})$	$P(\mathsf{W} = \mathsf{rainy} \mid \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{cloudy}$	$P(\mathsf{W} = \mathsf{cloudy} \mid \mathsf{h})$	$P(\mathsf{W} = \mathsf{cloudy} \mid \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{snowy}$	$P(\mathsf{W} = \mathsf{snowy} \mid \mathsf{h})$	$P(\mathsf{W} = \mathsf{snowy} \mid \neg\mathsf{h})$

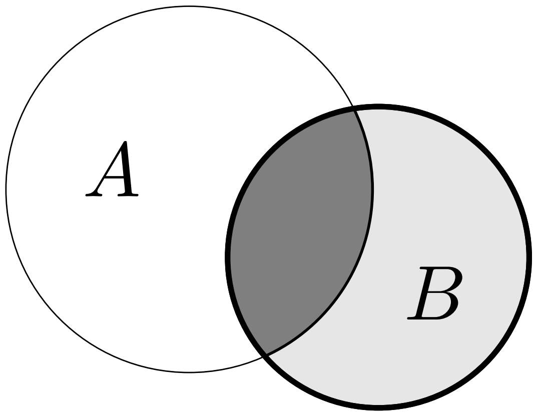
- Note that we are dealing with *two* distributions now!
  Therefore each column sums up to unity!

- Formal definition:

$$P(A \mid B) \;=\; \frac{P(A \wedge B)}{P(B)} \qquad \text{if} \quad P(B) > 0$$

Conditional Probabilities

$$\boxed{P(A \mid B) \;=\; \frac{P(A \wedge B)}{P(B)}}$$



- Product Rule: $P(A \wedge B) \;=\; P(A \mid B) \cdot P(B)$
- Also: $P(A \wedge B) \;=\; P(B \mid A) \cdot P(A)$
- A and B are *independent* iff

$$P(A \mid B) \;=\; P(A) \qquad \text{and} \qquad P(B \mid A) \;=\; P(B)$$

- Equivalently, iff the following equation holds true:

$$P(A \wedge B) \;=\; P(A) \cdot P(B)$$

Interpretation of Conditional Probabilities

Caution! Common misinterpretation:

"$P(A \mid B) = 0.8$ means, that $P(A) = 0.8$, given B holds."

This statement is wrong due to (at least) two facts:

- $P(A)$ is *always* the a-priori probability,
  never the probability of A given that B holds!

- $P(A \mid B) = 0.8$ is only applicable as long as no other evidence except B is present.
  If C becomes known, $P(A \mid B \wedge C)$ has to be determined.

  In general we have:

  $$P(A \mid B \wedge C) \neq P(A \mid B)$$

  E. g. $C \rightarrow A$ might apply.

Joint Probabilities

- Let $X_1, \dots, X_n$ be random variables over the same framce of descernment Ω and event algebra $\mathcal{E}$. Then $\vec{X} = (X_1, \dots, X_n)$ is called a *random vector* with

$$\vec{X}(\omega) \;=\; (X_1(\omega), \dots, X_n(\omega))$$

- Shorthand notation:

$$P(\vec{X} = (x_1, \dots, x_n)) \;=\; P(X_1 = x_1, \dots, X_n = x_n) \;=\; P(x_1, \dots, x_n)$$

- Definition:

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n) \;&=\; P\bigg(\Big\{\, \omega \in \Omega \;\Big|\; \bigwedge_{i=1}^{n} X_i(\omega) = x_i \,\Big\}\bigg) \\ &=\; P\bigg(\, \bigcap_{i=1}^{n} \{X_i = x_i\} \,\bigg) \end{aligned}$$

Joint Probabilities

- Example: $P(\mathsf{Headache}, \mathsf{Weather})$ is the *joint probability distribution* of both random variables and consists of the following table:

	$\mathsf{h} \mathrel{\widehat{=}} \mathsf{Headache} = \mathsf{true}$	$\neg\mathsf{h} \mathrel{\widehat{=}} \mathsf{Headache} = \mathsf{false}$
$\mathsf{Weather} = \mathsf{sunny}$	$P(\mathsf{W} = \mathsf{sunny} \wedge \mathsf{h})$	$P(\mathsf{W} = \mathsf{sunny} \wedge \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{rainy}$	$P(\mathsf{W} = \mathsf{rainy} \wedge \mathsf{h})$	$P(\mathsf{W} = \mathsf{rainy} \wedge \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{cloudy}$	$P(\mathsf{W} = \mathsf{cloudy} \wedge \mathsf{h})$	$P(\mathsf{W} = \mathsf{cloudy} \wedge \neg\mathsf{h})$
$\mathsf{Weather} = \mathsf{snowy}$	$P(\mathsf{W} = \mathsf{snowy} \wedge \mathsf{h})$	$P(\mathsf{W} = \mathsf{snowy} \wedge \neg\mathsf{h})$

- All table cells sum up to unity.

Calculating with Joint Probabilities

All desired probabilities can be computed from a joint probability distribution.

	toothache	¬toothache
cavity	0.04	0.06
¬cavity	0.01	0.89

- Example: $$\begin{aligned} P(\textsf{cavity} \lor \textsf{toothache}) &= P(\textsf{cavity} \land \textsf{toothache}) \\ &+ P(\neg\textsf{cavity} \land \textsf{toothache}) \\ &+ P(\textsf{cavity} \land \neg\textsf{toothache}) \;=\; 0.11 \end{aligned}$$

- Marginalizations: $$\begin{aligned} \mathrm{P}(\textsf{cavity}) &= P(\textsf{cavity} \land \textsf{toothache}) \\ &+ P(\textsf{cavity} \land \neg\textsf{toothache}) \;=\; 0.10 \end{aligned}$$

- Conditioning:

$$P(\textsf{cavity} \mid \textsf{toothache}) \;=\; \frac{P(\textsf{cavity} \land \textsf{toothache})}{P(\textsf{toothache})} \;=\; \frac{0.04}{0.04 + 0.01} \;=\; 0.80$$

Problems

- Easiness of computing all desired probabilities comes at an unaffordable price:

  Given n random variables with k possible values each, the joint probability distribution contains k^n entries which is infeasible in practical applications.

- Hard to handle.

- Hard to estimate.

Therefore:

1. Is there a more *dense* representation of joint probability distributions?

2. Is there a more *efficient* way of processing this representation?

- The answer is *no* for the general case, however, certain dependencies and independencies can be exploited to reduce the number of parameters to a practical size.

Stochastic Independence

- Two events A and B are *stochastically independent* iff

$$P(A \wedge B) \;=\; P(A) \cdot P(B)$$
$$\Leftrightarrow$$
$$P(A \mid B) \;=\; P(A) \;=\; P(A \mid \overline{B})$$

- Two random variables X and Y are *stochastically independent* iff

$$\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : \qquad P(X = x, Y = y) \;=\; P(X = x) \cdot P(Y = y)$$
$$\Leftrightarrow$$
$$\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : \qquad P(X = x \mid Y = y) \;=\; P(X = x)$$

- Shorthand notation: $P(X, Y) = P(X) \cdot P(Y)$.
  Note the formal difference between $P(A) \in [0, 1]$ and $P(X) \in [0, 1]^{|\mathrm{dom}(X)|}$.

Conditional Independence

- Let X, Y and Z be three random variables. We call X *and* Y *conditionally independent given* Z, iff the following condition holds:

$$\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : \forall z \in \mathrm{dom}(Z) :$$
$$P(X = x, Y = y \mid Z = z) \ = \ P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

- Shorthand notation: $X \perp\!\!\!\perp_P Y \mid Z$

- Let $\boldsymbol{X} = \{A_1, \ldots, A_k\}$, $\boldsymbol{Y} = \{B_1, \ldots, B_l\}$ and $\boldsymbol{Z} = \{C_1, \ldots, C_m\}$ be three disjoint sets of random variables. We call $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *conditionally independent given* $\boldsymbol{Z}$, iff

$$P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) \ = \ P(\boldsymbol{X} \mid \boldsymbol{Z}) \cdot P(\boldsymbol{Y} \mid \boldsymbol{Z}) \Leftrightarrow P(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}) \ = \ P(\boldsymbol{X} \mid \boldsymbol{Z})$$

- Shorthand notation: $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$

Conditional Independence

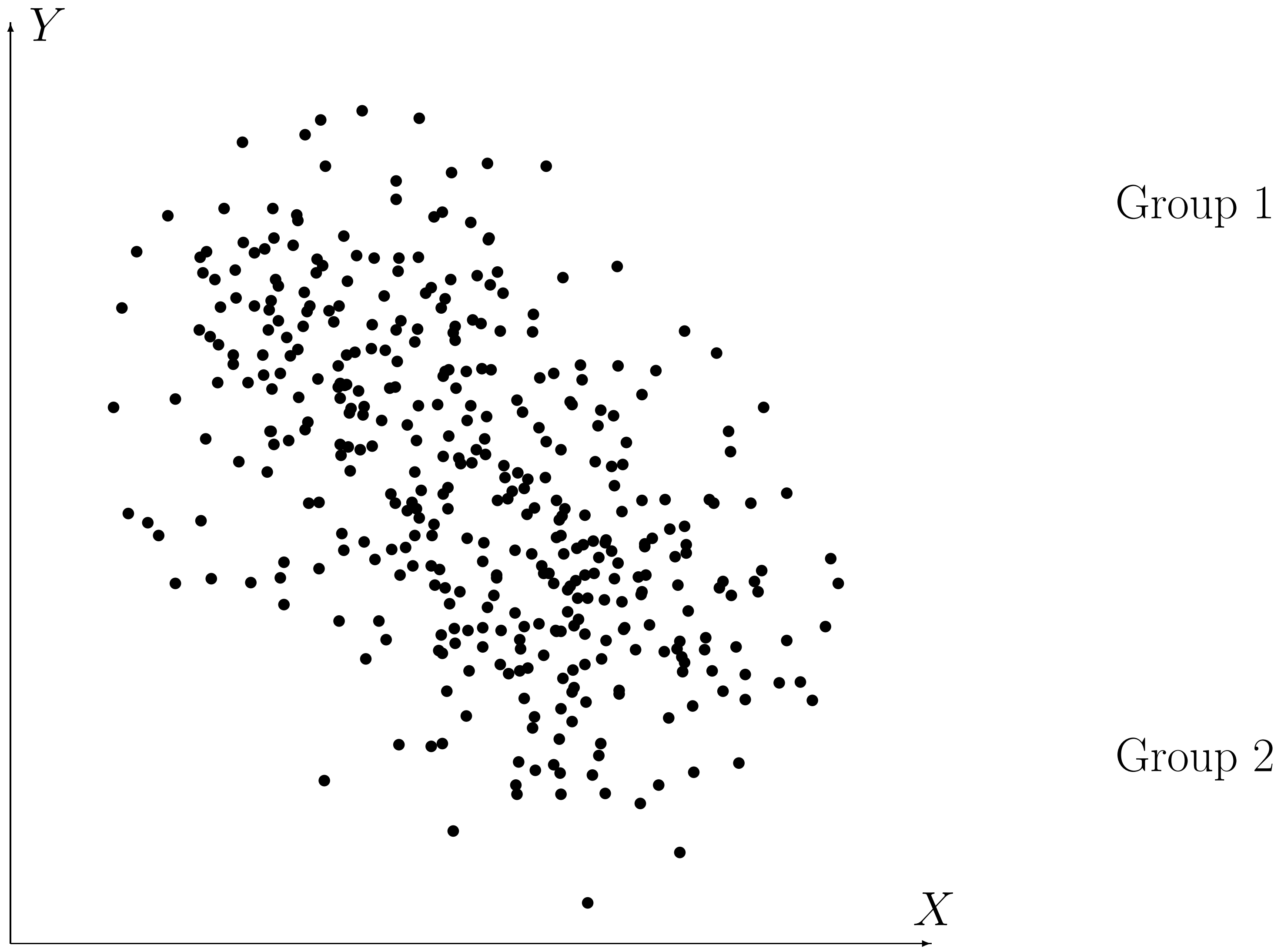
- The complete condition for $\boldsymbol{X} \perp\!\!\!\perp_P \boldsymbol{Y} \mid \boldsymbol{Z}$ would read as follows:

$$
\begin{aligned}
&\forall a_1 \in \mathrm{dom}(A_1) : \cdots \forall a_k \in \mathrm{dom}(A_k) : \\
&\quad \forall b_1 \in \mathrm{dom}(B_1) : \cdots \forall b_l \in \mathrm{dom}(B_l) : \\
&\quad\quad \forall c_1 \in \mathrm{dom}(C_1) : \cdots \forall c_m \in \mathrm{dom}(C_m) : \\
&\quad\quad\quad P(A_1 = a_1, \ldots, A_k = a_k, B_1 = b_1, \ldots, B_l = b_l \mid C_1 = c_1, \ldots, C_m = c_m) \\
&\quad\quad\quad\quad = P(A_1 = a_1, \ldots, A_k = a_k \mid C_1 = c_1, \ldots, C_m = c_m) \\
&\quad\quad\quad\quad\quad \cdot P(B_1 = b_1, \ldots, B_l = b_l \mid C_1 = c_1, \ldots, C_m = c_m)
\end{aligned}
$$

- Remarks:

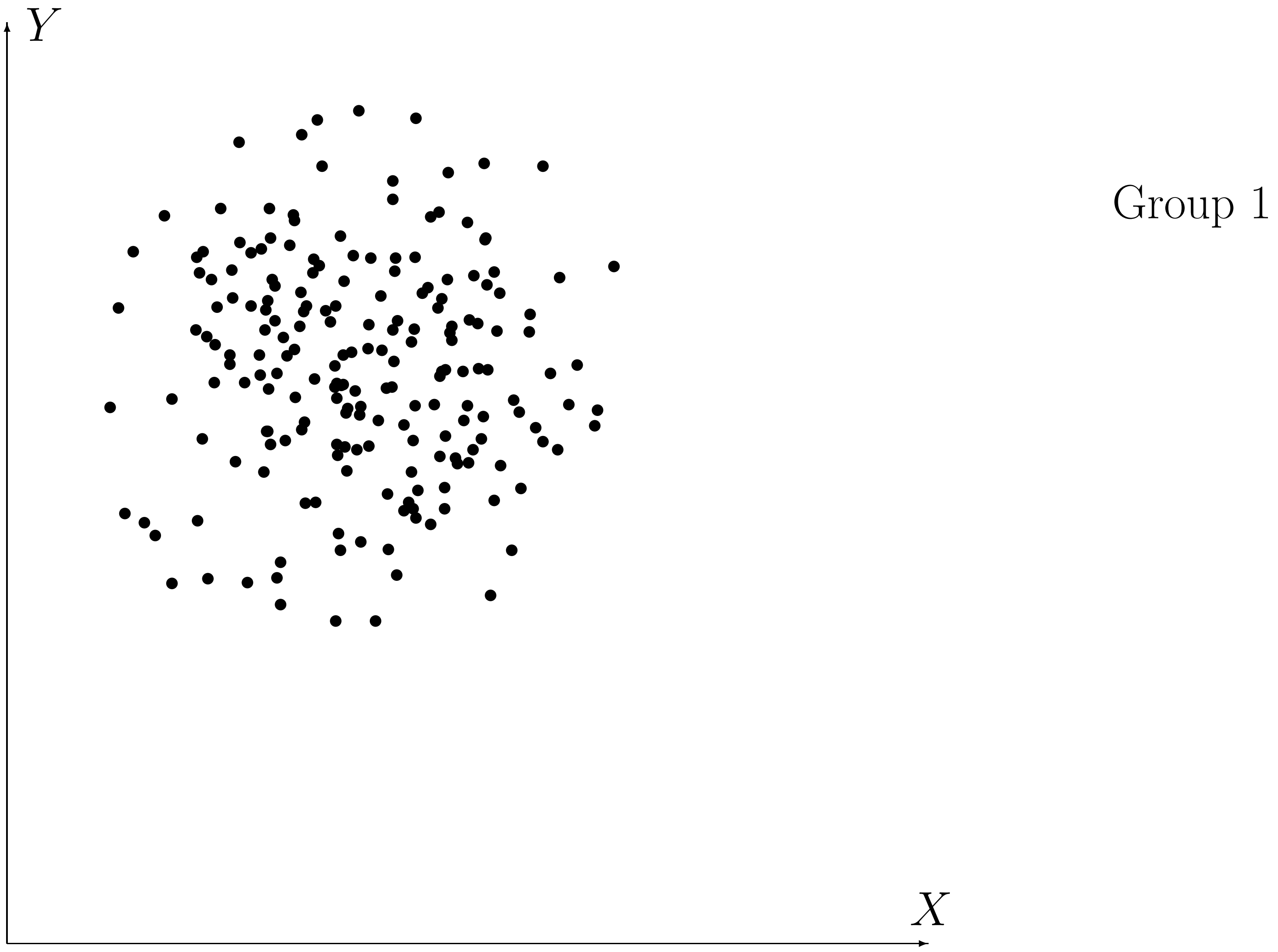
  1. If $\boldsymbol{Z} = \emptyset$ we get (unconditional) independence.
  2. We do not use curly braces ($\{\}$) for the sets if the context is clear. Likewise, we use X instead of $\boldsymbol{X}$ to denote sets.

Conditional Independence — Example 1



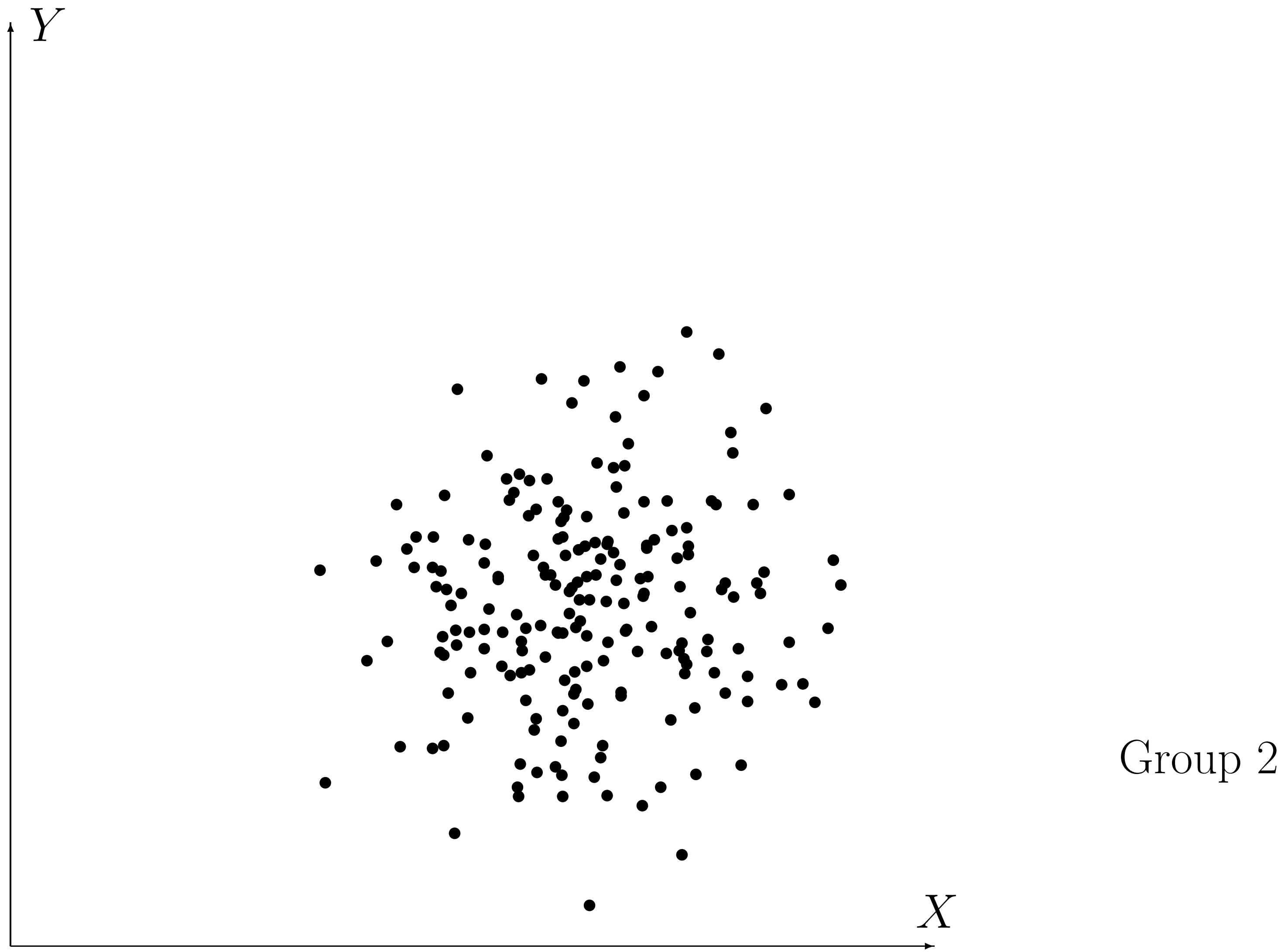
(Weak) Dependence in the entire dataset: X and Y dependent.

Conditional Independence — Example 1



No Dependence in Group 1: X and Y conditionally independent given Group 1.

Conditional Independence — Example 1



No Dependence in Group 2: X and Y conditionally independent given Group 2.

Conditional Independence — Example 2

- $\mathrm{dom}(G) = \{\mathsf{mal}, \mathsf{fem}\}$ Geschlecht (gender)
- $\mathrm{dom}(S) = \{\mathsf{sm}, \overline{\mathsf{sm}}\}$ Raucher (smoker)
- $\mathrm{dom}(M) = \{\mathsf{mar}, \overline{\mathsf{mar}}\}$ Verheiratet (married)
- $\mathrm{dom}(P) = \{\mathsf{preg}, \overline{\mathsf{preg}}\}$ Schwanger (pregnant)

p_{GSMP}		$\mathsf{G} = \mathsf{mal}$		$\mathsf{G} = \mathsf{fem}$	
		$\mathsf{S} = \mathsf{sm}$	$\mathsf{S} = \overline{\mathsf{sm}}$	$\mathsf{S} = \mathsf{sm}$	$\mathsf{S} = \overline{\mathsf{sm}}$
$\mathsf{M} = \mathsf{mar}$	$\mathsf{P} = \mathsf{preg}$	0	0	0.01	0.05
	$\mathsf{P} = \overline{\mathsf{preg}}$	0.04	0.16	0.02	0.12
$\mathsf{M} = \overline{\mathsf{mar}}$	$\mathsf{P} = \mathsf{preg}$	0	0	0.01	0.01
	$\mathsf{P} = \overline{\mathsf{preg}}$	0.10	0.20	0.07	0.21

Conditional Independence — Example 2

$$P(\mathsf{G}\!=\!\mathsf{fem}) \;=\; P(\mathsf{G}\!=\!\mathsf{mal}) \;=\; 0.5 \qquad\qquad P(\mathsf{P}\!=\!\mathsf{preg}) \;=\; 0.08$$

$$P(\mathsf{S}\!=\!\mathsf{sm}) \;=\; 0.25 \qquad\qquad P(\mathsf{M}\!=\!\mathsf{mar}) \;=\; 0.4$$

- Gender and Smoker are not independent:

$$P(\mathsf{G}\!=\!\mathsf{fem} \mid \mathsf{S}\!=\!\mathsf{sm}) \;=\; 0.44 \;\neq\; 0.5 \;=\; P(\mathsf{G}\!=\!\mathsf{fem})$$

- Gender and Marriage are marginally independent but conditionally dependent given Pregnancy:

$$P(\mathsf{fem}, \mathsf{mar} \mid \overline{\mathsf{preg}}) \approx 0.152 \quad \neq \quad 0.169 \approx P(\mathsf{fem} \mid \overline{\mathsf{preg}}) \cdot P(\mathsf{mar} \mid \overline{\mathsf{preg}})$$

Bayes Theorem

- Product Rule (for events A and B):

$$P(A \cap B) = P(A \mid B)P(B) \quad \text{and} \quad P(A \cap B) = P(B \mid A)P(A)$$

- Equating the right-hand sides:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- For random variables X and Y:

$$\forall x \forall y : \quad P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

- Generalization concerning background knowledge/evidence E:

$$P(Y \mid X, E) = \frac{P(X \mid Y, E)P(Y \mid E)}{P(X \mid E)}$$

Bayes Theorem — Application

$$P(\mathsf{toothache} \mid \mathsf{cavity}) = 0.4$$
$$P(\mathsf{cavity}) = 0.1 \qquad P(\mathsf{cavity} \mid \mathsf{toothache}) = \frac{0.4 \cdot 0.1}{0.05} = 0.8$$
$$P(\mathsf{toothache}) = 0.05$$

Why not estimate $P(\mathsf{cavity} \mid \mathsf{toothache})$ right from the start?

- Causal knowledge like $P(\mathsf{toothache} \mid \mathsf{cavity})$ is more robust than diagnostic knowledge $P(\mathsf{cavity} \mid \mathsf{toothache})$.
- The causality $P(\mathsf{toothache} \mid \mathsf{cavity})$ is independent of the a priori probabilities $P(\mathsf{toothache})$ and $P(\mathsf{cavity})$.
- If $P(\mathsf{cavity})$ rose in a caries epidemic, the causality $P(\mathsf{toothache} \mid \mathsf{cavity})$ would remain constant whereas both $P(\mathsf{cavity} \mid \mathsf{toothache})$ and $P(\mathsf{toothache})$ would increase according to $P(\mathsf{cavity})$.
- A physician, after having estimated $P(\mathsf{cavity} \mid \mathsf{toothache})$, would not know a rule for updating.

Relative Probabilities

Assumption:
We would like to consider the probability of the diagnosis $\mathsf{GumDisease}$ as well.

$$
\begin{aligned}
P(\mathsf{toothache} \mid \mathsf{gumdisease}) &= 0.7 \\
P(\mathsf{gumdisease}) &= 0.02
\end{aligned}
$$

Which diagnosis is more probable?

If we are interested in *relative probabilities* only (which may be sufficient for some decisions), $P(\mathsf{toothache})$ needs not to be estimated:

$$
\begin{aligned}
\frac{P(C \mid T)}{P(G \mid T)} &= \frac{P(T \mid C)P(C)}{P(T)} \cdot \frac{P(T)}{P(T \mid G)P(G)} \\
&= \frac{P(T \mid C)P(C)}{P(T \mid G)P(G)} = \frac{0.4 \cdot 0.1}{0.7 \cdot 0.02} \\
&= 28.57
\end{aligned}
$$

Normalization

If we are interested in the absolute probability of $P(C \mid T)$ but do not know $P(T)$, we may conduct a complete case analysis (according C) and exploit the fact that $P(C \mid T) + P(\neg C \mid T) = 1$.

$$P(C \mid T) = \frac{P(T \mid C)P(C)}{P(T)}$$

$$P(\neg C \mid T) = \frac{P(T \mid \neg C)P(\neg C)}{P(T)}$$

$$1 = P(C \mid T) + P(\neg C \mid T) = \frac{P(T \mid C)P(C)}{P(T)} + \frac{P(T \mid \neg C)P(\neg C)}{P(T)}$$

$$P(T) = P(T \mid C)P(C) + P(T \mid \neg C)P(\neg C)$$

Normalization

- Plugging into the equation for $P(C \mid T)$ yields:

$$P(C \mid T) = \frac{P(T \mid C)P(C)}{P(T \mid C)P(C) + P(T \mid \neg C)P(\neg C)}$$

- For general random variables, the equation reads:

$$P(Y=y \mid X=x) = \frac{P(X=x \mid Y=y)P(Y=y)}{\sum_{\forall y' \in \mathrm{dom}(Y)} P(X=x \mid Y=y')P(Y=y')}$$

- Note the "loop variable" y'. Do not confuse with y.

Multiple Evidences

- The patient complains about a toothache. From this first evidence the dentist infers:

$$P(\mathsf{cavity} \mid \mathsf{toothache}) = 0.8$$

- The dentist palpates the tooth with a metal probe which catches into a fracture:

$$P(\mathsf{cavity} \mid \mathsf{fracture}) = 0.95$$

- Both conclusions might be inferred via Bayes rule. But what does the combined evidence yield? Using Bayes rule further, the dentist might want to determine:

$$P(\mathsf{cavity} \mid \mathsf{toothache} \wedge \mathsf{fracture}) = \frac{P(\mathsf{toothache} \wedge \mathsf{fracture} \mid \mathsf{cavity}) \cdot P(\mathsf{cavity})}{P(\mathsf{toothache} \wedge \mathsf{fracture})}$$

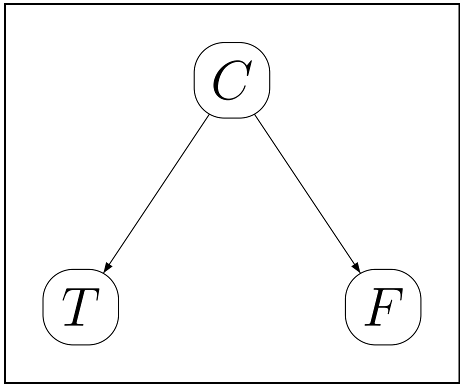
Multiple Evidences

Problem:
He needs $P(\mathsf{toothache} \wedge \mathsf{catch} \mid \mathsf{cavity})$, i. e. diagnostics knowledge for all combinations of symptoms in general. Better incorporate evidences step-by-step:

$$P(Y \mid X, E) = \frac{P(X \mid Y, E) P(Y \mid E)}{P(X \mid E)}$$

Abbreviations:

- C — cavity
- T — toothache
- F — fracture



Objective:
Computing $P(C \mid T, F)$ with just causal statements of the form $P(\,\cdot \mid C)$ and under exploitation of independence relations among the variables.

Multiple Evidences

- A priori: $P(C)$

- Evidence toothache: $P(C \mid T) = P(C)\,\dfrac{P(T \mid C)}{P(T)}$

- Evidence fracture: $P(C \mid T, F) = P(C \mid T)\,\dfrac{P(F \mid C, T)}{P(F \mid T)}$

$$T \perp\!\!\!\perp F \mid C \quad \Leftrightarrow \quad P(F \mid C, T) = P(F \mid C)$$

$$P(C \mid T, F) \;=\; P(C)\,\frac{P(T \mid C)}{P(T)}\,\frac{P(F \mid C)}{P(F \mid T)}$$

Seems that we still have to cope with symptom inter-dependencies?!

Multiple Evidences

- Compound equation from last slide:

$$P(C \mid T, F) \;=\; P(C)\,\frac{P(T \mid C)\;P(F \mid C)}{P(T)\;P(F \mid T)}$$

$$\;=\; P(C)\,\frac{P(T \mid C)\;P(F \mid C)}{P(F, T)}$$

- $P(F, T)$ is a normalizing constant and can be computed if $P(F \mid \neg C)$ and $P(T \mid \neg C)$ are known:

$$P(F, T) = \underbrace{P(F, T \mid C)}_{P(F|C)P(T|C)}\;P(C) + \underbrace{P(F, T \mid \neg C)}_{P(F|\neg C)P(T|\neg C)}\;P(\neg C)$$

- Therefore, we finally arrive at the following solution...

Multiple Evidences

$$P(C \mid F, T) = \frac{\boxed{P(C)}\ \boxed{P(T \mid C)}\ \boxed{P(F \mid C)}}{\boxed{P(F \mid C)}\ \boxed{P(T \mid C)}\ \boxed{P(C)} + \boxed{P(F \mid \neg C)}\ \boxed{P(T \mid \neg C)}\ \boxed{P(\neg C)}}$$

Note that we only use causal probabilities $P(\ \cdot \mid C)$ together with the a priori (marginal) probabilities $P(C)$ and $P(\neg C)$.

Multiple Evidences — Summary

Multiple evidences can be treated by reduction on

- a priori probabilities
- (causal) conditional probabilities for the evidence
- under assumption of conditional independence

General rule:

$$P(Z \mid X, Y) \;=\; \alpha\, P(Z)\, P(X \mid Z)\, P(Y \mid Z)$$

for X and Y conditionally independent given Z and with normalizing constant α.

Monty Hall Puzzle

Marylin Vos Savant in her riddle column in the New York Times:

You are a candidate in a game show and have to choose between three doors. Behind one of them is a Porsche, whereas behind the other two there are goats. After you chose a door, the host Monty Hall (who knows what is behind each door) opens another (not your chosen one) door with a goat. Now you have the choice between keeping your chosen door or choose the remaining one.

Which decision yields the best chance of winning the Porsche?

Monty Hall Puzzle

G	You win the Porsche.
R	You revise your decision.
A	Behind your initially chosen door is (and remains) the Porsche.

$$
\begin{aligned}
P(G \mid R) &= P(G, A \mid R) + P(G, \overline{A} \mid R) \\
&= P(G \mid A, R) P(A \mid R) + P(G \mid \overline{A}, R) P(\overline{A} \mid R) \\
&= 0 \cdot P(A \mid R) + 1 \cdot P(\overline{A} \mid R) \\
&= P(\overline{A} \mid R) = P(\overline{A}) = \frac{2}{3}
\end{aligned}
$$

$$
\begin{aligned}
P(G \mid \overline{R}) &= P(G, A \mid \overline{R}) + P(G, \overline{A} \mid \overline{R}) \\
&= P(G \mid A, \overline{R}) P(A \mid \overline{R}) + P(G \mid \overline{A}, \overline{R}) P(\overline{A} \mid \overline{R}) \\
&= 1 \cdot P(A \mid \overline{R}) + 0 \cdot P(\overline{A} \mid \overline{R}) \\
&= P(A \mid \overline{R}) = P(A) = \frac{1}{3}
\end{aligned}
$$

Simpson's Paradox

Example: C = Patient takes medication, E = patient recovers

	E	$\neg E$	$\sum$	Recovery rate
C	20	20	40	50%
$\neg C$	16	24	40	40%
$\sum$	36	44	80	

Men	E	$\neg E$	$\sum$	Rec.rate	Women	E	$\neg E$	$\sum$	Rec.rate
C	18	12	30	60%	C	2	8	10	20%
$\neg C$	7	3	10	70%	$\neg C$	9	21	30	30%
	25	15	40			11	29	40	

$$\begin{array}{rrcl} & P(E \mid C) & > & P(E \mid \neg C) \\ \text{but} & P(E \mid C, M) & < & P(E \mid \neg C, M) \\ & P(E \mid C, W) & < & P(E \mid \neg C, W) \end{array}$$

Probabilistic Reasoning

- Probabilistic reasoning is difficult and may be problematic:
  - $P(A \wedge B)$ is not determined simply by $P(A)$ and $P(B)$:
    $P(A) = P(B) = 0.5 \quad \Rightarrow \quad P(A \wedge B) \in [0, 0.5]$
  - $P(C \mid A) = x, P(C \mid B) = y \quad \Rightarrow \quad P(C \mid A \wedge B) \in [0, 1]$
    Probabilistic logic is *not truth functional*!
- Central problem: How does additional information affect the current knowledge?
  I. e., if $P(B \mid A)$ is known, what can be said about $P(B \mid A \wedge C)$?
- High complexity: n propositions $\rightarrow 2^n$ full conjunctives
- Hard to specify these probabilities.

Summary

- Uncertainty is inevitable in complex and dynamic scenarios that force agents to cope with ignorance.
- Probabilities express the agent's inability to vote for a definitive decision. They model the degree of belief.
- If an agent violates the axioms of probability, it may exhibit irrational behavior in certain circumstances.
- The Bayes rule is used to derive unknown probabilities from present knowledge and new evidence.
- Multiple evidences can be effectively included into computations exploiting conditional independencies.

Probabilistic Causal Networks

The Big Objective(s)

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

1. **How can (expert) knowledge of complex domains be efficiently represented?**

2. **How can inferences be carried out within these representations?**

3. **How can such representations be (automatically) extracted from collected data?**

We will deal with all three questions during the lecture.

Example 1: Planning in car manufacturing

Available information

- "Engine type e_1 can only be combined with transmission t_2 or t_5."
- "Transmission t_5 requires crankshaft c_2."
- "Convertibles have the same set of radio options as SUVs."

Possible questions/inferences:

- "Can a station wagon with engine e_4 be equipped with tire set y_6?"
- "Supplier S_8 failed to deliver on time. What production line has to be modified and how?"
- "Are there any peculiarties within the set of cars that suffered an aircondition failure?"

Example 2: Medical reasoning

Available information:

- “Malaria is much less likely than flu.”
- “Flu causes cough and fever.”
- “Nausea can indicate malaria as well as flu.”
- “Nausea never indicated pneunomia before.”

Possible questions/inferences

- “The patient has fever. How likely is he to have malaria?”
- “How much more likely does flu become if we can exclude malaria?”

Common Problems

Both scenarios share some severe problems:

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

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

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

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

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

Idea to Solve the Problems

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

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

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

Example: Car Manufacturing

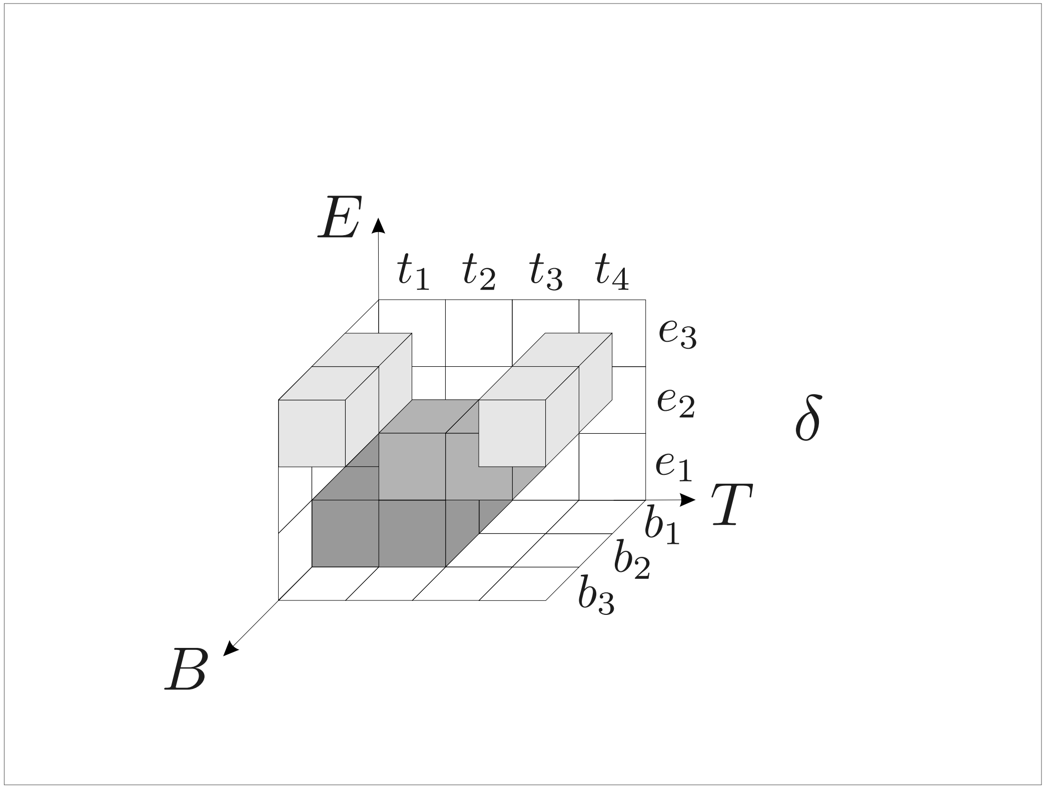
- Let us consider a car configuration is described by three attributes:
  - Engine E, $\mathrm{dom}(E) = \{e_1, e_2, e_3\}$
  - Breaks B, $\mathrm{dom}(B) = \{b_1, b_2, b_3\}$
  - Tires T, $\mathrm{dom}(T) = \{t_1, t_2, t_3, t_4\}$
- Therefore the set of all (theoretically) possible car configurations is:

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

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

Example: Car Manufacturing

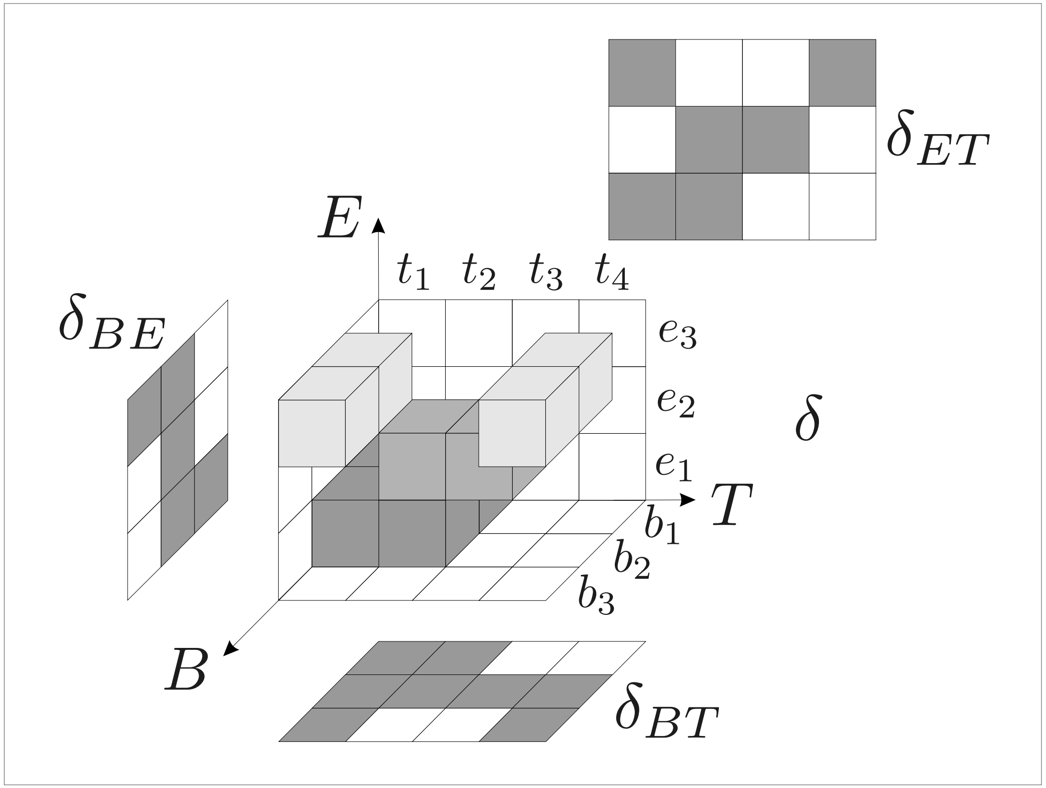
Possible car configurations



- Every cube designates a valid value combination.
- 10 car configurations in our model.
- Different colors are intended to distinguish the cubes only.

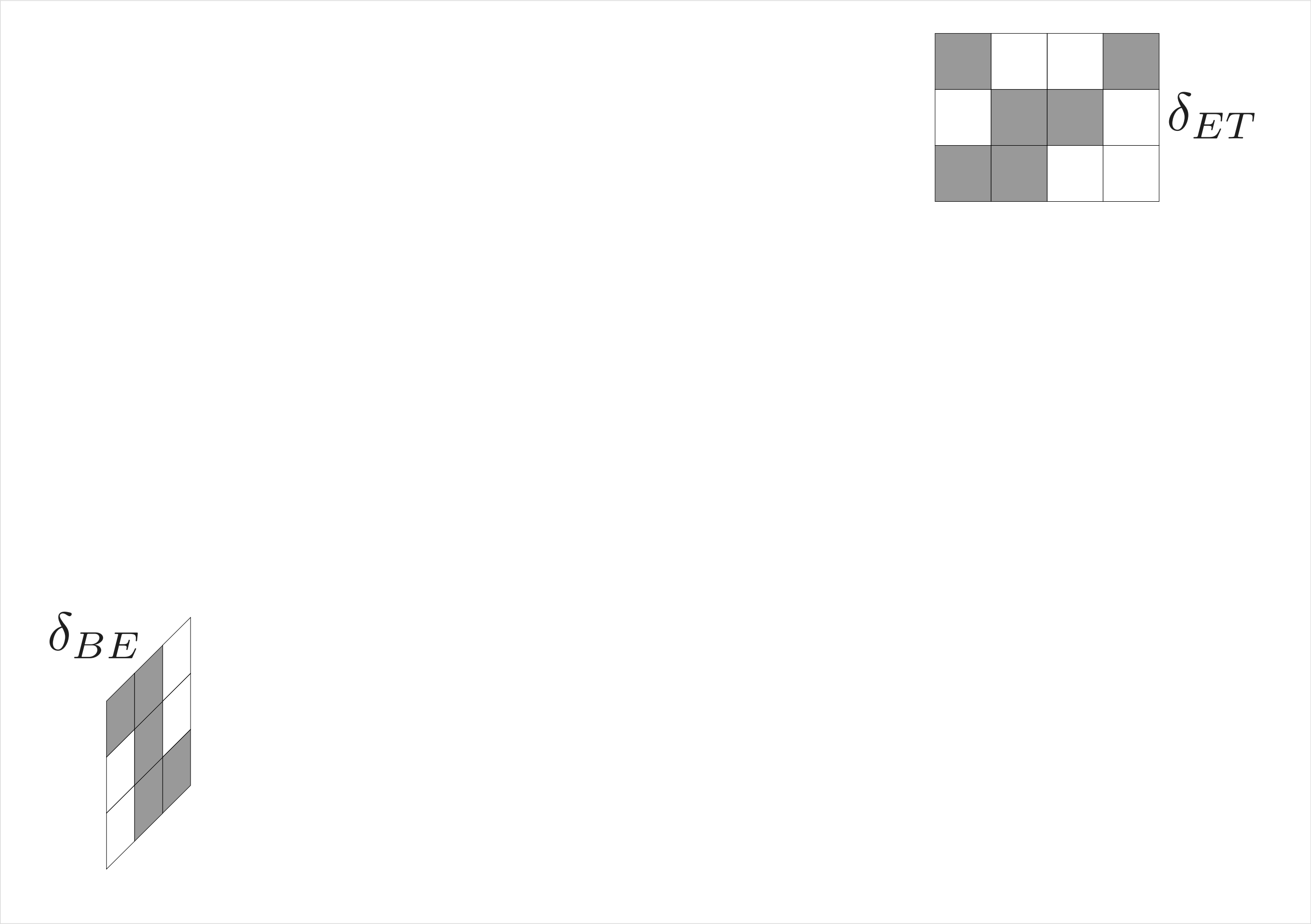
Example

2-D projections

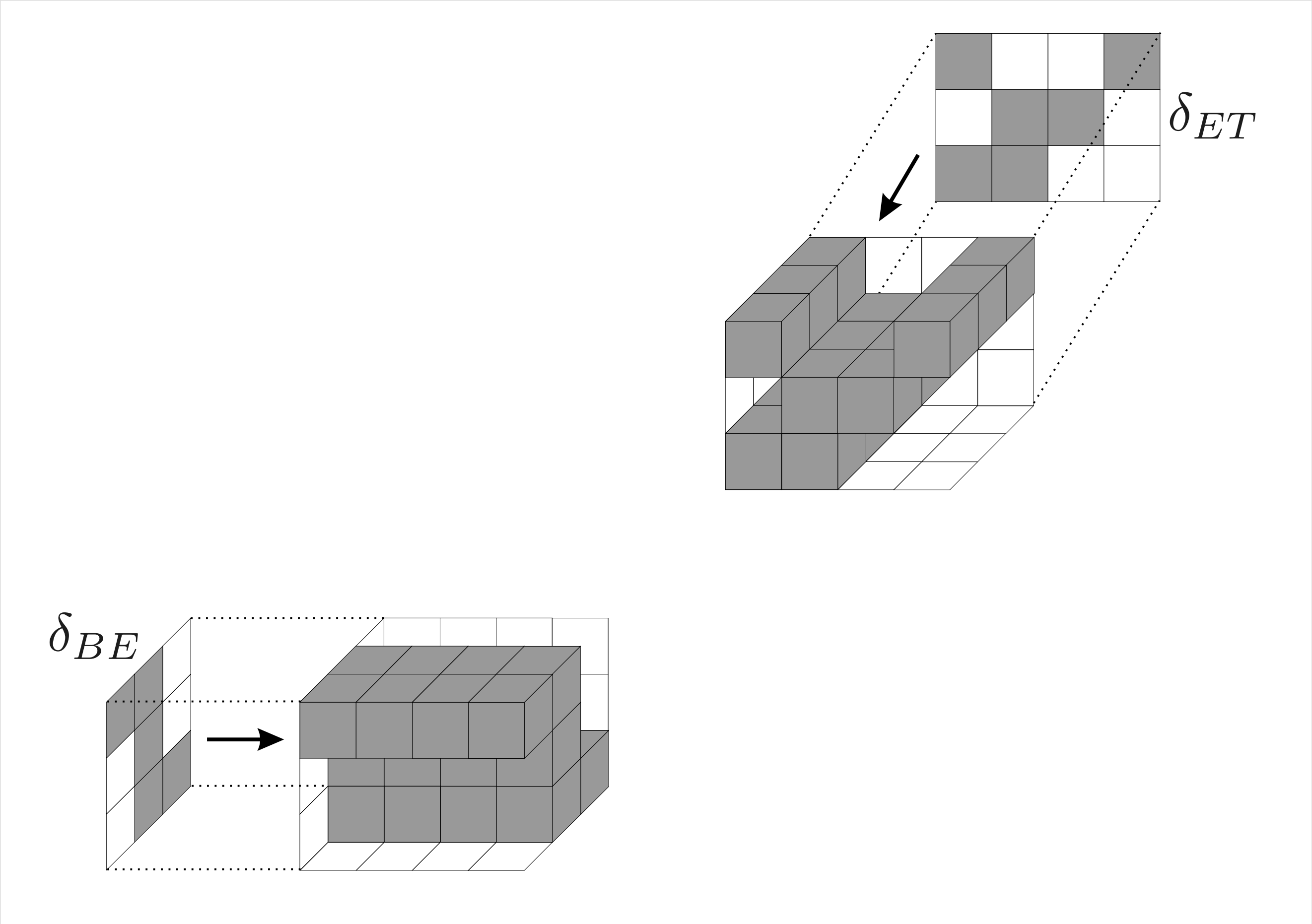


- Is it possible to reconstruct δ from the δ_i?

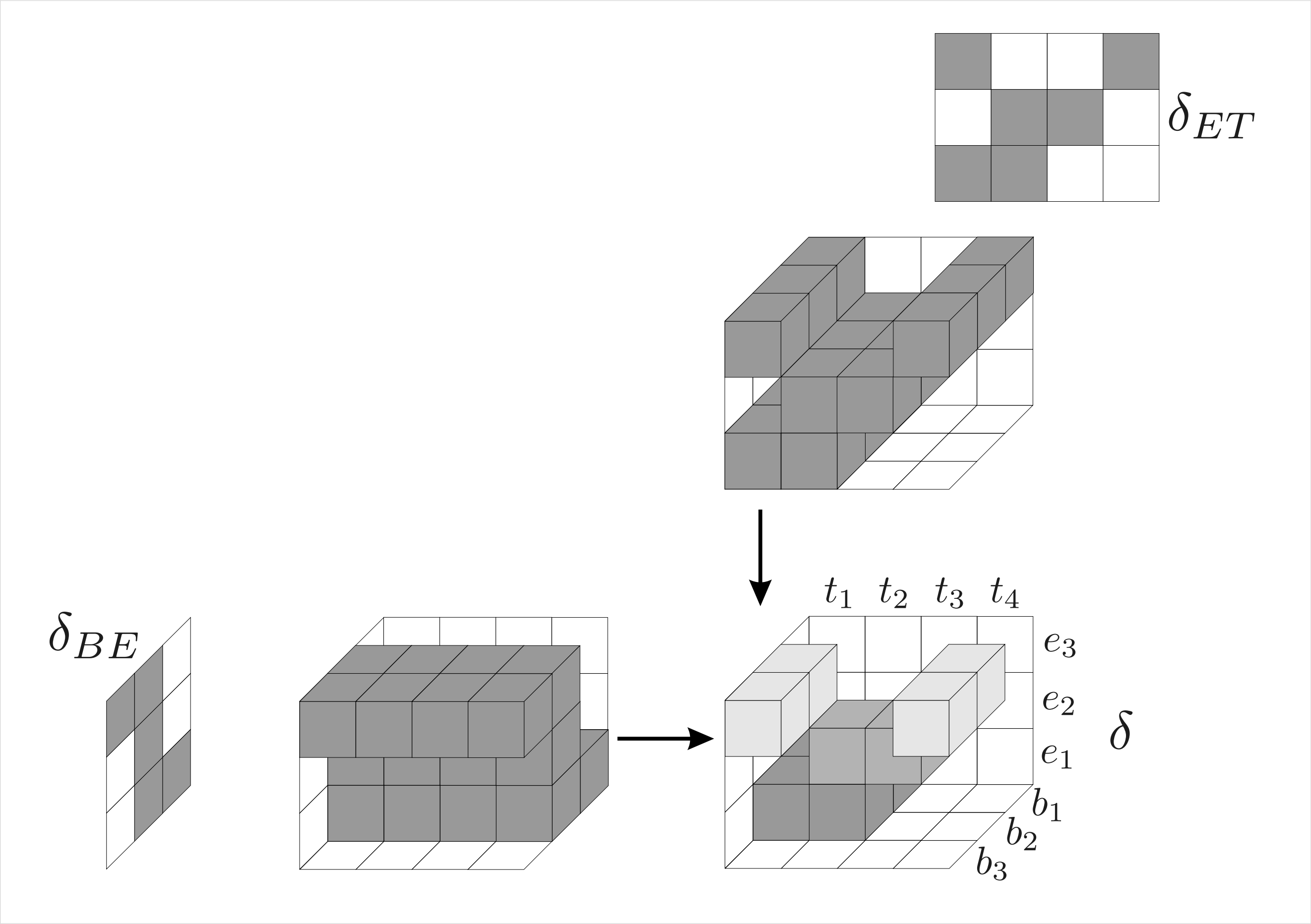
Example: Reconstruction of δ with δ_{BE} and δ_{ET}



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



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

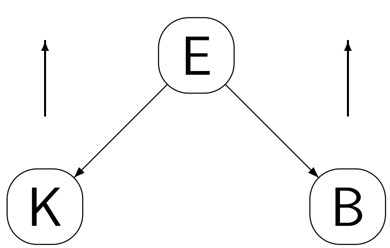


Example — Qualitative Aspects

- Lecture theatre in winter: Waiting for Mr. K and Mr. B.
  Not clear whether there is ice on the roads.

- 3 variables:
  - E road condition: $\mathrm{dom}(\mathsf{E}) = \{\mathsf{ice}, \neg\mathsf{ice}\}$
  - K K had an accident: $\mathrm{dom}(\mathsf{K}) = \{\mathsf{yes}, \mathsf{no}\}$
  - B B had an accident: $\mathrm{dom}(\mathsf{B}) = \{\mathsf{yes}, \mathsf{no}\}$

- Ignorance about these states is modelled via the observer's belief.

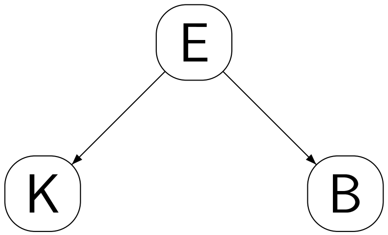


↓ E influences K and B
(the more ice the more accidents)

↑ Knowledge about accident increases belief in ice

Example

A priori knowledge	Evidence	Inferences
E unknown	B has accident	$\Rightarrow$ E $=$ ice more likely
		$\Rightarrow$ K has accident more likely
E $=$ $\neg$ice	B has accident	$\Rightarrow$ no change in belief about E
		$\Rightarrow$ no change in belief about accident of K
E unknown		K and B dependent
E known		K and B independent



Causal Dependence vs. Reasoning

Rule: A entails B with certainty x: $\boxed{A \xrightarrow{x} B}$

- **Deduction** ($\rightarrow$):
  A and $A \xrightarrow{x} B$, therefore B more likely as effect (causality)

- **Abduction** ($\leftarrow$):
  B and $A \xrightarrow{x} B$, therefore A more likely as cause (no causality)

For this reason, the notion "dependency model" is to be preferred to "causal network".

Objective

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

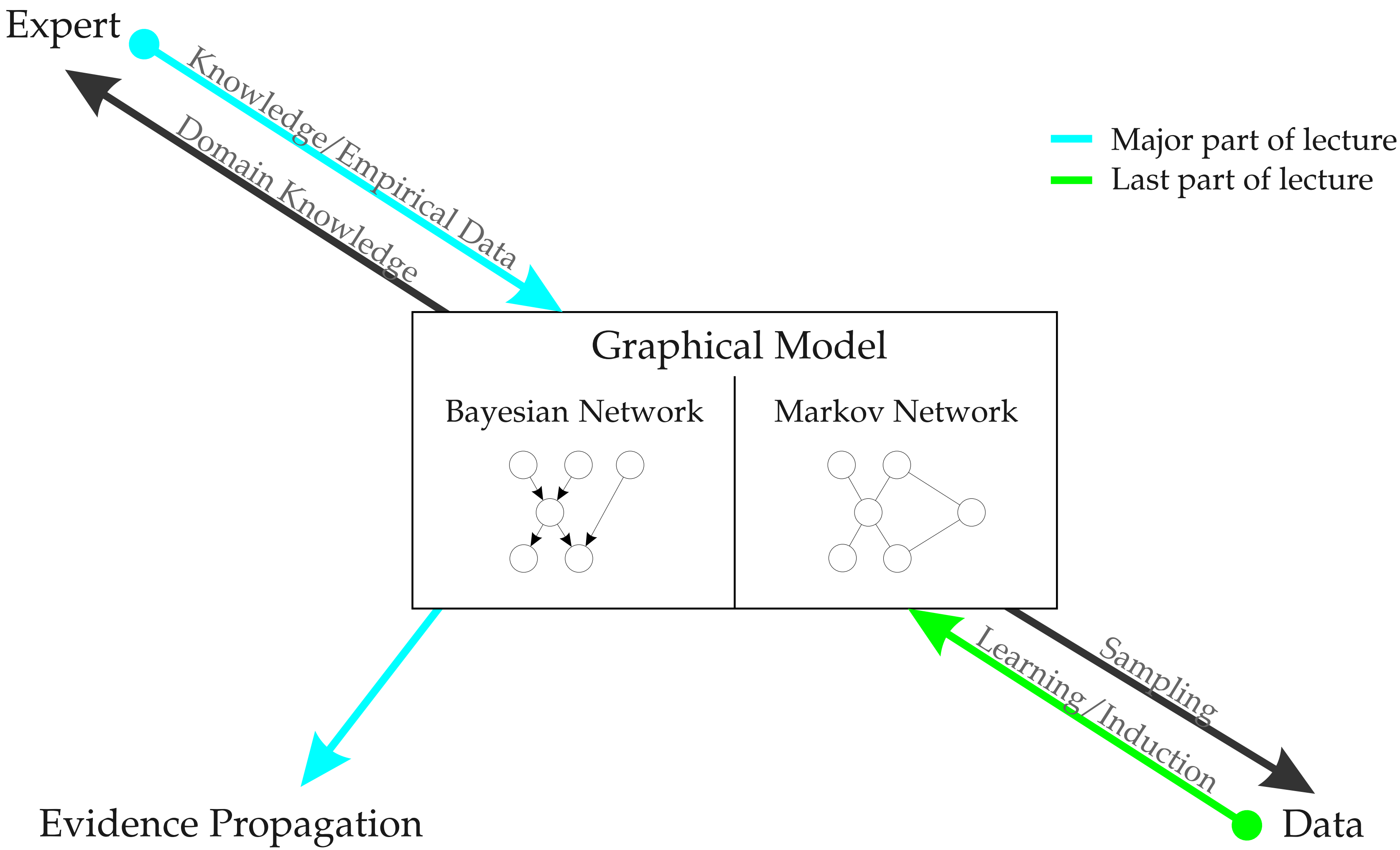
- The collective size of those sub-structures is much smaller than that of the original distribution P.

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

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

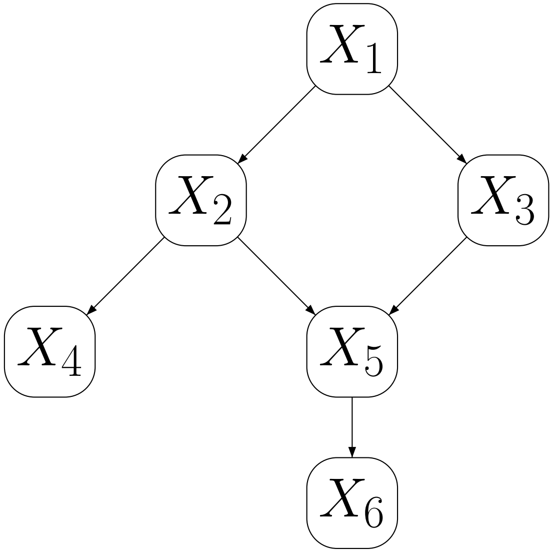
where c_i is an instantiation of C_i and $\Psi_i(c_i) \in \mathbb{R}^+$ a *factor potential.*

The Big Picture / Lecture Roadmap



Probabilistic Causal Networks

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

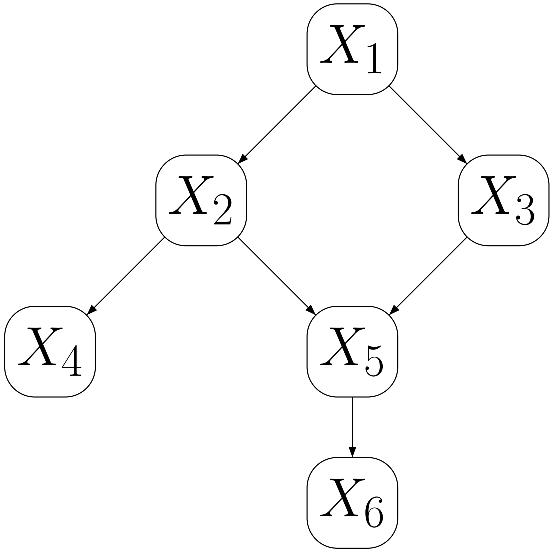


In general (according chain rule):

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

Probabilistic Causal Networks

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



According graph (independence structure):

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

Formal Framework

Nomenclature for the next slides:

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

Formal Framework

- Let Ω^* be the real universe of objects under consideration (e. g. population of people, collection of cars, customer transactions, etc.). Then the random vector $\vec{X} = (X_1, \ldots, X_n)$ *describes* each element $\omega^* \in \Omega^*$ in terms of the universe of discourse Ω:

$$\vec{X} : \Omega^* \to \Omega \quad \text{with} \quad \vec{X}(\omega^*) = (X_1(\omega^*), \ldots, X_n(\omega^*))$$

- If $(\Omega^*, \mathcal{E}, Q)$ is an intrinsic probability space acting in the background, then it induces — in combination with $\vec{X}$ — a probability measure P over Ω:

$$\begin{aligned}
&\forall (x_1, \ldots, x_n) \in \Omega : \\
&\qquad P(\{(x_1, \ldots, x_n)\}) = P(X_1 = x_1, \ldots, X_n = x_n) \\
&\qquad\qquad = Q(\{\omega^* \in \Omega^* \mid \bigwedge_{i=1}^{n} X_i = x_i\})
\end{aligned}$$

Formal Framework

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

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

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

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

Constructing a DAG

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

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

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

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

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

return G

Belief Network

- A *Belief Network* (V, E, P) consists of a set $V = \{X_1, \dots, X_n\}$ of random variables and a set E of directed edges between the variables.
- Each variable has a finite set of mutual exclusive and collectively exhaustive states.
- The variables in combination with the edges form a directed, acyclich graph.
- Each variable with parent nodes $B_1, \dots, B_m$ is assigned a potential table $P(A \mid B_1, \dots, B_m)$.
- Note, that the connections between the nodes not necessarily express a causal relationship.
- For every belief network, the following equation holds:

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

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

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

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

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

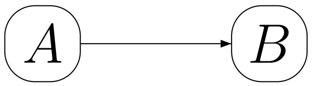
- A and B constitute random variables w. r. t. $(\Omega^*, \mathcal{E}, Q)$.

$$\Omega = \Omega_A \times \Omega_B \qquad A : \Omega^* \to \Omega_A, \quad B : \Omega^* \to \Omega_B$$

- A, B and $(\Omega^*, \mathcal{E}, Q)$ induce the probability space $(\Omega, 2^\Omega, P)$ with

$$P\Big(\{(a, b)\}\Big) = Q\Big(\{\omega^* \in \Omega^* \mid A(\omega^*) = a \wedge B(\omega^*) = b\}\Big) :$$

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



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

We are dealing with a belief network.

Example

Choice of universe of discourse

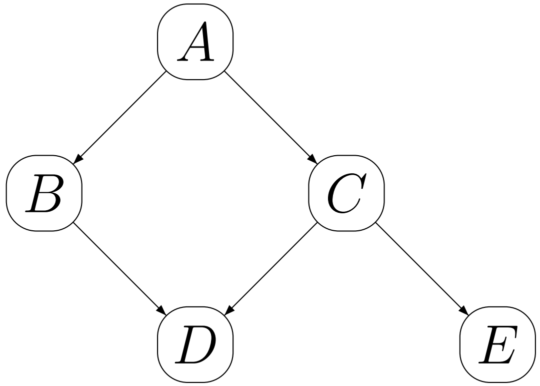
	Variable	Domain
A	metastatic cancer	$\{a_1, a_2\}$
B	increased serum calcium	$\{b_1, b_2\}$
C	brain tumor	$\{c_1, c_2\}$
D	coma	$\{d_1, d_2\}$
E	headache	$\{e_1, e_2\}$

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

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

$$|\Omega| = 32$$

Analysis of dependencies



Choice of probability parameters

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

Shorthand notation

- 11 values to store instead of 31
- Consult experts, textbooks, case studies, surveys, etc.

Calculation of conditional probabilities

Calculation of marginal probabilities

Crux of the Matter

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

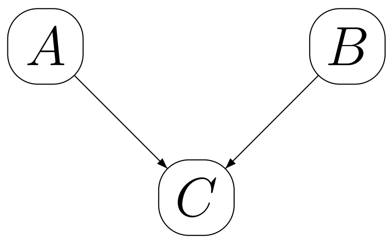
- Computational complexities
  $\rightarrow$ exploit independencies

Problem:

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

- How can we determine $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$ solely using the graph structure?

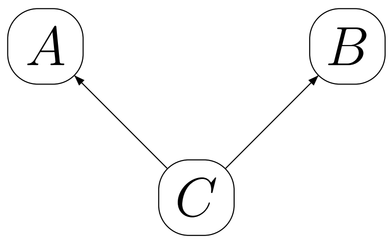
Converging Connection



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

- If C is not instantiated (i. e., no value specified/observed), A and B are marginally independent.
- After instantiation (observation) of C the variables A and B become conditionally dependent given C.
- Evidence can only be transferred over a converging connection if the variable in between (or one of its successors) is initialized.

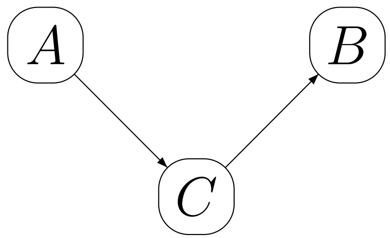
Diverging Connection



Diagnosis	
A	body temperature
B	cough
C	disease

- If C is unknown, knowledge about A ist relevant for B and vice versa, i. e. A and B are marginally dependent.
- However, if C is observed, A and B become conditionally independent given C.
- A influences B via C. If C is known it in a way blocks the information from flowing from A to B, thus rendering A and B (conditionally) independent.

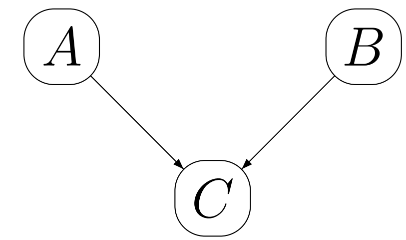
Serial Connection



Accidents	
A	rain
B	accident risk
C	road conditions

- Analog scenario to case 2
- A influences C and C influences B. Thus, A influences B.
  If C is known, it blocks the path between A and B.

Formal Representation



Converging Connection: Marginal Independence

- Decomposition according to graph:

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

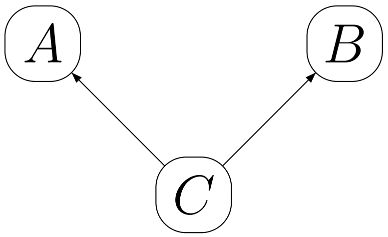
- Embedded Independence:

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

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

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

Formal Representation



Diverging Connection: Conditional Independence

- Decomposition according to graph:

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

- Embedded Independence:

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

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

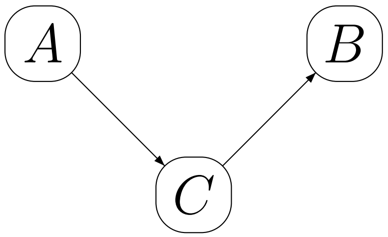
- Alternative derivation:

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

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

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

Serial Connection: Conditional Independence



- Decomposition according to graph:

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

- Embedded Independence:

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

Formal Representation

Trivial Cases:

- Marginal Independence:

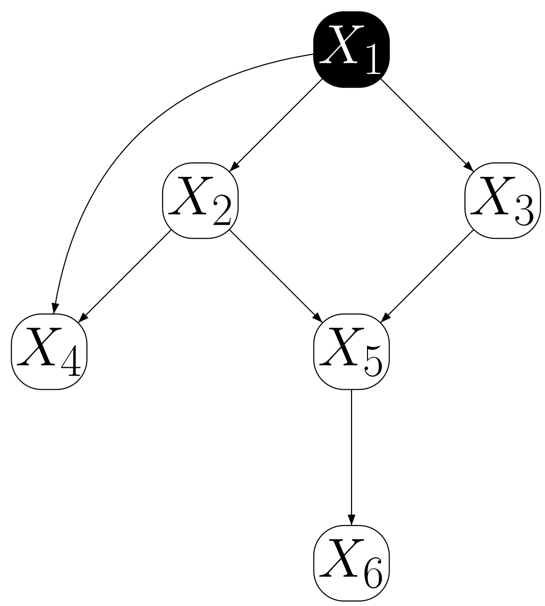
$$A \qquad\qquad B \qquad\qquad P(A, B) \;=\; P(A) \cdot P(B)$$

- Marginal Dependence:

$$A \longrightarrow B \qquad\qquad P(A, B) \;=\; P(B \mid A) \cdot P(A)$$

Question: Are X_2 and X_3 independent given X_1?

evidence $X_1 = x_1$



d-Separation

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

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

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

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

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

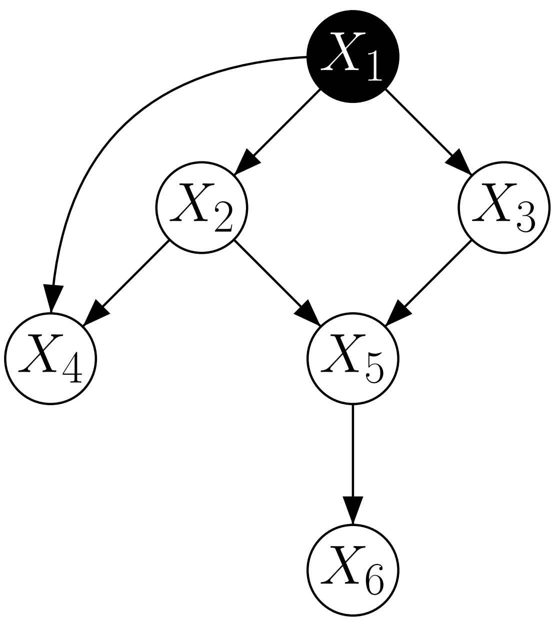
d) Two edges meeting head-to-head in Y are blocked if neither Y nor its successors are in S.

Relation to Conditional independence

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

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

Application to the previous example:



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

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

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

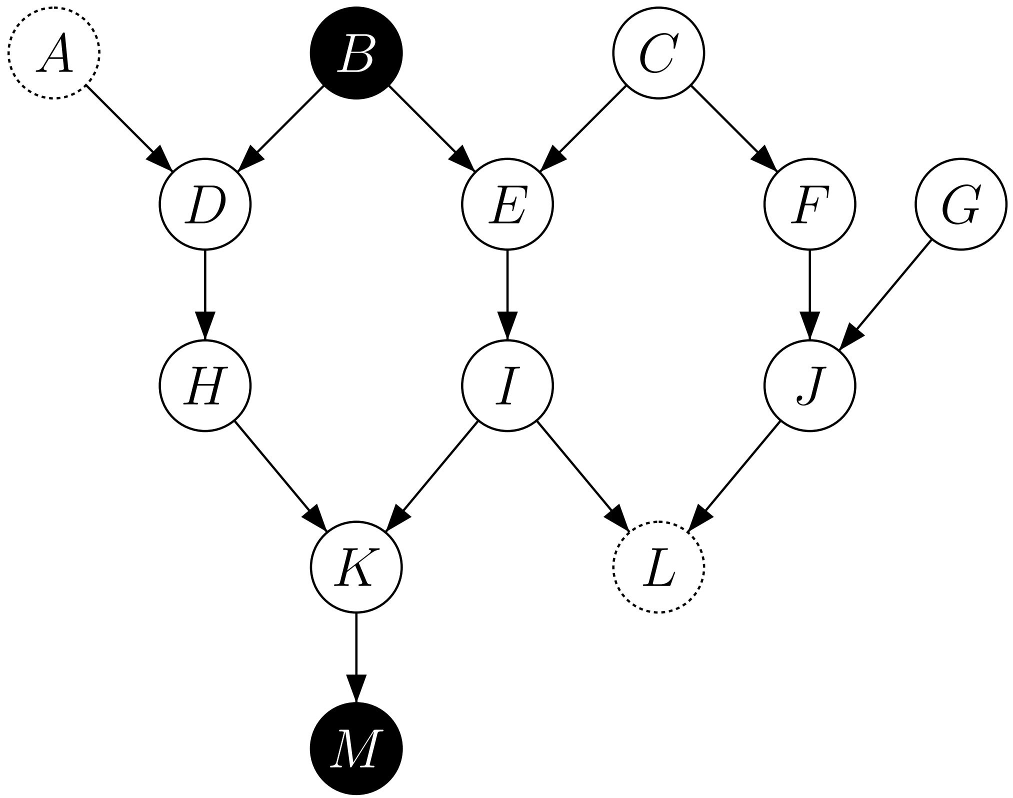
π_3 $X_4 \leftarrow X_1 \rightarrow X_3$ tail-to-tail
$X_2 \rightarrow X_4 \leftarrow X_1$ head-to-head
both connections are blocked $\;\Rightarrow\; \pi_3$ is blocked

Example (cont.)

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

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

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



Are A and L conditionally independent given $\{B, M\}$?

Algebraic structure of CI statements

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

Repetition

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

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

Often, the following (equivalent) notation is used:

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

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

(Semi-)Graphoid-Axioms

Let $(\Omega, \mathcal{E}, P)$ be a probability space and W, X, Y and Z four disjoint subsets of random variables (over Ω). Then the propositions

a) Symmetry: $(X \perp\!\!\!\perp_P Y \mid Z) \Rightarrow (Y \perp\!\!\!\perp_P X \mid Z)$

b) Decomposition: $(W \cup X \perp\!\!\!\perp_P Y \mid Z) \Rightarrow (W \perp\!\!\!\perp_P Y \mid Z) \wedge (X \perp\!\!\!\perp_P Y \mid Z)$

c) Weak Union: $(W \cup X \perp\!\!\!\perp_P Y \mid Z) \Rightarrow (X \perp\!\!\!\perp_P Y \mid Z \cup W)$

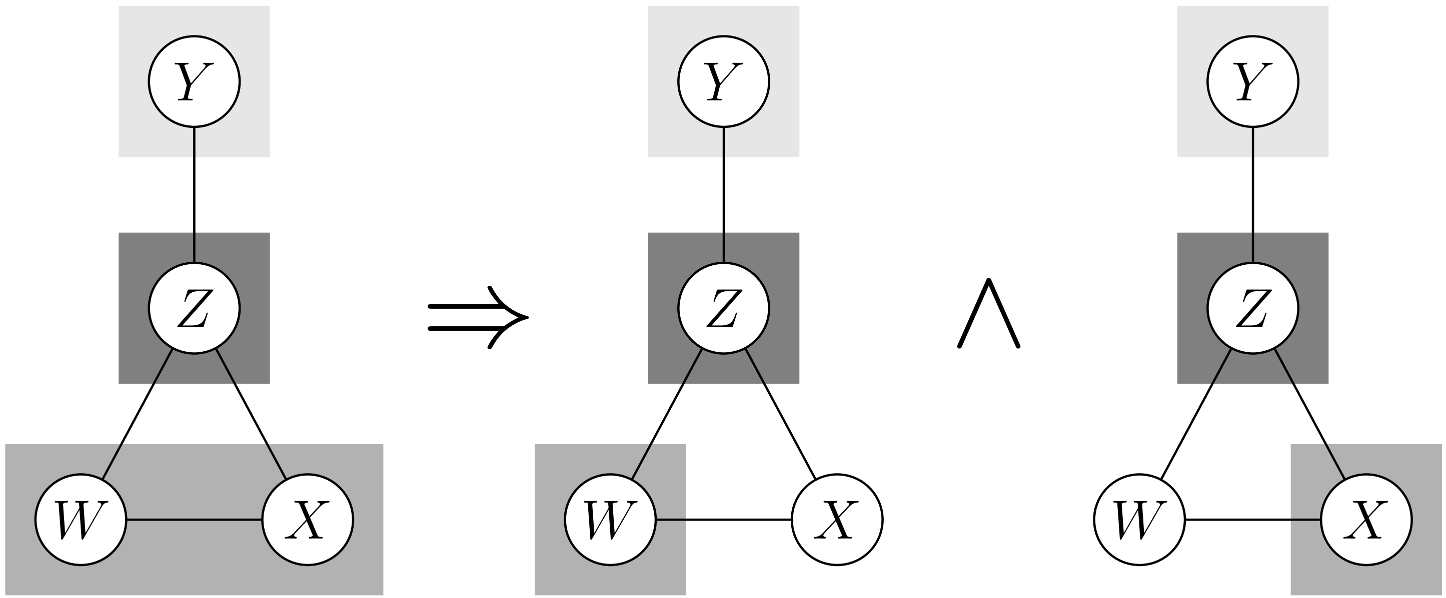
d) Contraction: $(X \perp\!\!\!\perp_P Y \mid Z \cup W) \wedge (W \perp\!\!\!\perp_P Y \mid Z) \Rightarrow (W \cup X \perp\!\!\!\perp_P Y \mid Z)$

are called the **Semi-Graphoid Axioms**. The above propositions and

e) Intersection: $(W \perp\!\!\!\perp_P Y \mid Z \cup X) \wedge (X \perp\!\!\!\perp_P Y \mid Z \cup W) \Rightarrow (W \cup X \perp\!\!\!\perp_P Y \mid Z)$

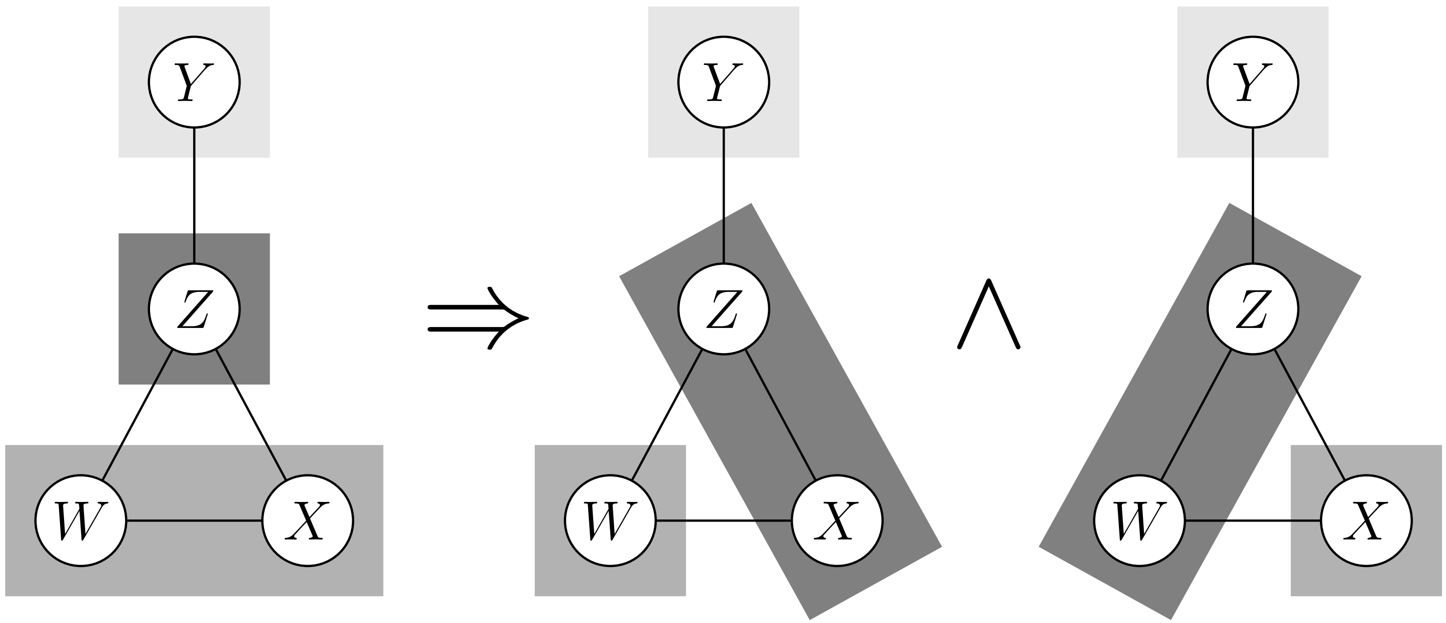
are called the **Graphoid Axioms**.

Decomposition



Drawings adapted from [Castillo *et al.* 1997].

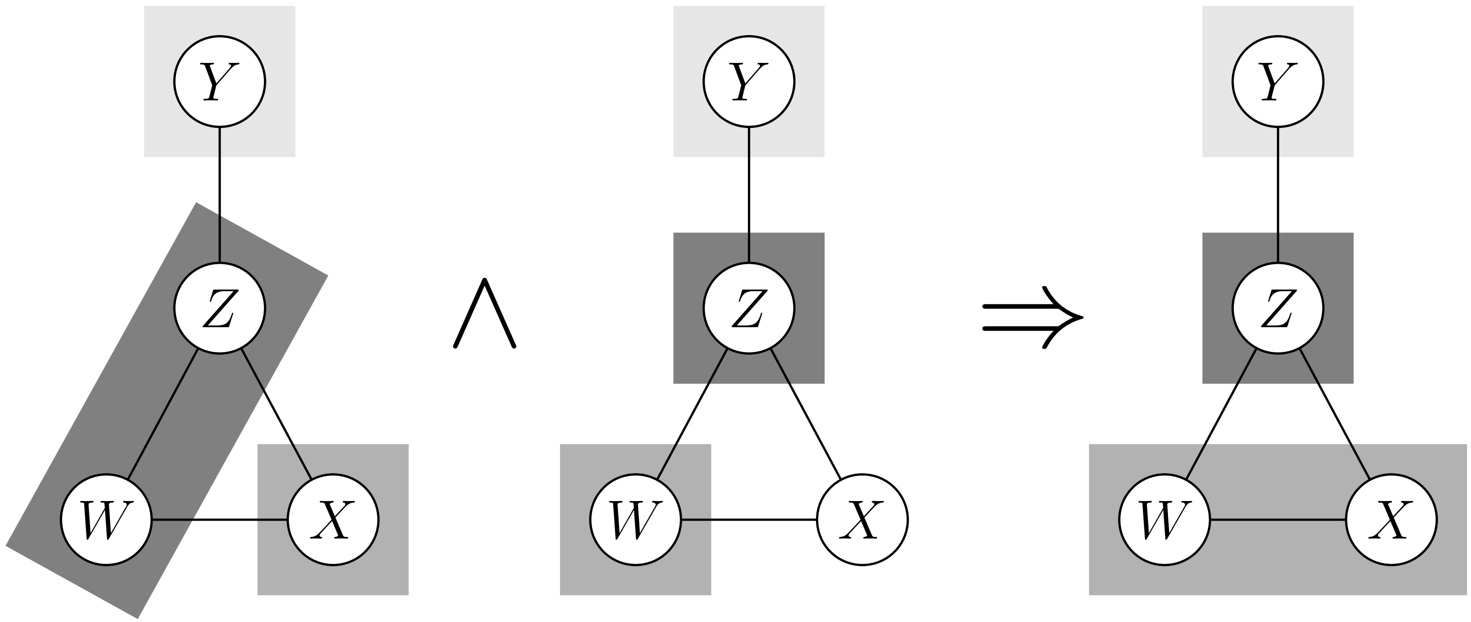
Weak Union



Learning irrelevant information W cannot render irrelevant information X relevant.

Drawings adapted from [Castillo *et al.* 1997].

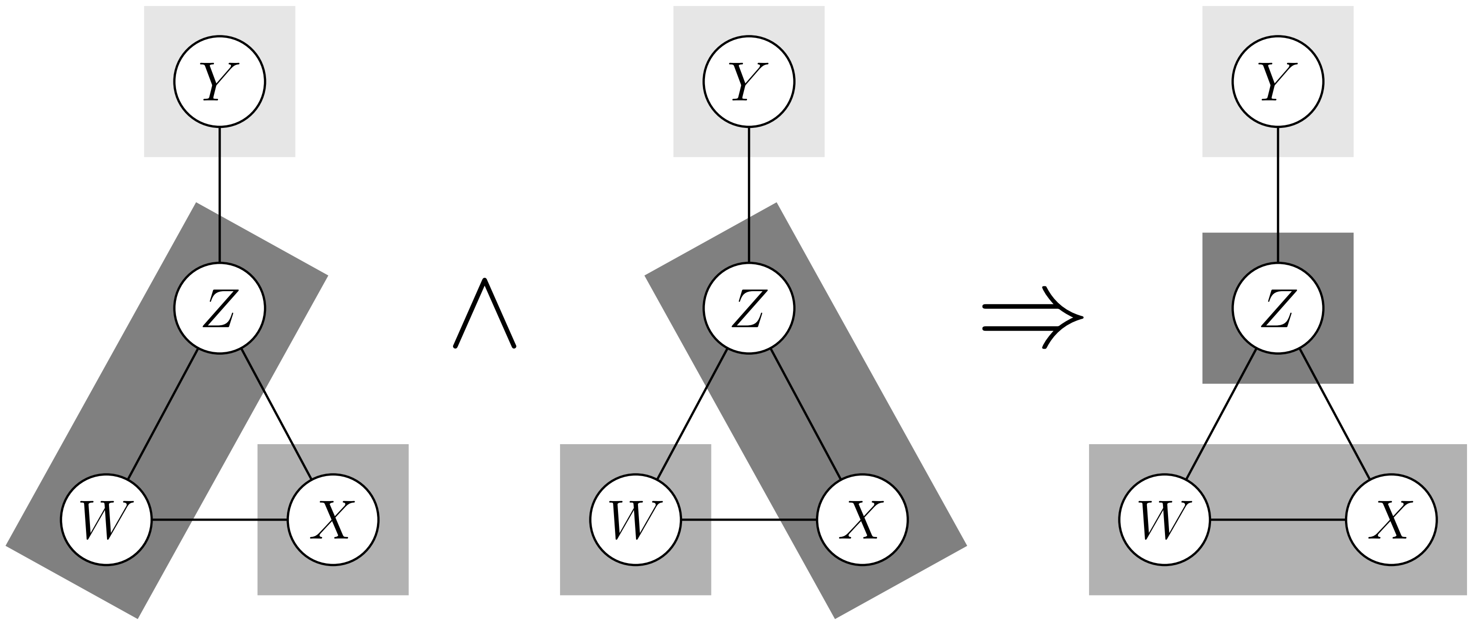
Contraction



If X is irrelevant (to Y) after having learnt some irrelevant information W, then X must have been irrelevant before.

Drawings adapted from [Castillo *et al.* 1997].

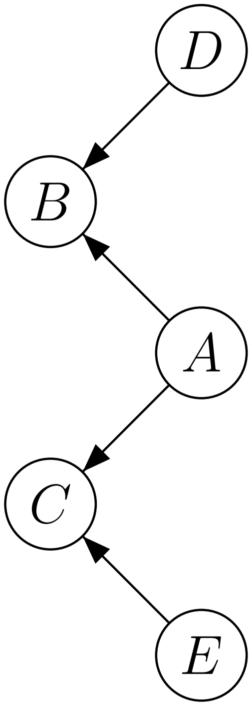
Intersection



Unless W affects Y when X is known or X affects Y when W is known, neither X nor W nor their combination can affect Y.

Drawings adapted from [Castillo *et al.* 1997].

Example



Proposition: $B \perp\!\!\!\perp C \mid A$

Proof: $D \perp\!\!\!\perp A, C \mid \emptyset \quad \wedge \quad B \perp\!\!\!\perp C \mid A, D$

$\stackrel{\text{w. union}}{\Longrightarrow} \quad D \perp\!\!\!\perp C \mid A \quad \wedge \quad B \perp\!\!\!\perp C \mid A, D$

$\stackrel{\text{symm.}}{\Longleftrightarrow} \quad C \perp\!\!\!\perp D \mid A \quad \wedge \quad C \perp\!\!\!\perp B \mid A, D$

$\stackrel{\text{contr.}}{\Longrightarrow} \quad C \perp\!\!\!\perp B, D \mid A$

$\stackrel{\text{decomp.}}{\Longrightarrow} \quad C \perp\!\!\!\perp B \mid A$

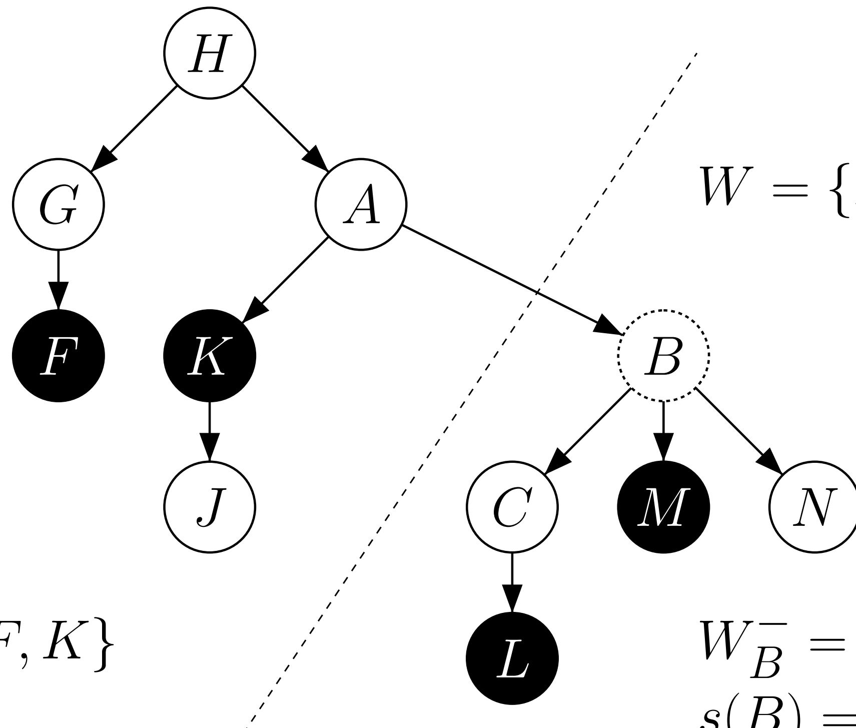
$\stackrel{\text{symm.}}{\Longleftrightarrow} \quad B \perp\!\!\!\perp C \mid A$

Propagation in Belief Networks

Objective

- **Given:** Belief network (V, E, P) with tree structure and $P(V) > 0$.
  Set $W \subseteq V$ of instantiated variables where
  a priori knowledge $W \neq \emptyset$ is allowed
- **Desired:** $P(B \mid W)$ for all $B \in V$
- **Notation:**

W_B^-	subset of those variables of W that belong to the subtree of (V, E) that has root B
W_B^+	$= W \setminus W_B^-$
$s(B)$	set of direct successors of B
Ω_B	domain of B
b^*	value that B is instantiated with



$W = \{F, K, L, M\}$

$W_B^+ = \{F, K\}$

$W_B^- = \{L, M\}$

$s(B) = \{C, M, N\}$

Example

$$
\begin{aligned}
P(B = b \mid W) &= P(b \mid W_B^- \cup W_B^+) \qquad \text{with } B \notin W \\
&= \frac{P(W_B^- \cup W_B^+ \mid b)\; P(b)}{P(W_B^- \cup W_B^+)} \\
&= \frac{P(W_B^- \mid b)\; P(W_B^+ \mid b)\; P(b)}{P(W_B^- \cup W_B^+)} \\
&= \frac{P(W_B^- \mid b)\; P(b \mid W_B^+)}{P(W_B^- \cup W_B^+)\; P(W_B^+)} \\
&= \beta_{B,W} \quad \underbrace{P(W_B^- \mid b)}_{\text{Evidence from ''`below''}} \quad \underbrace{P(b \mid W_B^+)}_{\text{Evidence from ''`above''}}
\end{aligned}
$$

Example

Since we ignore the constant $\beta_{B,W}$ for the derivations below, the following designations are used instead of $P(\cdot)$:

π-values and λ-values

Let $B \in V$ be a variable and $b \in \Omega_B$ a value of its domain. We define the π- and λ-values as follows:

$$\lambda(b) = \begin{cases} P(W_B^- \mid b) & \text{if } B \notin W \\ \quad 1 & \text{if } B \in W \wedge b^* = b \\ \quad 0 & \text{if } B \in W \wedge b^* \neq b \end{cases}$$

$$\pi(b) = P(b \mid W_B^+)$$

Example

$$\lambda(b) = \prod_{C \in s(B)} P(W_C^- \mid b) \qquad \text{if } B \in W$$

$$\lambda(b) = 1 \qquad \text{if } B \text{ leaf in } (V, E)$$

$$\pi(b) = P(b) \qquad \text{if } B \text{ root in } (V, E)$$

$$P(b \mid W) = \alpha_{B,W} \cdot \lambda(b) \cdot \pi(b)$$

λ-message

Let $B \in V$ be an attribute and $C \in s(B)$ its direct children with the respective domains $\mathrm{dom}(B) = \{B_1, \ldots, b_i, \ldots, b_k\}$ and $\mathrm{dom}(C) = \{c_1, \ldots, c_j, \ldots, c_m\}$.

$$\lambda_{C \to B}(b_i) \overset{\text{Def}}{=} \sum_{j=1}^{m} P(c_j \mid b_i) \cdot \lambda(c_j), \qquad i = 1, \ldots, k$$

The vector

$$\vec{\lambda}_{C \to B} \overset{\text{Def}}{=} \Big(\lambda_{C \to B}(b_i)\Big)_{i=1}^{k}$$

is called λ-*message* from C to B.

Example

Let $B \in V$ an attribute an $b \in \mathrm{dom}(B)$ a value of its domain.
Then

$$\lambda(b) = \begin{cases} \rho_{B,W} \cdot \prod\limits_{C \in s(B)} \lambda_C(b) & \text{if } B \notin W \\ 1 & \text{if } B \in W \wedge b = b^* \\ 0 & \text{if } B \in W \wedge b \neq b^* \end{cases}$$

with $\rho_{B,W}$ being a positive constant.

π-message

Let $B \in V$ be a non-root node in (V, E) and $A \in V$ its parent with domain $\mathrm{dom}(A) = \{a_1, \dots, a_j, \dots, a_m\}$.

$$j = 1, \dots, m :$$

$$\pi_{A \to B}(a_j) \overset{\text{Def}}{=} \begin{cases} \pi(a_j) \cdot \prod\limits_{C \in s(A) \setminus \{B\}} \lambda_C(a_j) & \text{if } A \notin W \\ 1 & \text{if } A \in W \wedge a = a^* \\ 0 & \text{if } A \in W \wedge a \neq a^* \end{cases}$$

The vector

$$\vec{\pi}_{A \to B} \overset{\text{Def}}{=} \Big(\pi_{A \to B}(a_j)\Big)_{j=1}^{m}$$

is called π-*message* from A to B.

Example

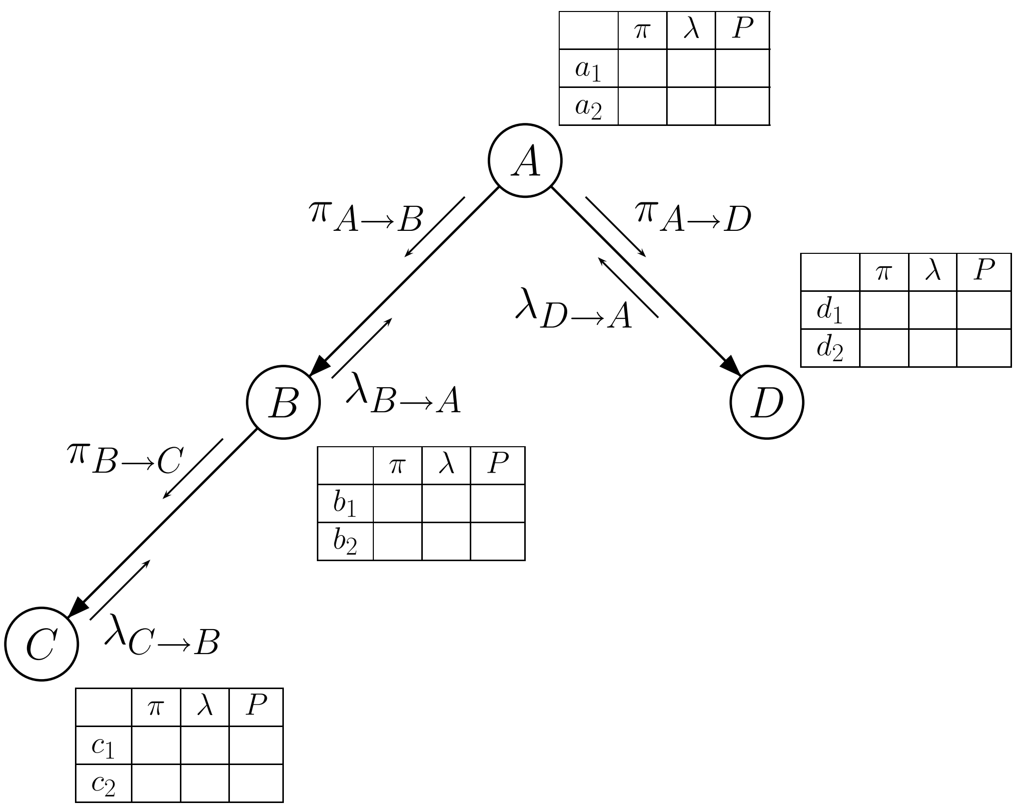
Let $B \in V$ be a non-root node in (V, E) and A the parent node of B.
Further let $b \in \mathrm{dom}(B)$ be a value of B's domain.

$$\pi(b) \;=\; \mu_{B,W} \;\cdot \sum_{a \in \mathrm{dom}(A)} P(b \mid a) \cdot \pi_{A \to B}(a)$$

Let $A \notin W$ a non-instantiated attribute and $P(V) > 0$.

$$\begin{aligned} \pi_{A \to B}(a_j) \;&=\; \pi(a_j) \;\cdot \prod_{C \in s(A) \setminus \{B\}} \lambda_{C \to A}(a_j) \\ &=\; \tau_{B,W} \cdot \frac{P(a_j \mid W)}{\lambda_{B \to A}(a_j)} \end{aligned}$$

Belief Tree:



Parameters:

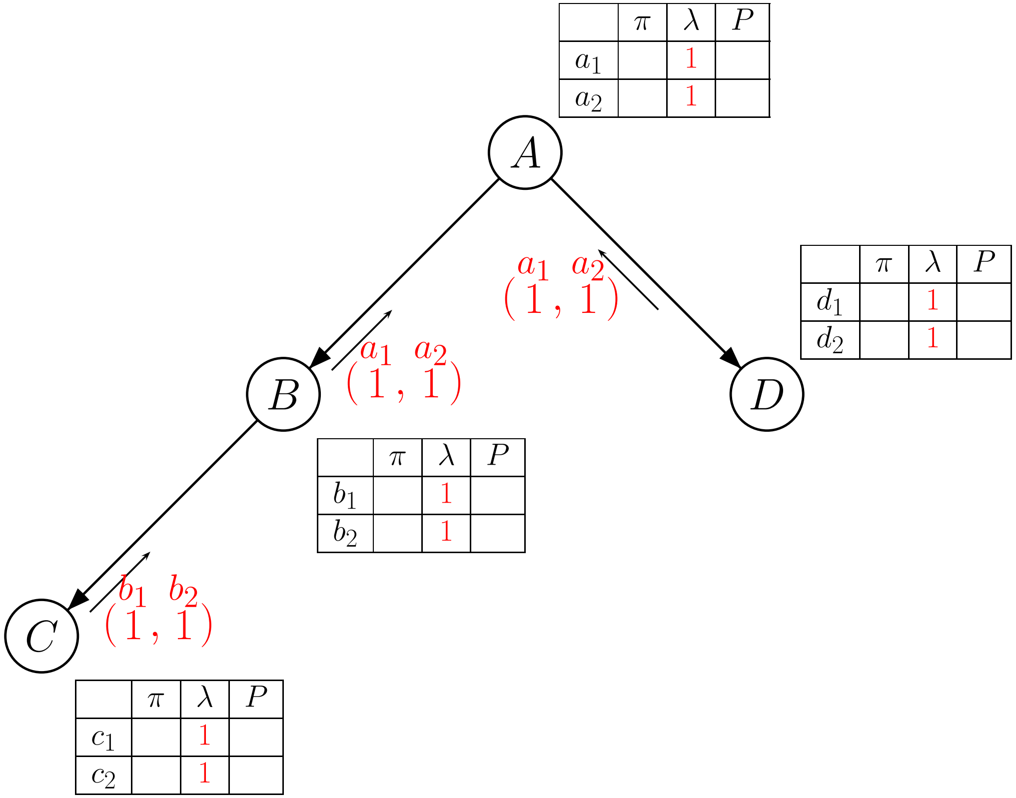
$$
\begin{array}{rr}
P(a_1) = 0.1 & P(b_1 \mid a_1) = 0.7 \\
 & P(b_1 \mid a_2) = 0.2 \\
P(d_1 \mid a_1) = 0.8 & P(c_1 \mid b_1) = 0.4 \\
P(d_1 \mid a_2) = 0.4 & P(c_1 \mid b_2) = 0.001
\end{array}
$$

Desired:

$$\forall X \in \{A, B, C, D\} : \ P(X \mid \emptyset) = ?$$

Propagation in Belief Trees (2)

Belief Tree:

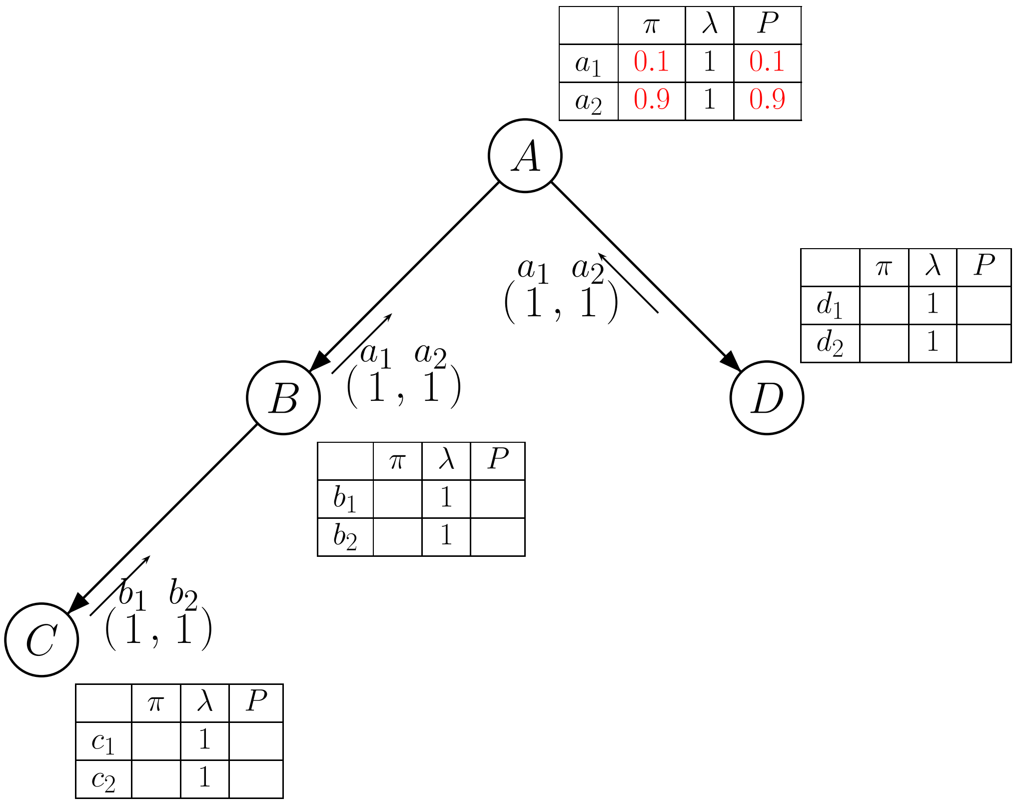


Initialization Phase:

- Set all λ-messages and λ-values to 1.

Propagation in Belief Trees (3)

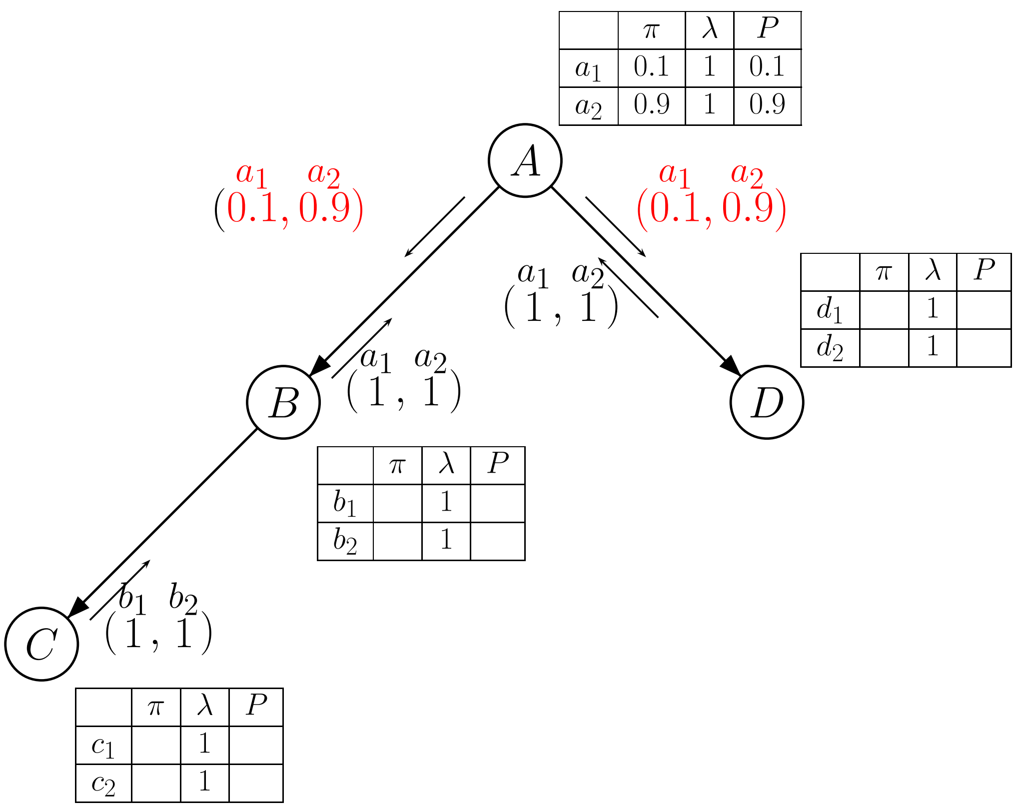
Belief Tree:



Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$

Belief Tree:



	π	λ	P
a_1	0.1	1	0.1
a_2	0.9	1	0.9

	π	λ	P
b_1		1	
b_2		1	

	π	λ	P
c_1		1	
c_2		1	

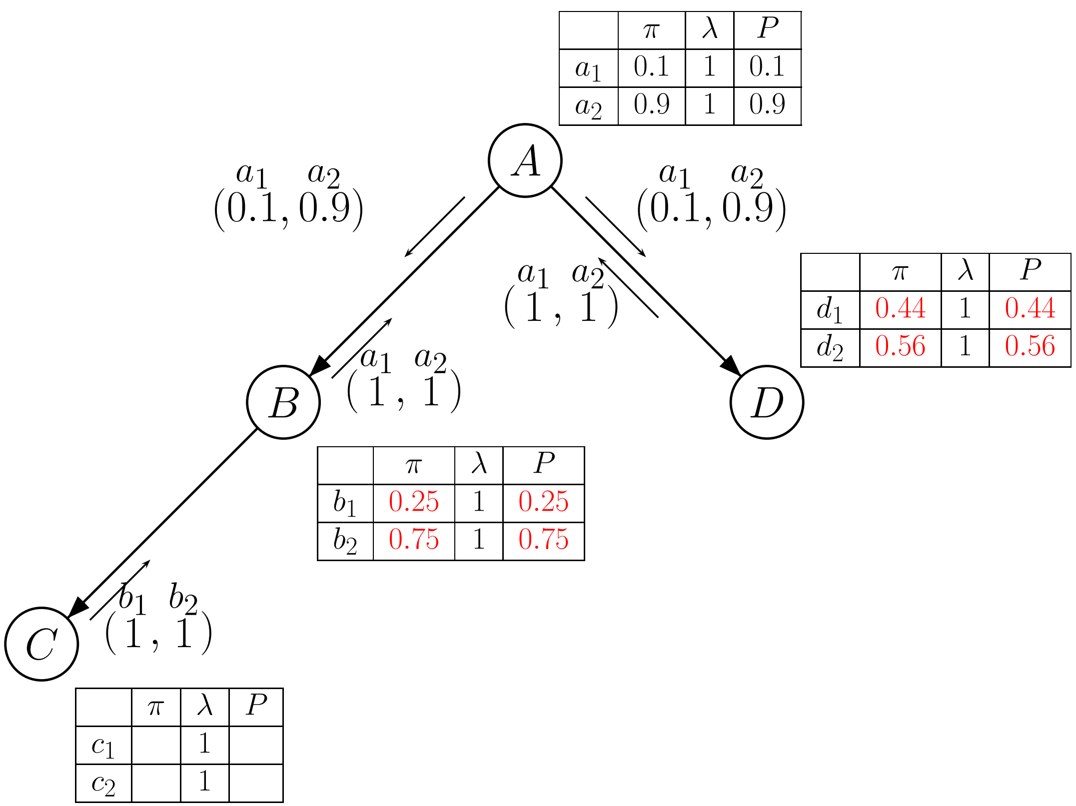
	π	λ	P
d_1		1	
d_2		1	

Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$.
- A sends π-messages to B and D.

Propagation in Belief Trees (5)

Belief Tree:



	π	λ	P
a_1	0.1	1	0.1
a_2	0.9	1	0.9

	π	λ	P
b_1	0.25	1	0.25
b_2	0.75	1	0.75

	π	λ	P
c_1		1	
c_2		1	

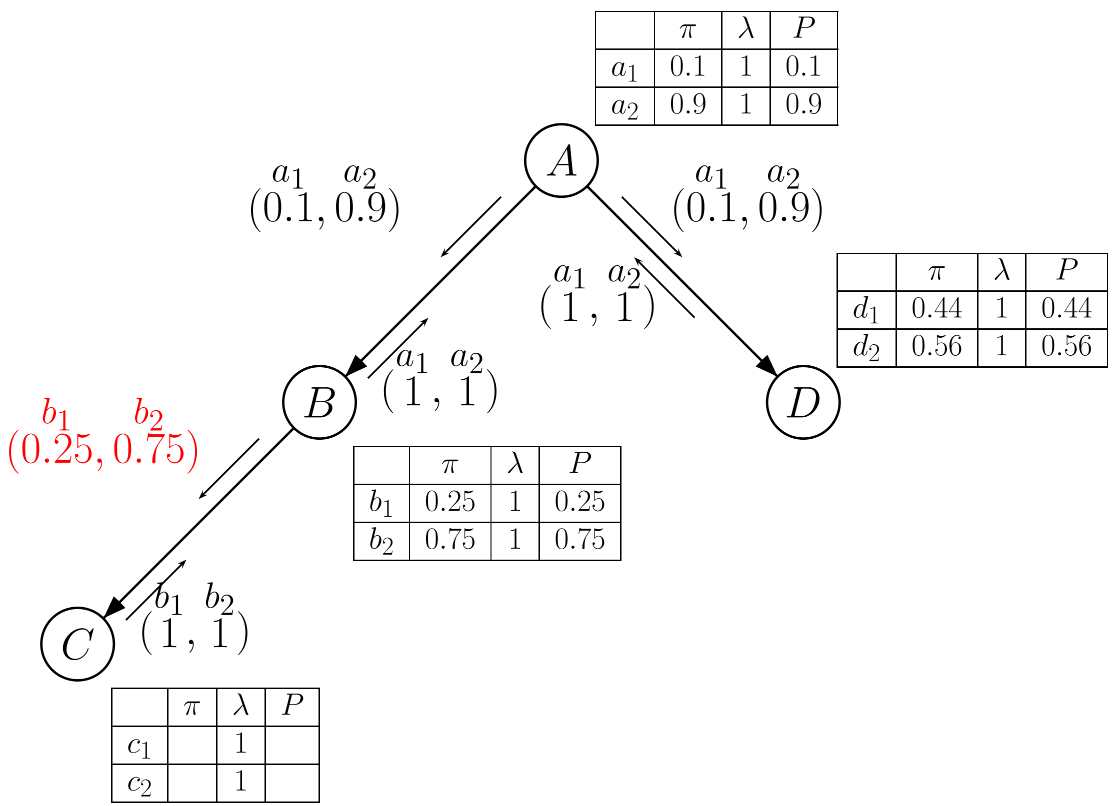
	π	λ	P
d_1	0.44	1	0.44
d_2	0.56	1	0.56

Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$.
- A sends π-messages to B and D.
- B and D update their π-values.

Propagation in Belief Trees (6)

Belief Tree:



	π	λ	P
a_1	0.1	1	0.1
a_2	0.9	1	0.9

	π	λ	P
d_1	0.44	1	0.44
d_2	0.56	1	0.56

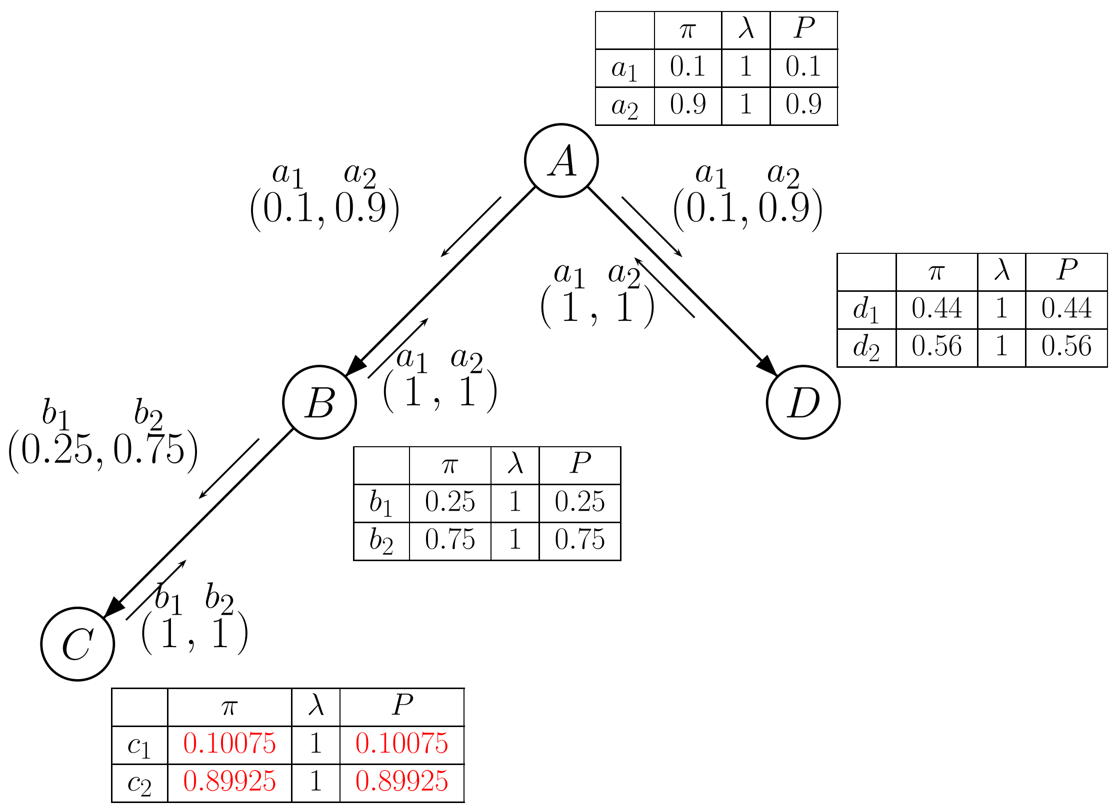
	π	λ	P
b_1	0.25	1	0.25
b_2	0.75	1	0.75

	π	λ	P
c_1		1	
c_2		1	

Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$.
- A sends π-messages to B and D.
- B and D update their π-values.
- B sends π-message to C.

Belief Tree:



	π	λ	P
a_1	0.1	1	0.1
a_2	0.9	1	0.9

	π	λ	P
d_1	0.44	1	0.44
d_2	0.56	1	0.56

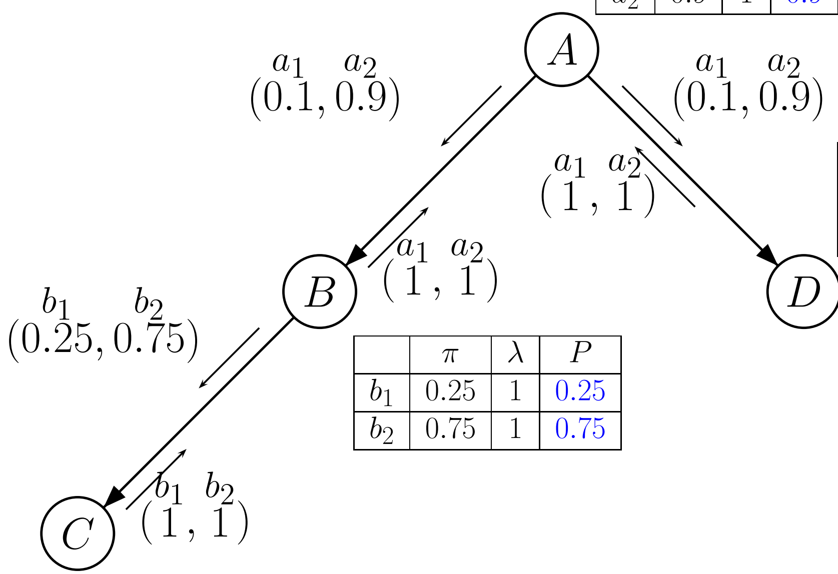
	π	λ	P
b_1	0.25	1	0.25
b_2	0.75	1	0.75

	π	λ	P
c_1	0.10075	1	0.10075
c_2	0.89925	1	0.89925

Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$.
- A sends π-messages to B and D.
- B and D update their π-values.
- B sends π-message to C.
- C updates it π-value.

Belief Tree:



	π	λ	P
a_1	0.1	1	0.1
a_2	0.9	1	0.9

	π	λ	P
d_1	0.44	1	0.44
d_2	0.56	1	0.56

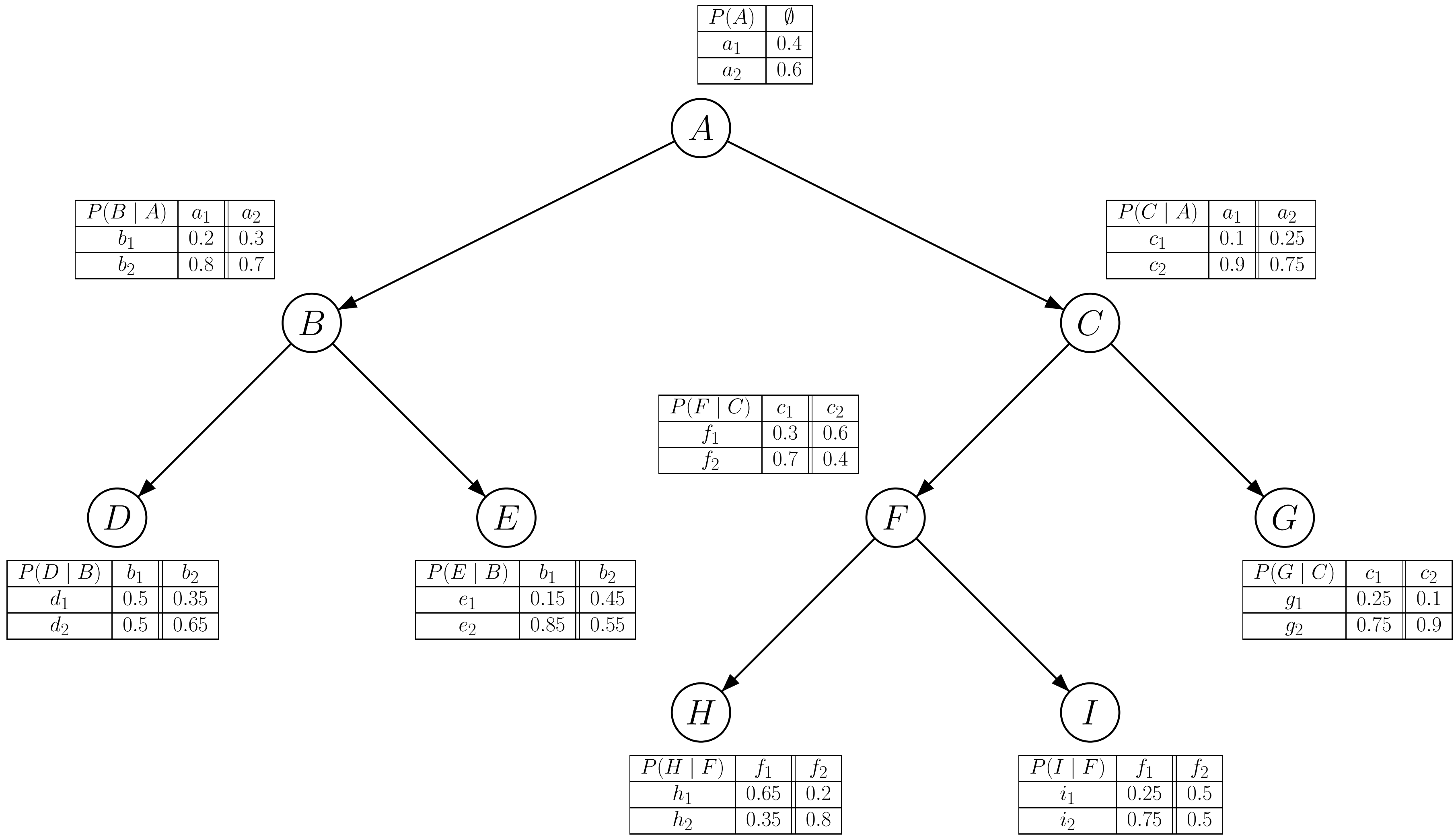
	π	λ	P
b_1	0.25	1	0.25
b_2	0.75	1	0.75

	π	λ	P
c_1	0.10075	1	0.10075
c_2	0.89925	1	0.89925

Initialization Phase:

- Set all λ-messages and λ-values to 1.
- $\pi(a_1) = P(a_1)$ and $\pi(a_2) = P(a_2)$.
- A sends π-messages to B and D.
- B and D update their π-values.
- B sends π-message to C.
- C updates it π-value.
- Initialization finished.

Larger Network (1): Parameters



$P(A)$	$\emptyset$
a_1	0.4
a_2	0.6

$P(B \mid A)$	a_1	a_2
b_1	0.2	0.3
b_2	0.8	0.7

$P(C \mid A)$	a_1	a_2
c_1	0.1	0.25
c_2	0.9	0.75

$P(F \mid C)$	c_1	c_2
f_1	0.3	0.6
f_2	0.7	0.4

$P(D \mid B)$	b_1	b_2
d_1	0.5	0.35
d_2	0.5	0.65

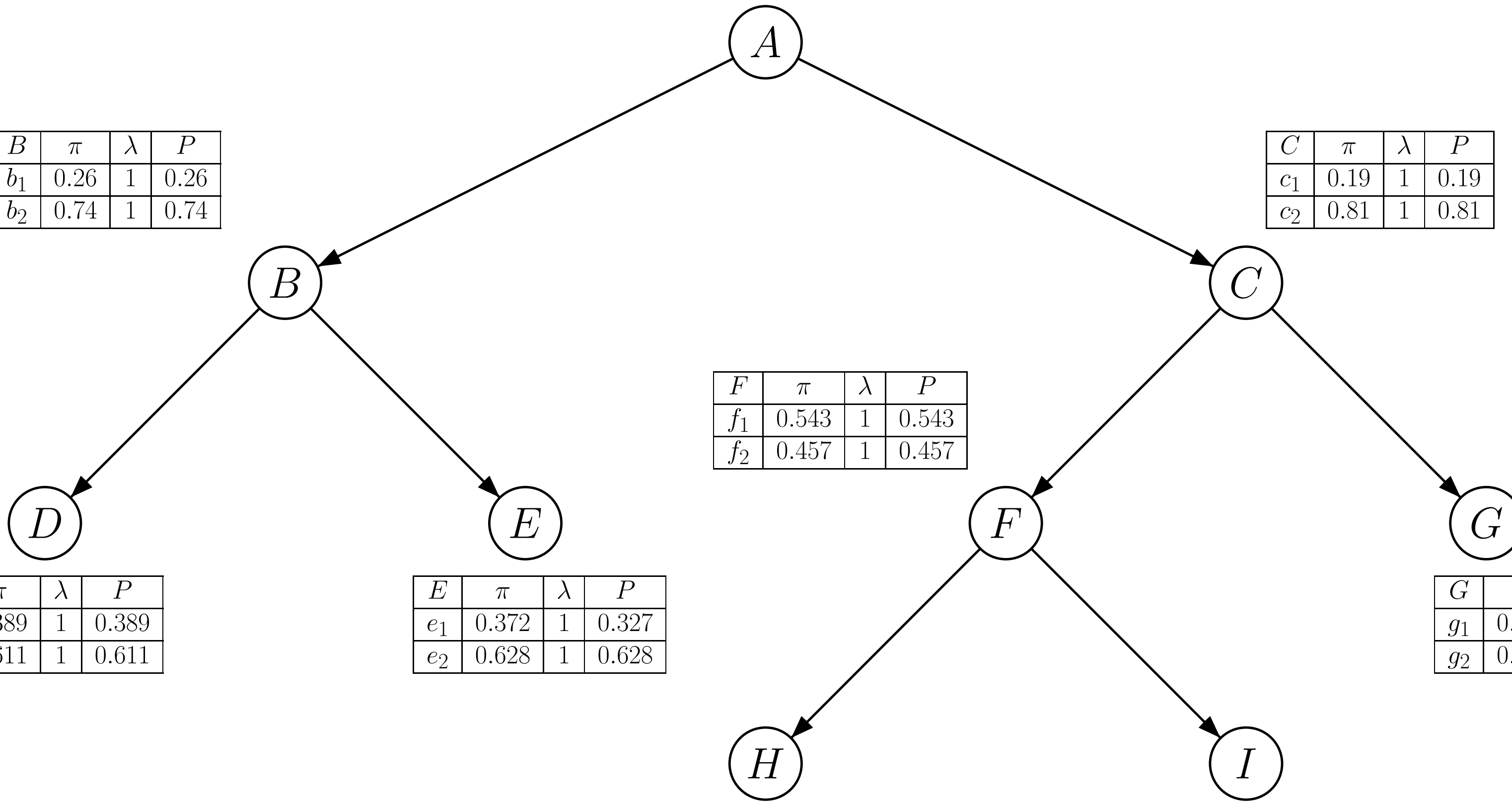
$P(E \mid B)$	b_1	b_2
e_1	0.15	0.45
e_2	0.85	0.55

$P(G \mid C)$	c_1	c_2
g_1	0.25	0.1
g_2	0.75	0.9

$P(H \mid F)$	f_1	f_2
h_1	0.65	0.2
h_2	0.35	0.8

$P(I \mid F)$	f_1	f_2
i_1	0.25	0.5
i_2	0.75	0.5

Larger Network (2): After Initialization



A	π	λ	P
a_1	0.4	1	0.4
a_2	0.6	1	0.6

B	π	λ	P
b_1	0.26	1	0.26
b_2	0.74	1	0.74

C	π	λ	P
c_1	0.19	1	0.19
c_2	0.81	1	0.81

F	π	λ	P
f_1	0.543	1	0.543
f_2	0.457	1	0.457

D	π	λ	P
d_1	0.389	1	0.389
d_2	0.611	1	0.611

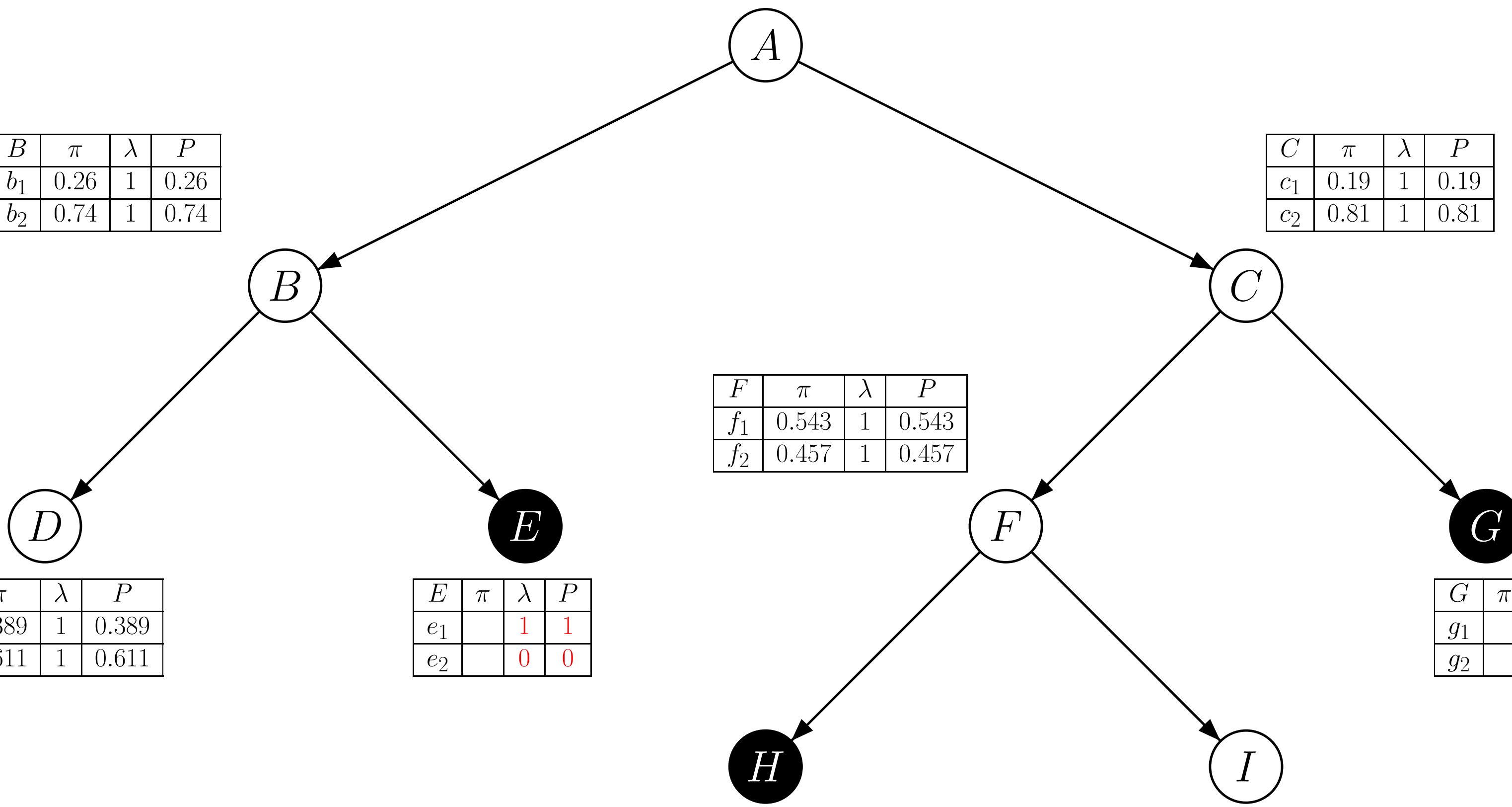
E	π	λ	P
e_1	0.372	1	0.327
e_2	0.628	1	0.628

G	π	λ	P
g_1	0.1285	1	0.1285
g_2	0.8715	1	0.8715

H	π	λ	P
h_1	0.4444	1	0.4444
h_2	0.5556	1	0.5556

I	π	λ	P
i_1	0.3643	1	0.3643
i_2	0.6357	1	0.6357

Larger Network (3): Set Evidence e_1, g_1, h_1



A	π	λ	P
a_1	0.4	1	0.4
a_2	0.6	1	0.6

B	π	λ	P
b_1	0.26	1	0.26
b_2	0.74	1	0.74

C	π	λ	P
c_1	0.19	1	0.19
c_2	0.81	1	0.81

D	π	λ	P
d_1	0.389	1	0.389
d_2	0.611	1	0.611

E	π	λ	P
e_1		1	1
e_2		0	0

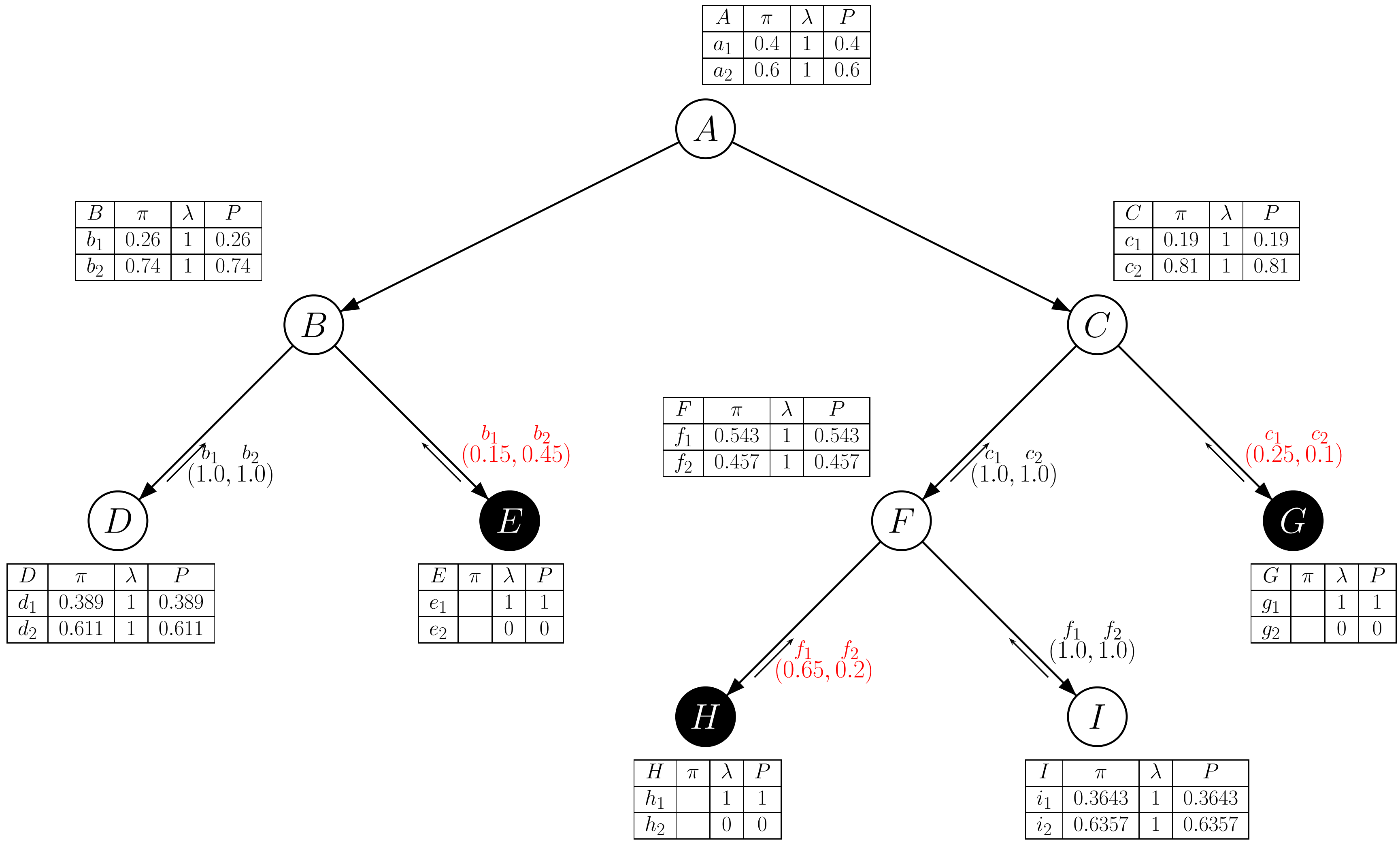
F	π	λ	P
f_1	0.543	1	0.543
f_2	0.457	1	0.457

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3643	1	0.3643
i_2	0.6357	1	0.6357

Larger Network (4): Propagate Evidence



A	π	λ	P
a_1	0.4	1	0.4
a_2	0.6	1	0.6

B	π	λ	P
b_1	0.26	1	0.26
b_2	0.74	1	0.74

C	π	λ	P
c_1	0.19	1	0.19
c_2	0.81	1	0.81

F	π	λ	P
f_1	0.543	1	0.543
f_2	0.457	1	0.457

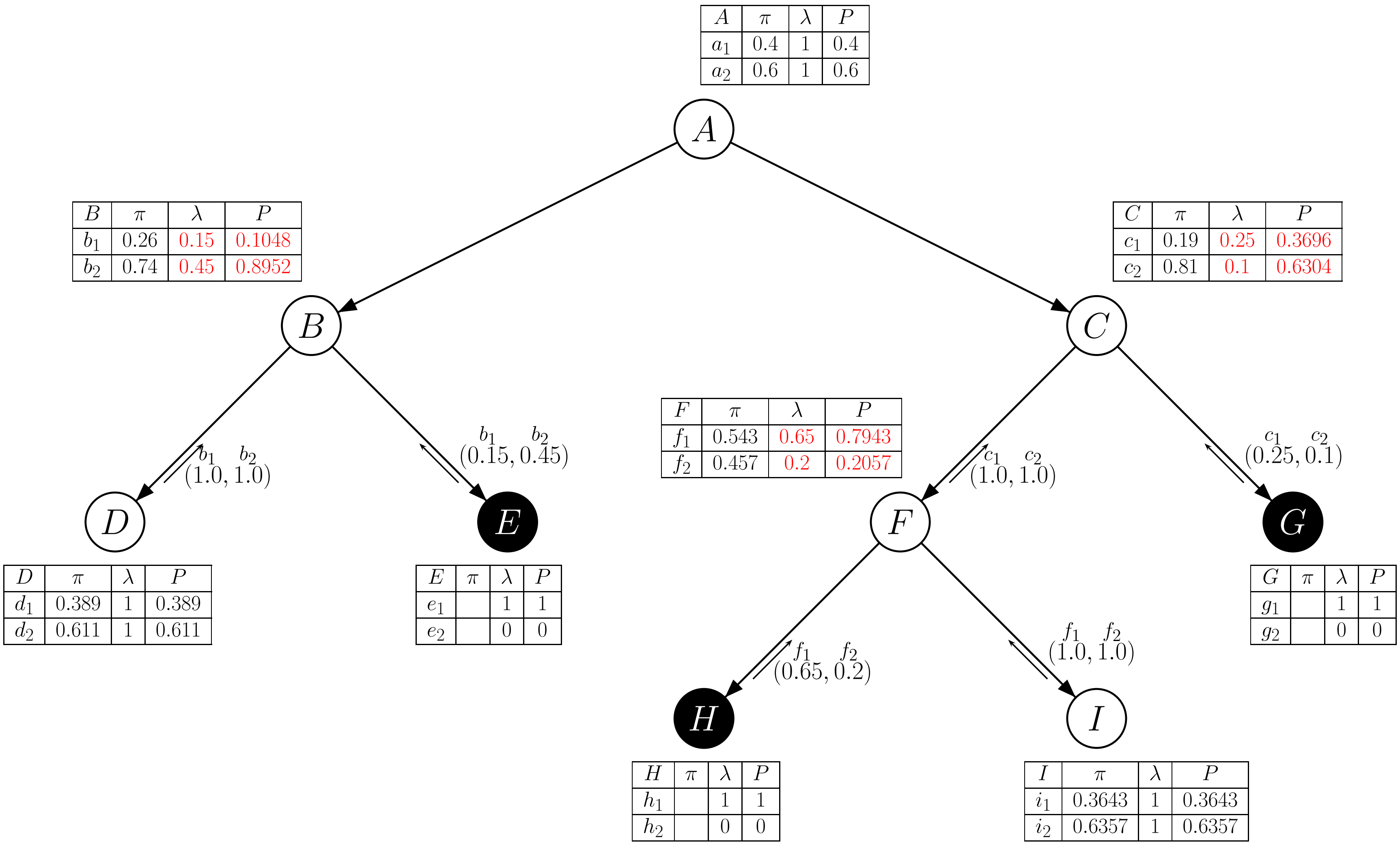
D	π	λ	P
d_1	0.389	1	0.389
d_2	0.611	1	0.611

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3643	1	0.3643
i_2	0.6357	1	0.6357



A	π	λ	P
a_1	0.4	1	0.4
a_2	0.6	1	0.6

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.25	0.3696
c_2	0.81	0.1	0.6304

F	π	λ	P
f_1	0.543	0.65	0.7943
f_2	0.457	0.2	0.2057

D	π	λ	P
d_1	0.389	1	0.389
d_2	0.611	1	0.611

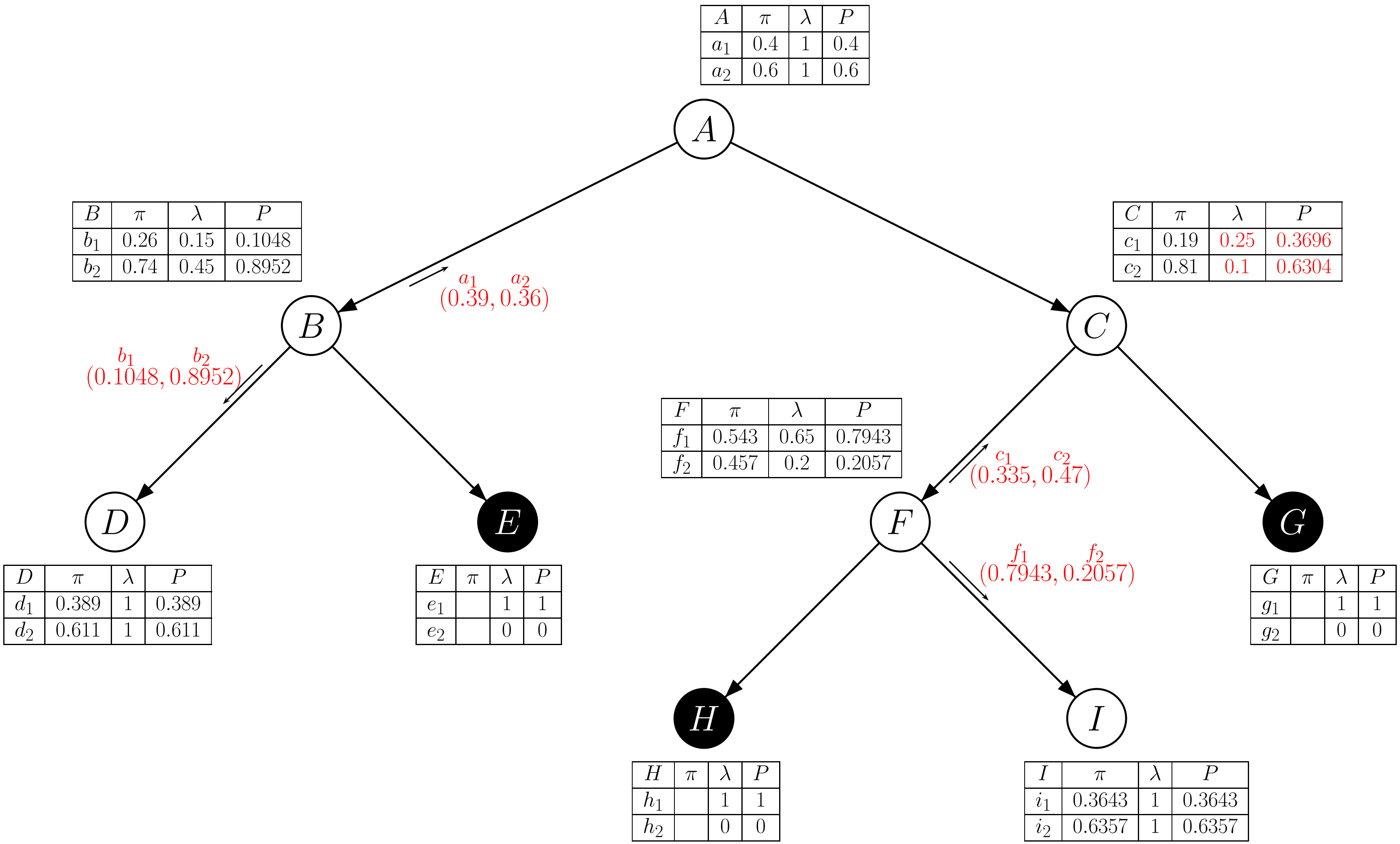
E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3643	1	0.3643
i_2	0.6357	1	0.6357

Larger Network (6): Propagate Evidence, cont.



A	π	λ	P
a_1	0.4	1	0.4
a_2	0.6	1	0.6

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.25	0.3696
c_2	0.81	0.1	0.6304

F	π	λ	P
f_1	0.543	0.65	0.7943
f_2	0.457	0.2	0.2057

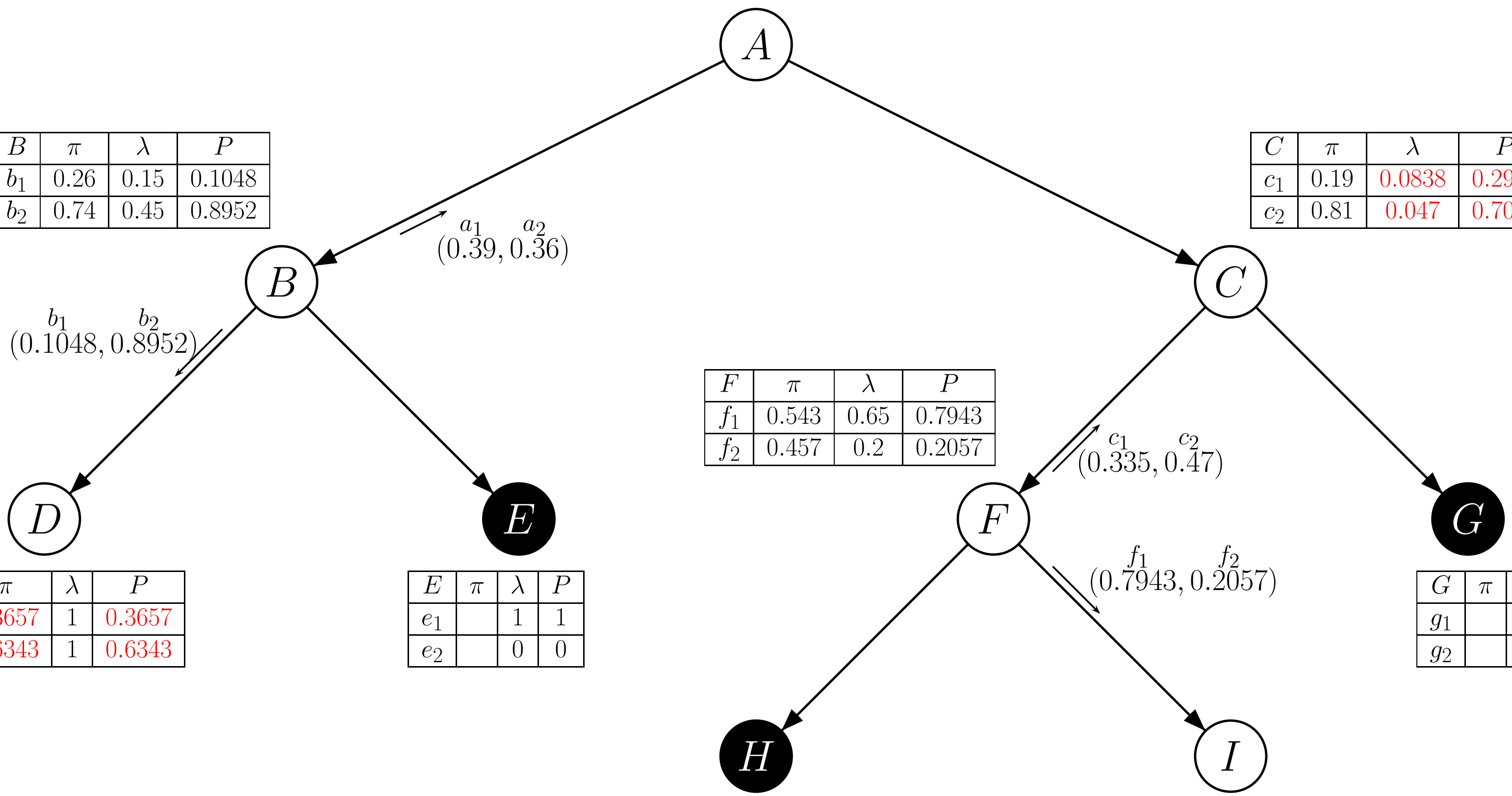
D	π	λ	P
d_1	0.389	1	0.389
d_2	0.611	1	0.611

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3643	1	0.3643
i_2	0.6357	1	0.6357



A	π	λ	P
a_1	0.4	0.39	0.4194
a_2	0.6	0.36	0.5806

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.0838	0.2948
c_2	0.81	0.047	0.7052

F	π	λ	P
f_1	0.543	0.65	0.7943
f_2	0.457	0.2	0.2057

D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

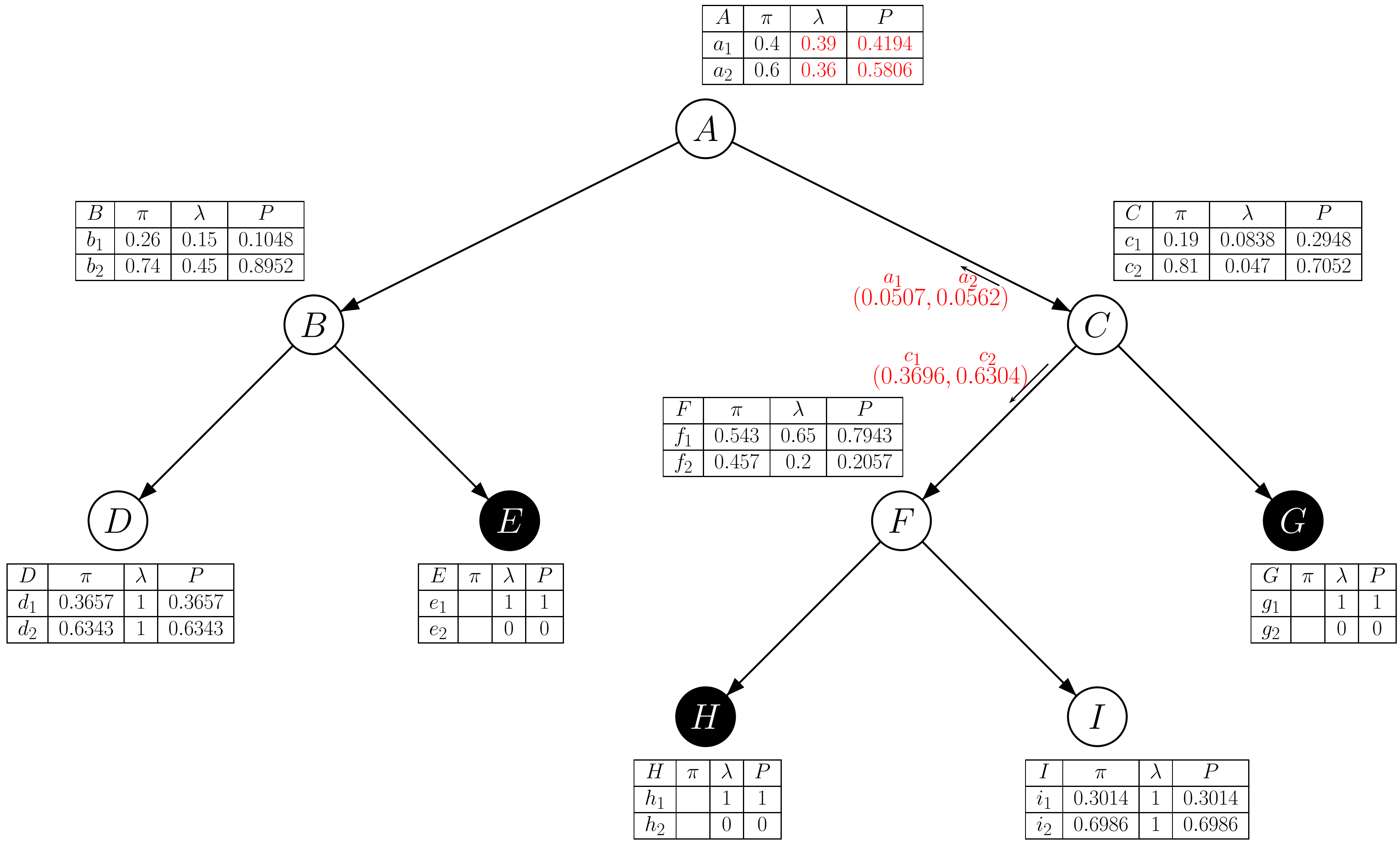
E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3014	1	0.3014
i_2	0.6986	1	0.6986

Larger Network (8): Propagate Evidence, cont.



A	π	λ	P
a_1	0.4	0.39	0.4194
a_2	0.6	0.36	0.5806

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.0838	0.2948
c_2	0.81	0.047	0.7052

F	π	λ	P
f_1	0.543	0.65	0.7943
f_2	0.457	0.2	0.2057

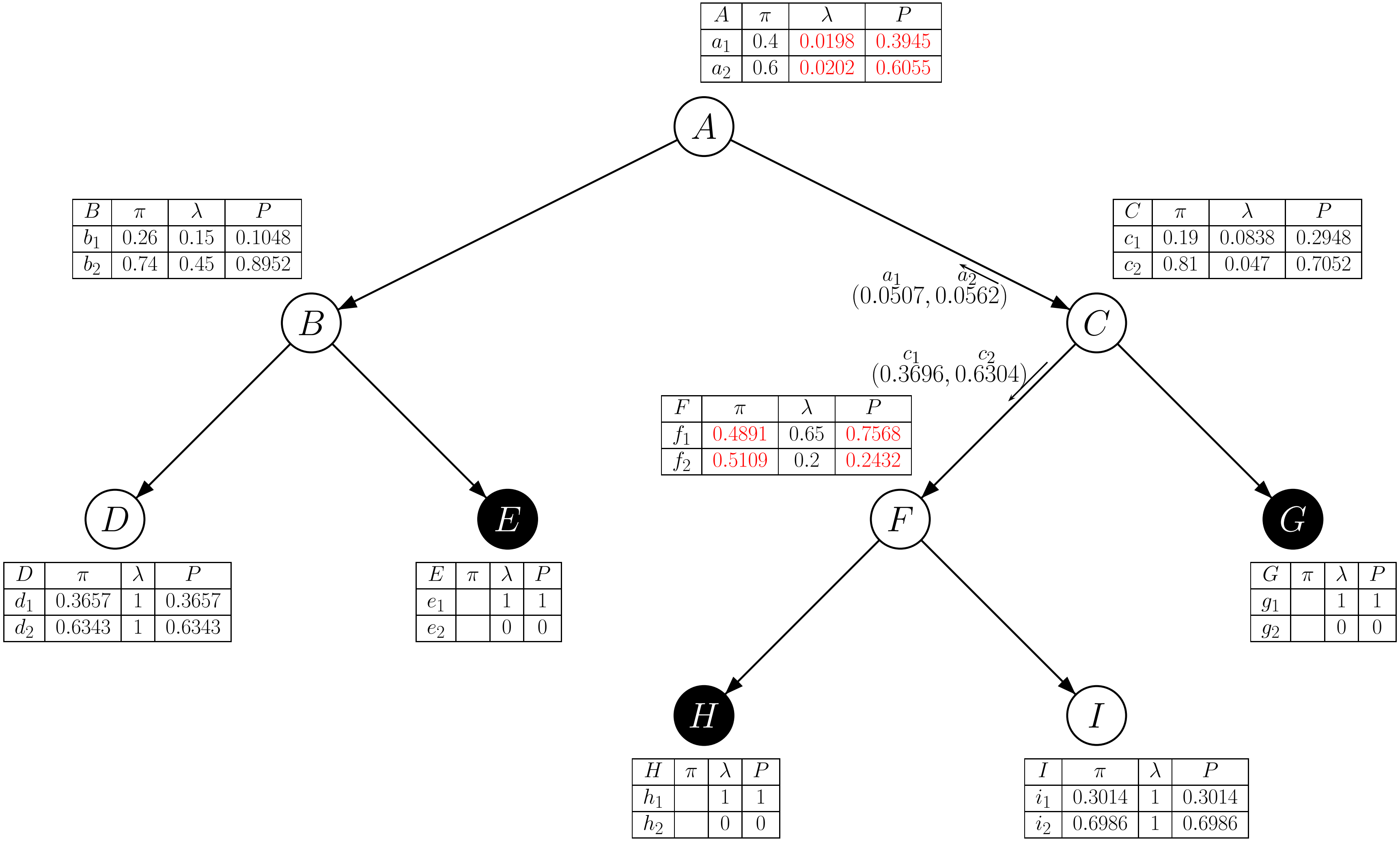
D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3014	1	0.3014
i_2	0.6986	1	0.6986



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.0838	0.2948
c_2	0.81	0.047	0.7052

F	π	λ	P
f_1	0.4891	0.65	0.7568
f_2	0.5109	0.2	0.2432

D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

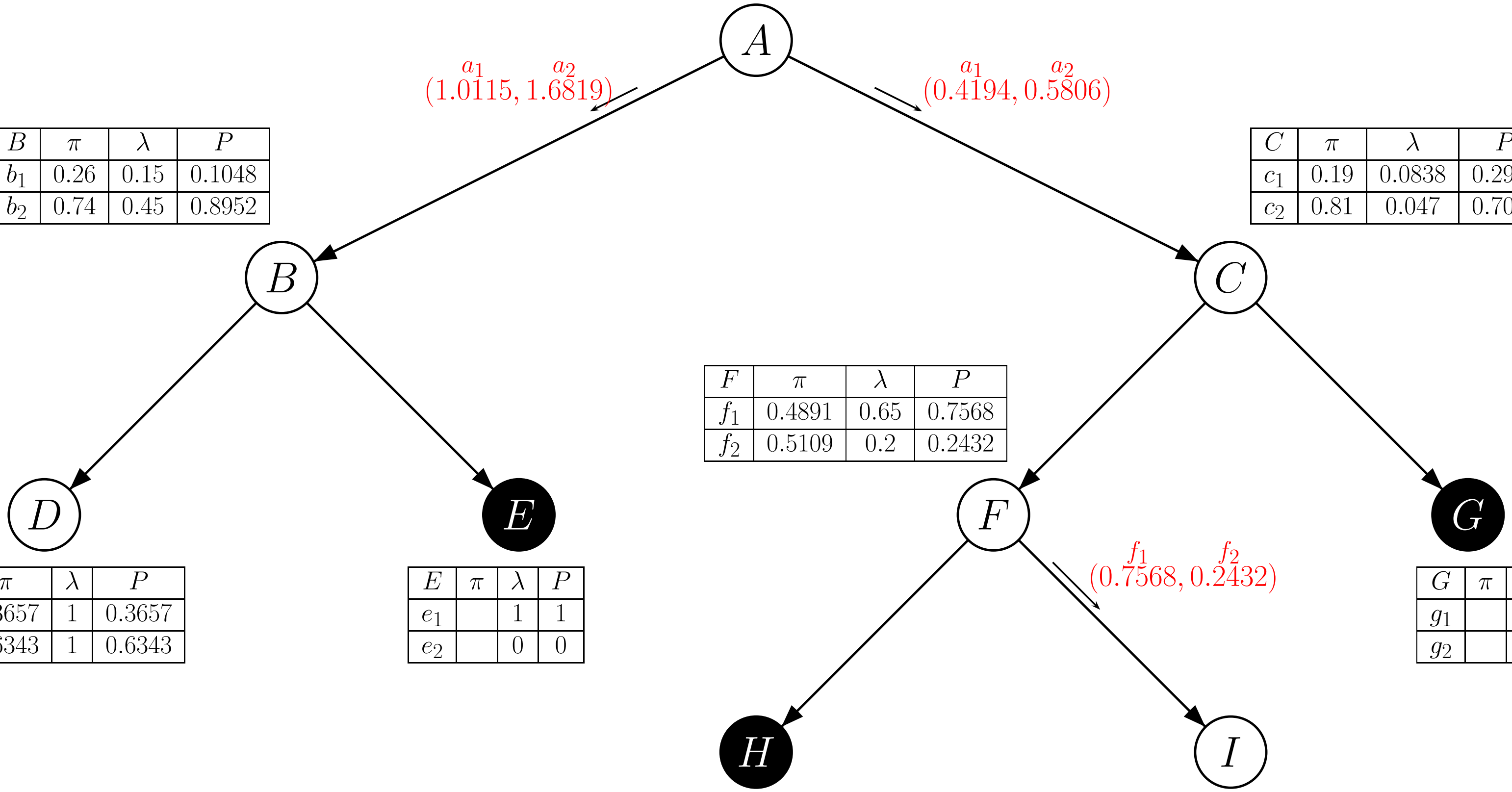
E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3014	1	0.3014
i_2	0.6986	1	0.6986

Larger Network (10): Propagate Evidence, cont.



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.26	0.15	0.1048
b_2	0.74	0.45	0.8952

C	π	λ	P
c_1	0.19	0.0838	0.2948
c_2	0.81	0.047	0.7052

F	π	λ	P
f_1	0.4891	0.65	0.7568
f_2	0.5109	0.2	0.2432

D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

E	π	λ	P
e_1		1	1
e_2		0	0

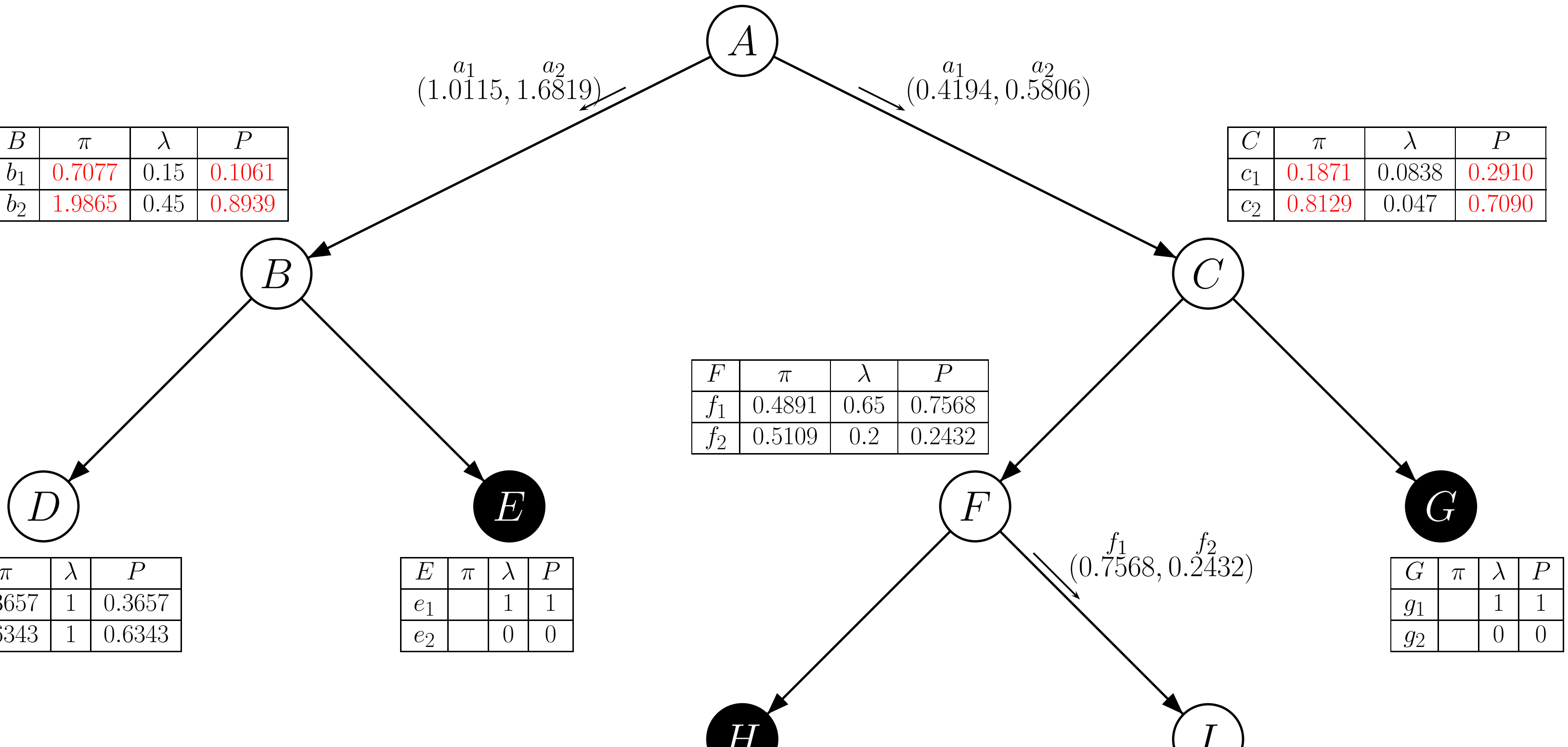
G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3014	1	0.3014
i_2	0.6986	1	0.6986

Larger Network (11): Propagate Evidence, cont.

A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055



B	π	λ	P
b_1	0.7077	0.15	0.1061
b_2	1.9865	0.45	0.8939

C	π	λ	P
c_1	0.1871	0.0838	0.2910
c_2	0.8129	0.047	0.7090

F	π	λ	P
f_1	0.4891	0.65	0.7568
f_2	0.5109	0.2	0.2432

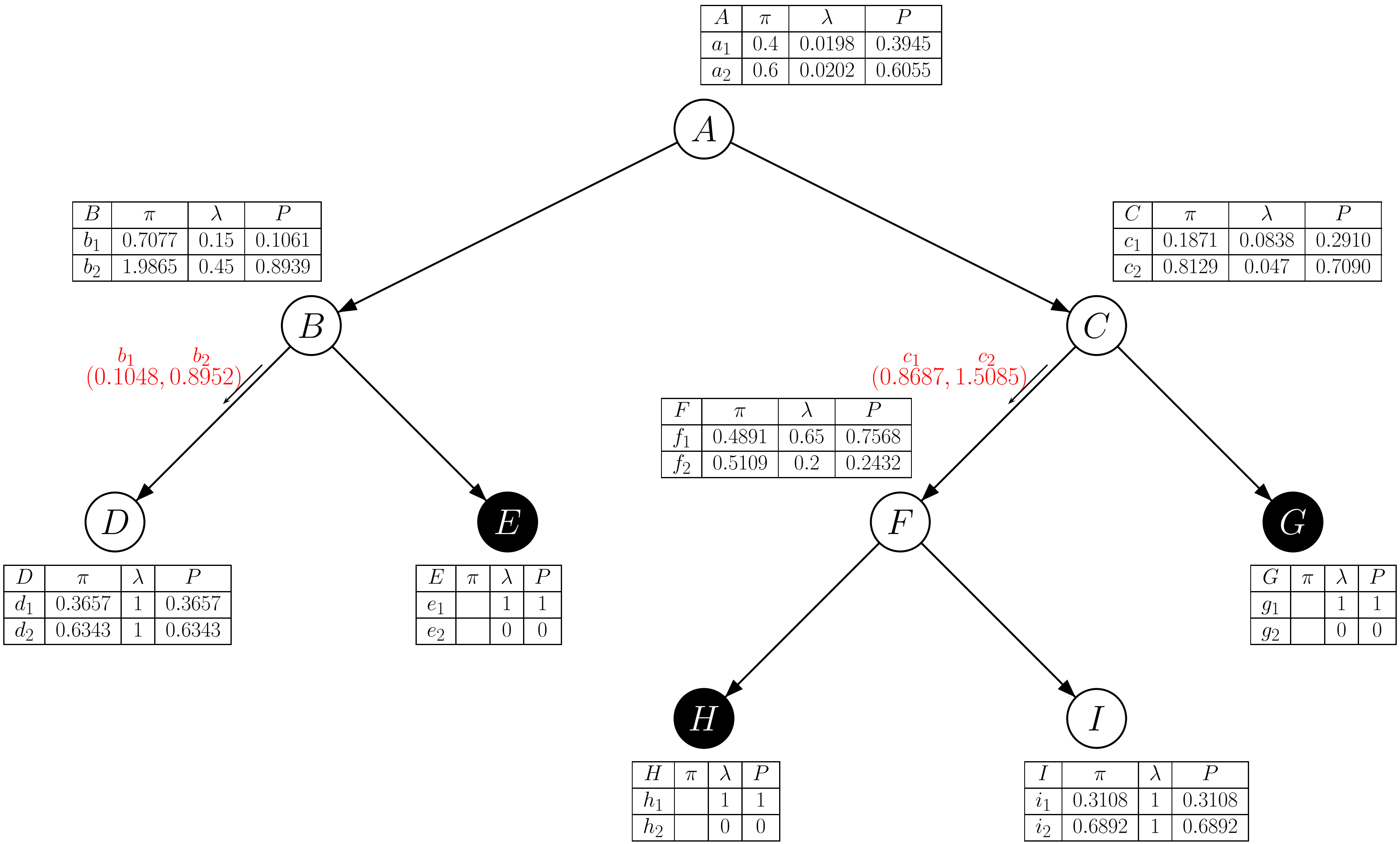
D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3108	1	0.3108
i_2	0.6892	1	0.6892



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.7077	0.15	0.1061
b_2	1.9865	0.45	0.8939

C	π	λ	P
c_1	0.1871	0.0838	0.2910
c_2	0.8129	0.047	0.7090

F	π	λ	P
f_1	0.4891	0.65	0.7568
f_2	0.5109	0.2	0.2432

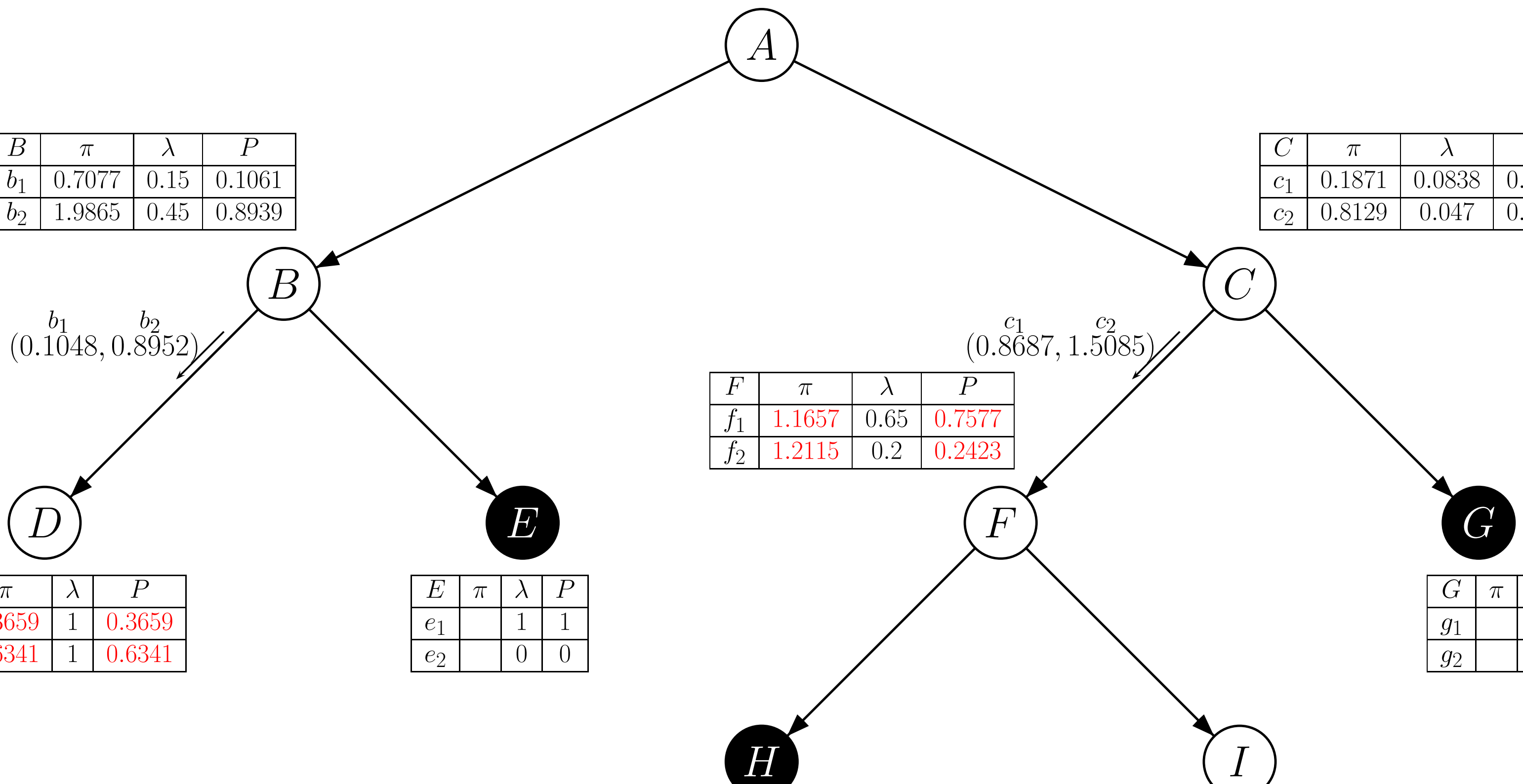
D	π	λ	P
d_1	0.3657	1	0.3657
d_2	0.6343	1	0.6343

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3108	1	0.3108
i_2	0.6892	1	0.6892



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.7077	0.15	0.1061
b_2	1.9865	0.45	0.8939

C	π	λ	P
c_1	0.1871	0.0838	0.2910
c_2	0.8129	0.047	0.7090

F	π	λ	P
f_1	1.1657	0.65	0.7577
f_2	1.2115	0.2	0.2423

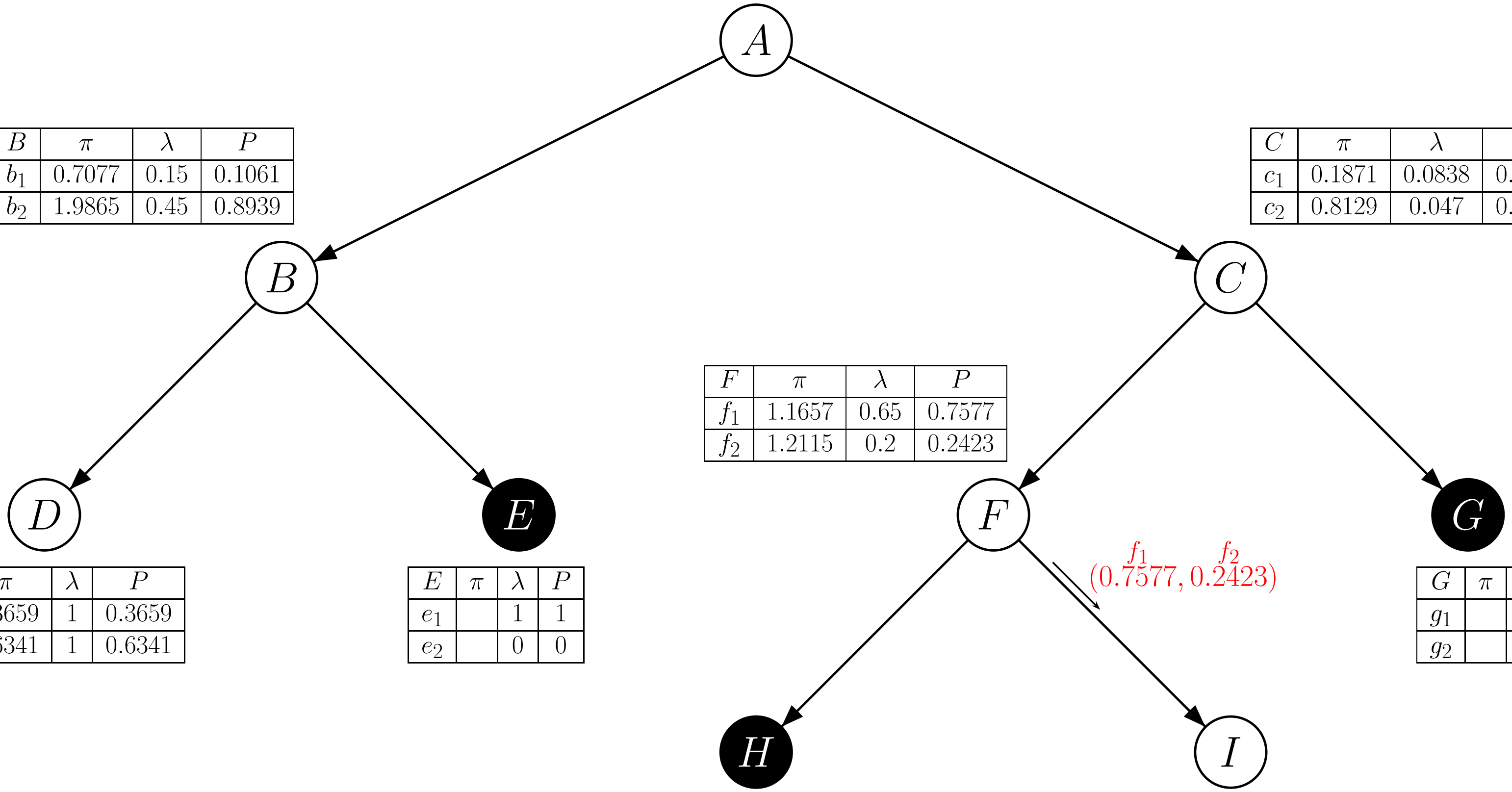
D	π	λ	P
d_1	0.3659	1	0.3659
d_2	0.6341	1	0.6341

E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3108	1	0.3108
i_2	0.6892	1	0.6892



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.7077	0.15	0.1061
b_2	1.9865	0.45	0.8939

C	π	λ	P
c_1	0.1871	0.0838	0.2910
c_2	0.8129	0.047	0.7090

F	π	λ	P
f_1	1.1657	0.65	0.7577
f_2	1.2115	0.2	0.2423

D	π	λ	P
d_1	0.3659	1	0.3659
d_2	0.6341	1	0.6341

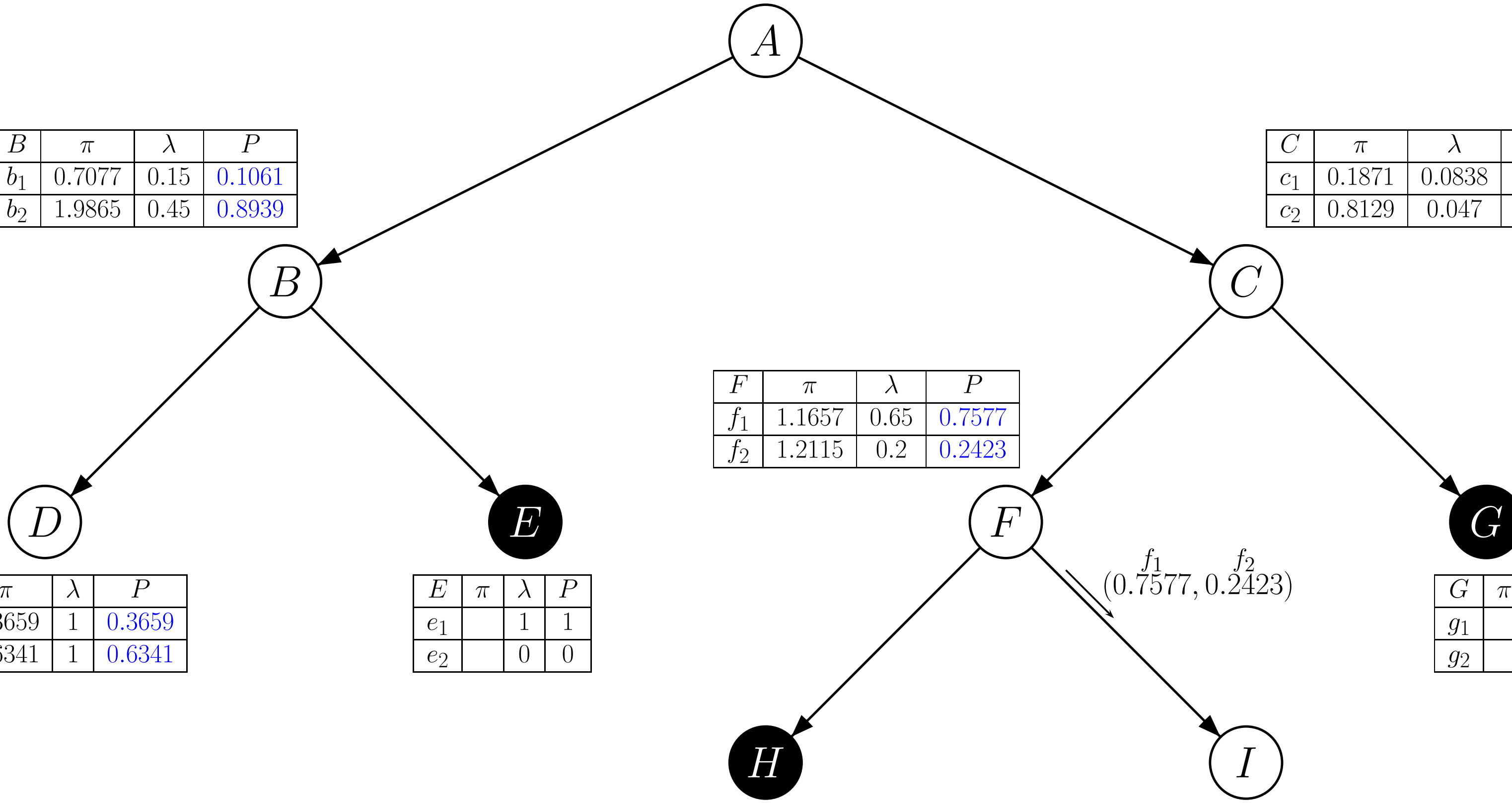
E	π	λ	P
e_1		1	1
e_2		0	0

G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3108	1	0.3108
i_2	0.6892	1	0.6892

Larger Network (15): Finished



A	π	λ	P
a_1	0.4	0.0198	0.3945
a_2	0.6	0.0202	0.6055

B	π	λ	P
b_1	0.7077	0.15	0.1061
b_2	1.9865	0.45	0.8939

C	π	λ	P
c_1	0.1871	0.0838	0.2910
c_2	0.8129	0.047	0.7090

F	π	λ	P
f_1	1.1657	0.65	0.7577
f_2	1.2115	0.2	0.2423

D	π	λ	P
d_1	0.3659	1	0.3659
d_2	0.6341	1	0.6341

E	π	λ	P
e_1		1	1
e_2		0	0

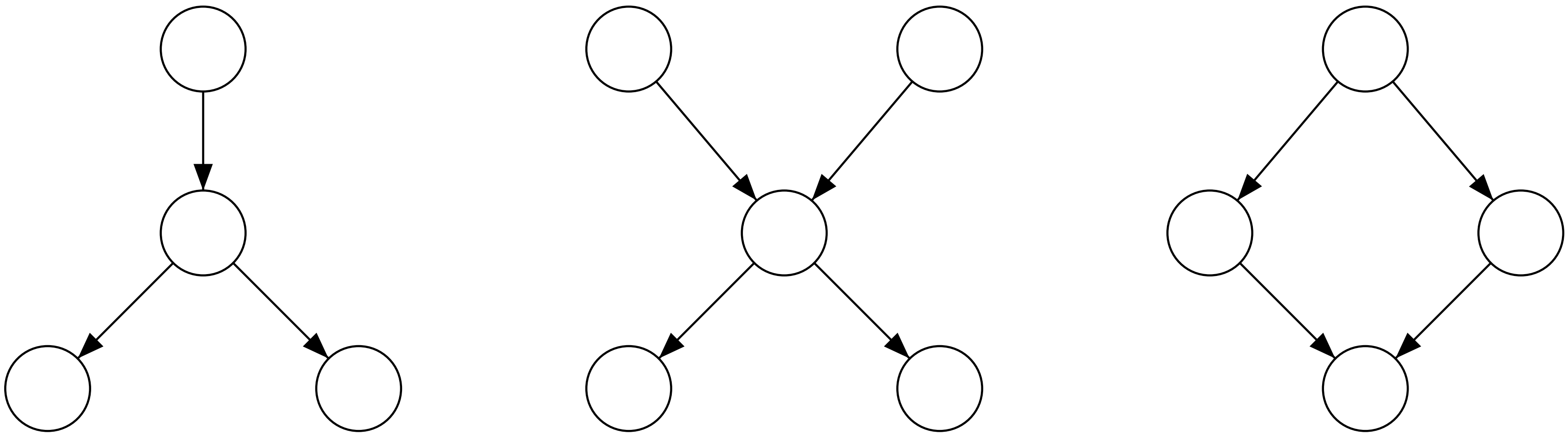
G	π	λ	P
g_1		1	1
g_2		0	0

H	π	λ	P
h_1		1	1
h_2		0	0

I	π	λ	P
i_1	0.3106	1	0.3106
i_2	0.6894	1	0.6894

Propagation in Clique Trees

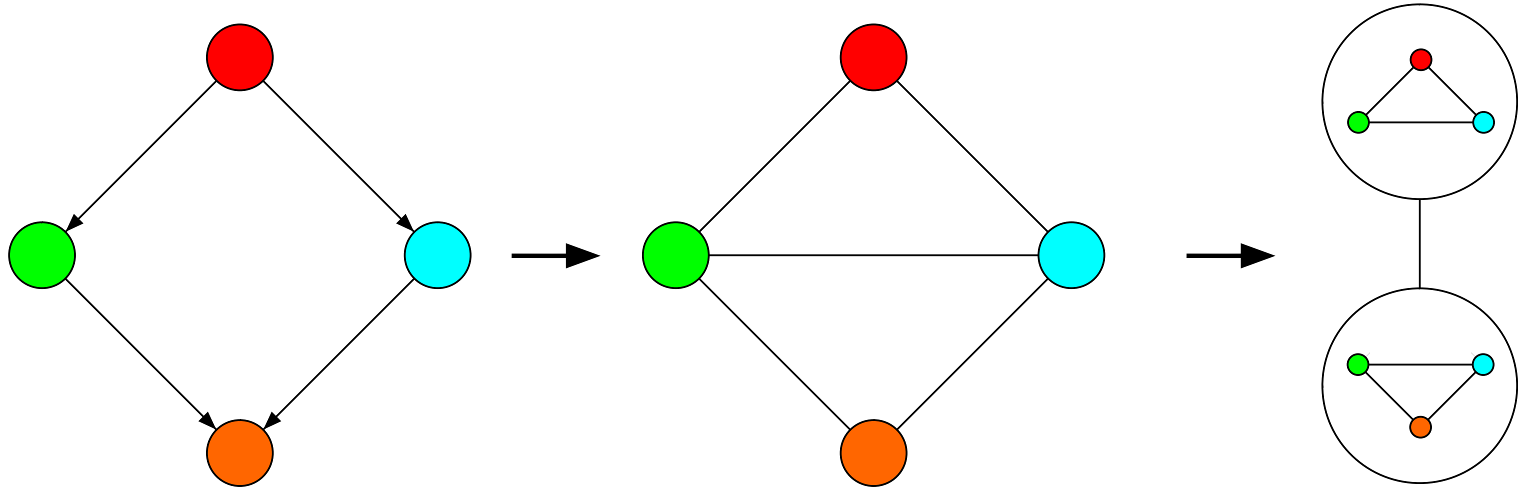
Problems



- The propagation algorithm as presented can only deal with *trees*.
- Can be extended to *polytrees* (i. e. singly connected graphs with multiple parents per node).
- However, it cannot handle networks that contains loops.

Idea

- Combine nodes of the original (primary) graph structure
- These groups form the nodes of a secondary structure
- Find a transformation that yields tree structure



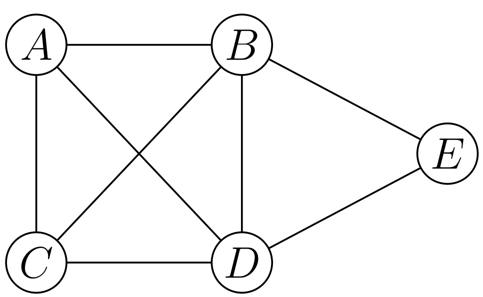
Complete Graph

An undirected Graph $G = (V, E)$ is called *complete*, if every pair of (distinct) nodes is connected by an edge.

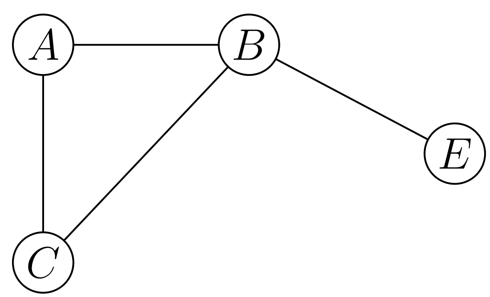
Induced Subgraph

Let $G = (V, E)$ be an undirected graph and $W \subseteq V$ a selection of nodes. Then, $G_W = (W, E_W)$ is called the *subgraph of G induced by W* with E_W being

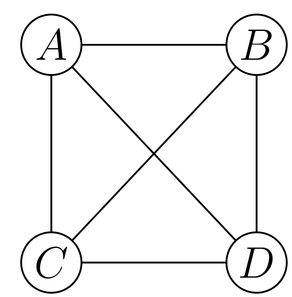
$$E_W = \{(u, v) \in E \mid u, v \in W\}.$$



Incomplete graph



Subgraph (W, E_W)
with $W = \{A, B, C, E\}$



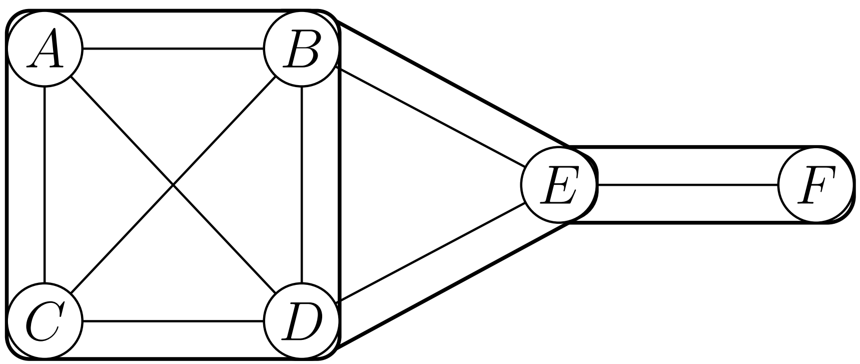
Complete (sub)graph

Complete Set, Clique

Let $G = (V, E)$ be an undirected graph. A set $W \subseteq V$ is called *complete* iff it induces a complete subgraph. It is further called a *clique*, iff W is maximal, i. e. it is not possible to add a node to W without violating the completeness condition.

a) W is complete $\Leftrightarrow$ W induces a complete subgraph

b) W is a clique $\Leftrightarrow$ W is complete and maximal



3 cliques

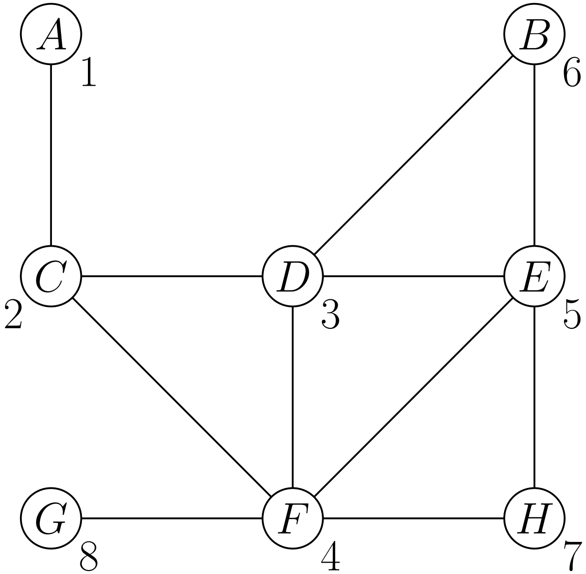
$$C_1 = \{A, B, C, D\}$$
$$C_2 = \{B, D, E\}$$
$$C_3 = \{E, F\}$$

Perfect Ordering

Let $G = (V, E)$ be an undirected graph with n nodes and $\alpha = \langle v_1, \ldots, v_n \rangle$ a total ordering on V. Then, α is called *perfect*, if the following sets

$$\operatorname{adj}(v_i) \cap \{v_1, \ldots, v_{i-1}\} \qquad i = 1, \ldots, n$$

are complete, where $\operatorname{adj}(v_i) = \{w \mid (v_i, w) \in E\}$ returns the adjacent nodes of v_i.



$\alpha = \langle A,\ C,\ D,\ F,\ E,\ B,\ H,\ G \rangle$

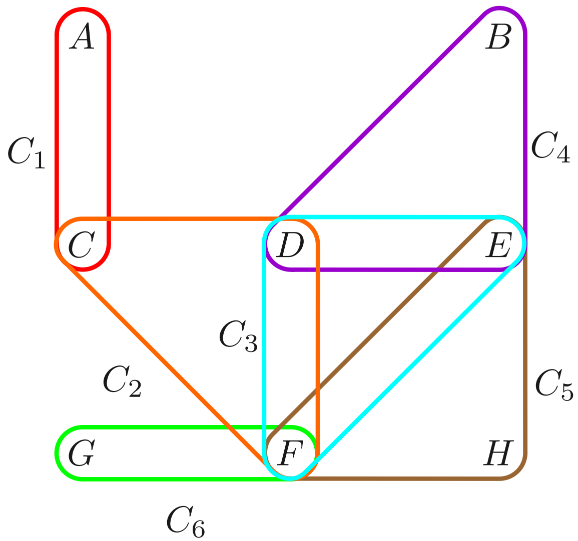
i	$\operatorname{adj}(v_i)$	$\operatorname{adj}(v_i) \cap \{v_1, \ldots, v_{i-1}\}$		
1	$\{C\}$	$\{C\} \cap \emptyset$	$= \emptyset$	complete
2	$\{A, D, F\}$	$\{A\} \cap \{A, D, F\}$	$= \{A\}$	complete
3	$\{C, B, E, F\}$	$\{A, C\} \cap \{C, B, E, F\}$	$= \{C\}$	complete
4	$\{G, C, D, E, H\}$	$\{A, C, D\} \cap \{G, C, D, E, H\}$	$= \{C, D\}$	complete
5	$\{B, D, F, H\}$	$\{A, C, D, F\} \cap \{B, D, F, H\}$	$= \{D, F\}$	complete
6	$\{D, E\}$	$\{A, C, D, F, E\} \cap \{D, E\}$	$= \{D, E\}$	complete
7	$\{F, E\}$	$\{A, C, D, F, E, B\} \cap \{F, E\}$	$= \{F, E\}$	complete
8	$\{F\}$	$\{A, C, D, F, E, B, H\} \cap \{F\}$	$= \{F\}$	complete

α is a perfect ordering

Running Intersection Property

Let $G = (V, E)$ be an undirected graph with p cliques. An ordering of these cliques has the *running intersection property (RIP)*, if for every $j > 1$ there exists an $i < j$ such that:

$$C_j \ \cap \ \left(C_1 \cup \cdots \cup C_{j-1}\right) \ \subseteq \ C_i$$

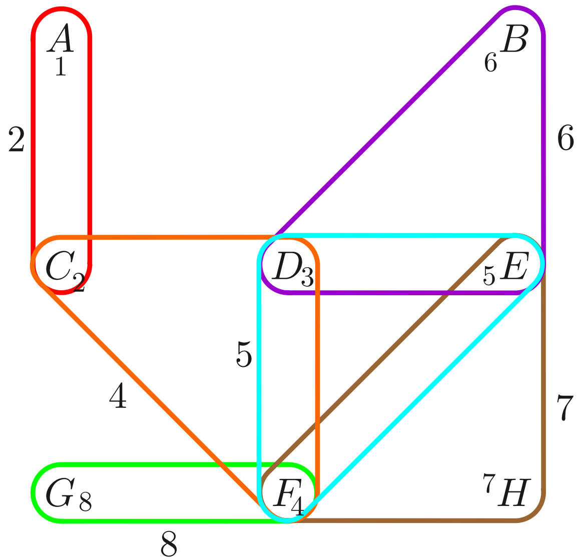


$\xi = \langle C_1,\ C_2,\ C_3,\ C_4,\ C_5,\ C_6 \rangle$

j				i
2	$C_2 \cap C_1$	$= \{C\}$	$\subseteq C_1$	1
3	$C_3 \cap (C_1 \cup C_2)$	$= \{D, F\}$	$\subseteq C_2$	2
4	$C_4 \cap (C_1 \cup C_2 \cup C_3)$	$= \{D, E\}$	$\subseteq C_3$	3
5	$C_5 \cap (C_1 \cup C_2 \cup C_3 \cup C_4)$	$= \{E, F\}$	$\subseteq C_3$	3
6	$C_6 \cap (C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5)$	$= \{F\}$	$\subseteq C_5$	5

ξ has running intersection property

If a node ordering α of an undirected graph $G = (V, E)$ is perfect and the cliques of G are ordered according to the highest rank (w. r. t. α) of the containing nodes, then this clique ordering has RIP.

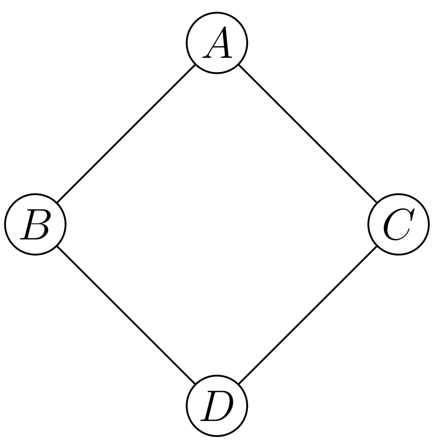


Clique	Rank		
$\{A, C\}$	$\max\{\alpha(A), \alpha(C)\}$	$= 2$	$\rightarrow C_1$
$\{C, D, F\}$	$\max\{\alpha(C), \alpha(D), \alpha(F)\}$	$= 4$	$\rightarrow C_2$
$\{D, E, F\}$	$\max\{\alpha(D), \alpha(E), \alpha(F)\}$	$= 5$	$\rightarrow C_3$
$\{B, D, E\}$	$\max\{\alpha(B), \alpha(D), \alpha(E)\}$	$= 6$	$\rightarrow C_4$
$\{F, E, H\}$	$\max\{\alpha(F), \alpha(E), \alpha(H)\}$	$= 7$	$\rightarrow C_5$
$\{F, G\}$	$\max\{\alpha(F), \alpha(G)\}$	$= 8$	$\rightarrow C_6$

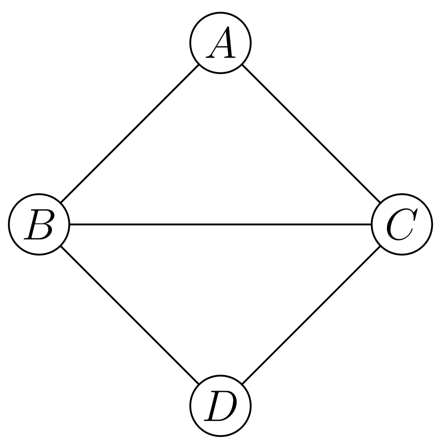
How to get a perfect ordering?

Triangulated Graph

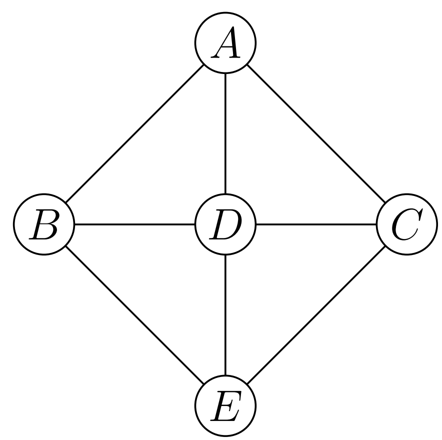
An undirected graph is called *triangulated* if every simple loop (i. e. path with identical start and end node but with any other node occurring at most once) of length greater 3 has a chord.



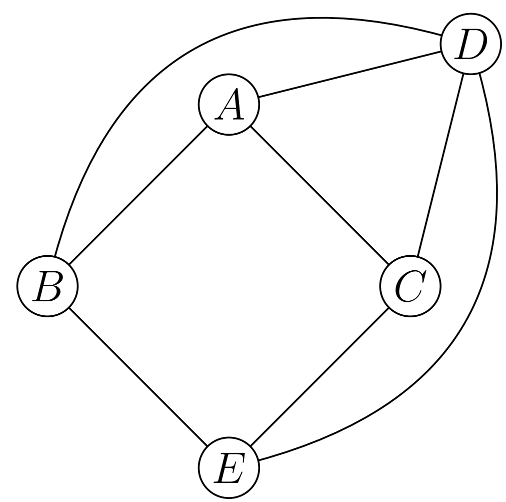
not triangulated



triangulated



not triangulated

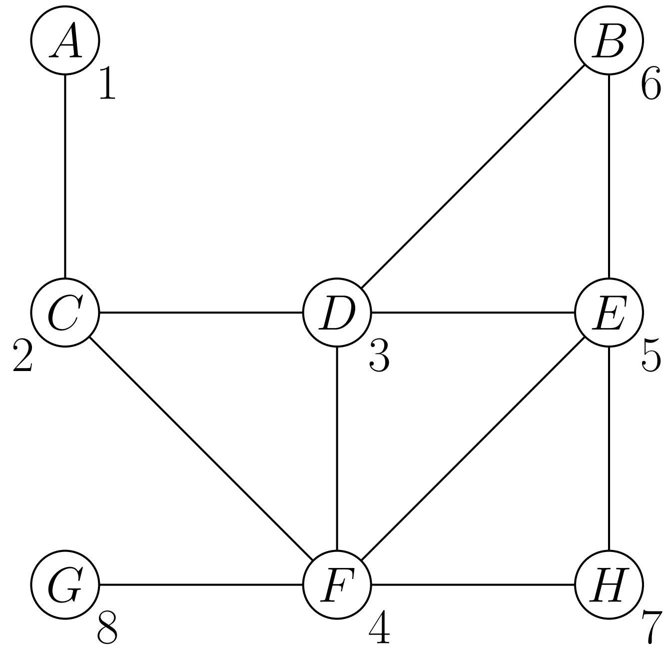


no chord for $\langle A, B, E, C \rangle$

Triangulated Graphs (2)

Maximum Cardinality Search

Let $G = (V, E)$ be an undirected graph. An ordering according *maximum cardinality search (MCS)* is obtained by first assigning 1 to an arbitray node. If n numbers are assigned the node that is connected to most of the nodes already numbered gets assigned number $n + 1$.



3 can be assigned to D or F

6 can be assigned to H or B

An undirected graph is triangulated iff the ordering obtained by MCS is perfect.

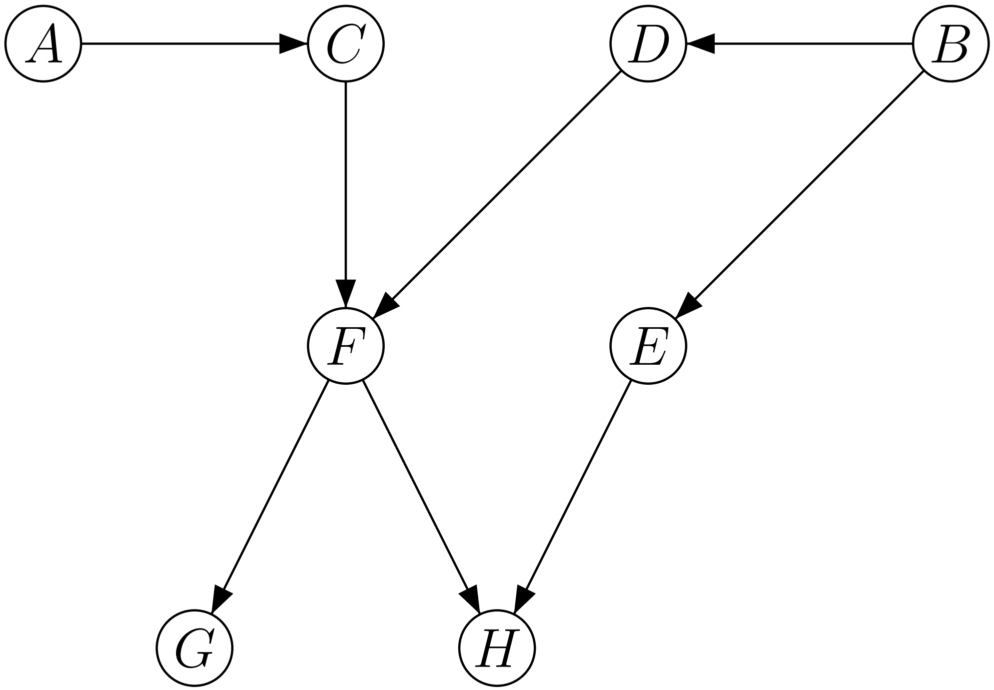
To check whether a graph is triangulated is efficient to implement. The optimization problem that is related to the triangulation task is NP-hard. However, there are good heuristics.

Moral Graph (Repetition)

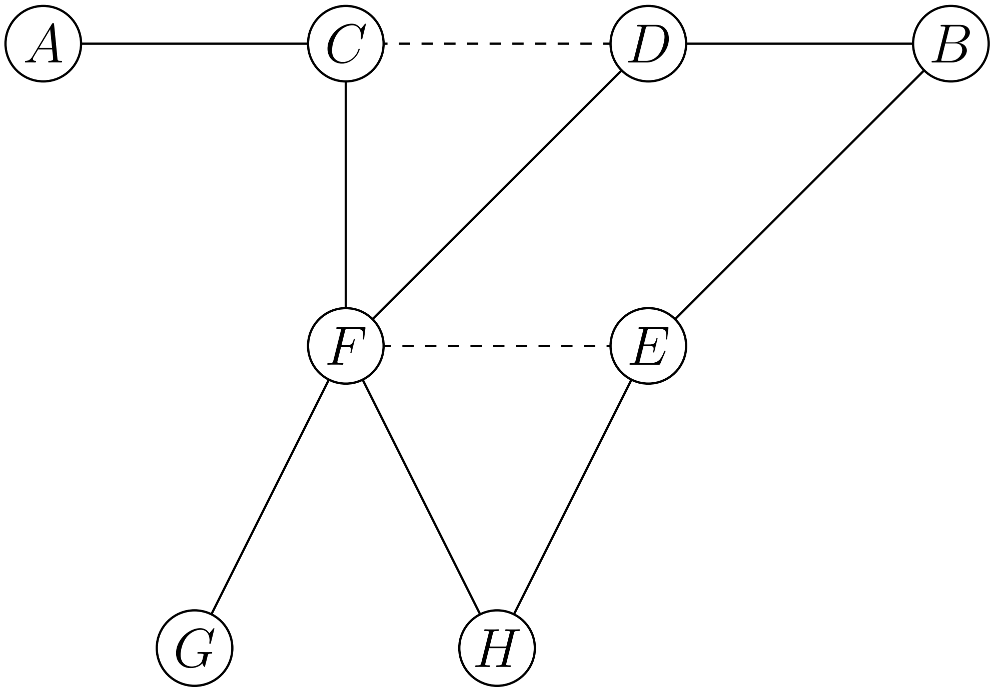
Let $G = (V, E)$ be a directed acyclic graph. If $u, w \in W$ are parents of $v \in V$ connect u and w with an (arbitrarily oriented) edge. After the removal of all edge directions the resulting graph $G_m = (V, E')$ is called the *moral graph* of G.

Join-Tree Construction (1)

Given directed graph.

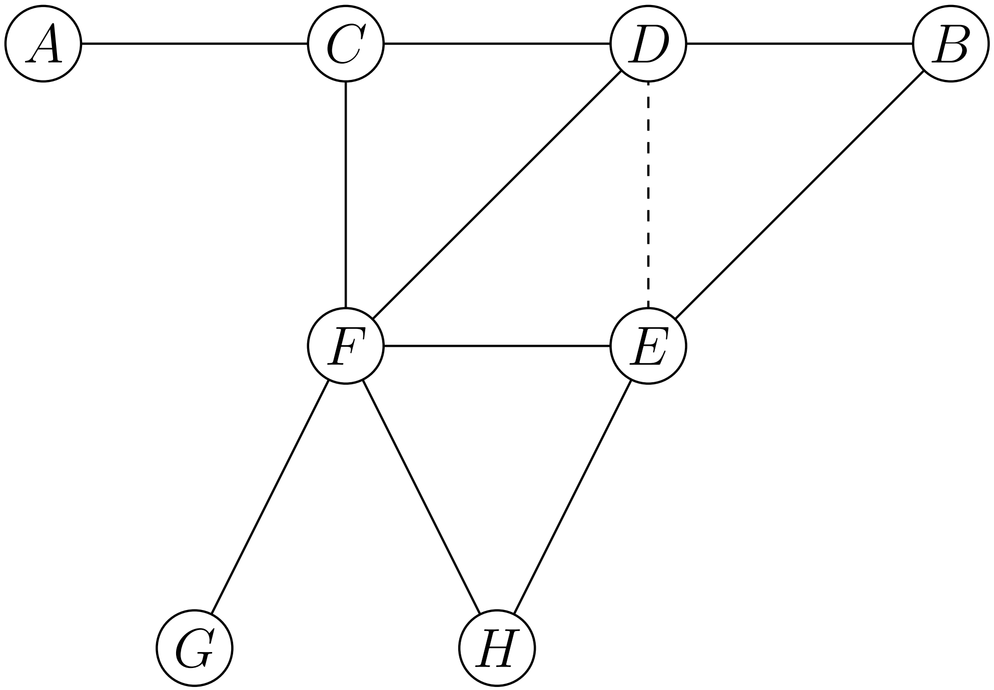


Join-Tree Construction (2)



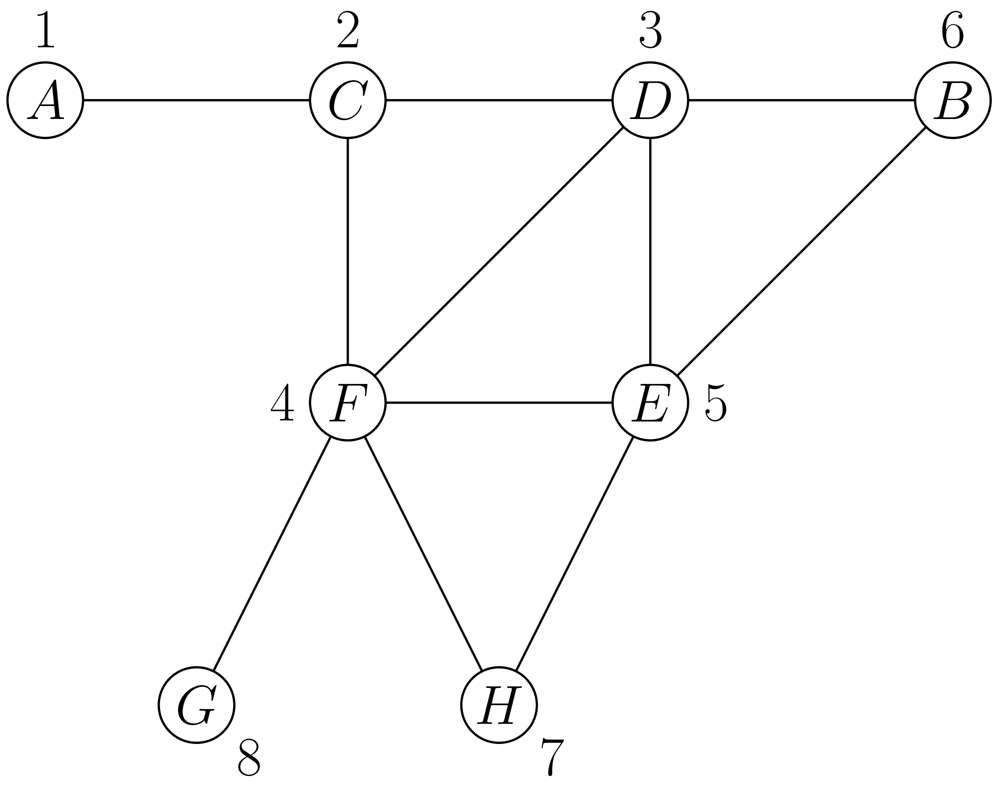
- Moral graph

Join-Tree Construction (3)



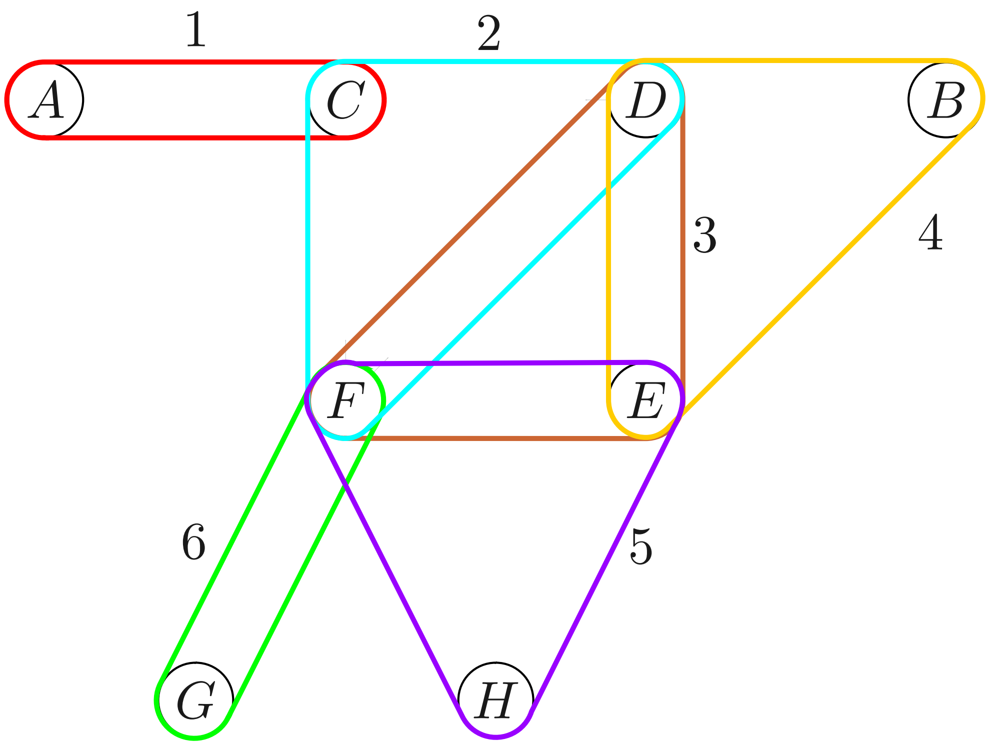
- Moral graph
- Triangulated graph

Join-Tree Construction (4)



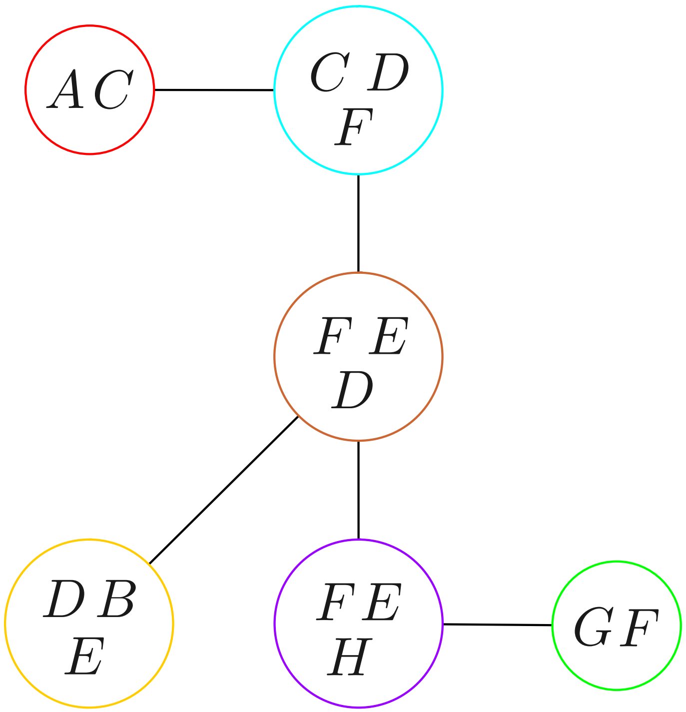
- Moral graph
- Triangulated graph
- MCS yields perfect ordering

Join-Tree Construction (5)



- Moral graph
- Triangulated graph
- MCS yields perfect ordering
- Clique order has RIP

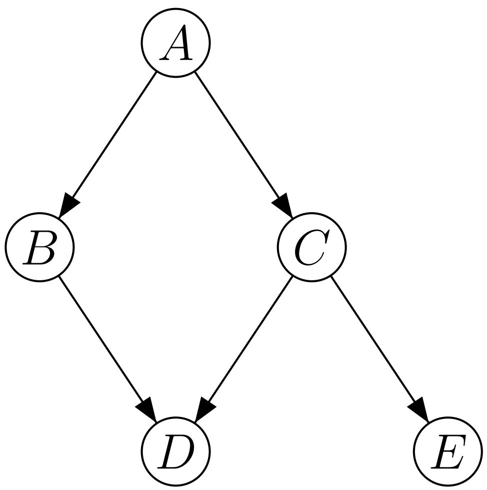
Join-Tree Construction (6)



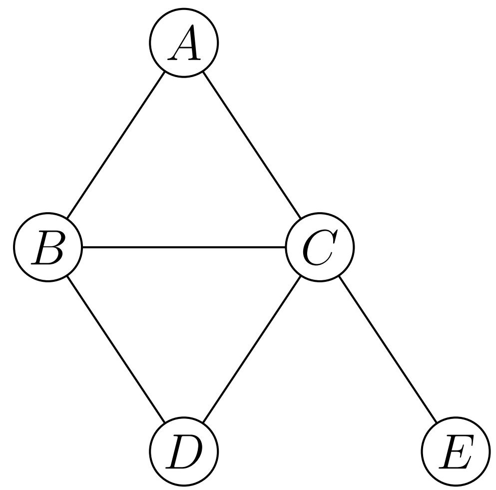
- Moral graph
- Triangulated graph
- MCS yields perfect ordering
- Clique order has RIP
- Form a join-tree

  Two cliques can be connected if they have a non-empty intersection. The generation of the tree follows the RIP. In case of a tie, connect cliques with the largest intersection. (e. g. DBE—FED instead of DBE—CFD) Break remaining ties arbitrarily.

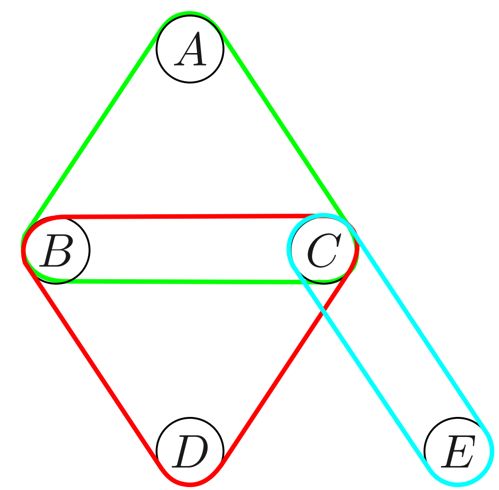
Example: Metastatic Cancer



Dependencies



Moralization/Triangulation



MCS, hyper graph

Clique tree with separator sets

Propagation on Cliques (2)

Quantitative knowledge:

(a, b, c)	$P(a, b, c)$
a_1, b_1, c_1	0.032
a_2, b_1, c_1	0.008
$\vdots$	$\vdots$
a_2, b_2, c_2	0.608

(b, c, d)	$P(b, c, d)$
b_1, c_1, d_1	0.032
b_2, c_1, d_1	0.032
$\vdots$	$\vdots$
b_2, c_2, d_2	0.608

(c, e)	$P(b, c, d)$
c_1, e_1	0.064
c_2, e_1	0.552
c_1, e_2	0.016
c_2, e_2	0.368

Potential representation:

$$\begin{aligned} P(A, B, C, D, E,) &= P(A \mid \emptyset)P(B \mid A)P(C \mid A)P(B \mid BC)P(E \mid C) \\ &= \frac{P(A, B, C)P(B, C, D), P(C, E)}{P(BC)P(C)} \end{aligned}$$

Propagation on Cliques (3)

Propagation:

- $P(d_1) = 0.32, \quad \text{evidence } E = e_1, \quad \text{desired:} \quad P^*(\ldots) = P(\cdot \mid \{e_1\})$

$$P^*(c) = P(c \mid e_1) \qquad \text{conditional marginal distribution}$$

$$P^*(b,c,d) = \frac{P(b,c,d)}{P(c)} P^*(c) \qquad \text{multipl./division with separation prob.}$$

$$P(b,c), \quad P^*(b,c) \qquad \text{calculate marginal distributions}$$

$$P^*(a,b,c) = \frac{P(a,b,c)}{P(b,c)} P^*(b,c) \qquad \text{multipl./division with separation prob.}$$

$$P^*(d_1) = P(d_1 \mid e_1) = 0.33$$

Potential Representation

Let $V = \{X_j\}$ be a set of random variables $X_j : \Omega \to \text{dom}(X_j)$ and P the joint distribution over V. Further, let

$$\{W_i \mid W_i \subseteq V, 1 \leq i \leq p\}$$

a family of subsets of V with associated functions

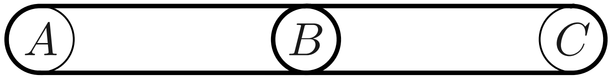
$$\Psi_i : \bigtimes_{X_j \in W_i} \text{dom}(X_j) \to \mathbb{R}$$

It is said that $P(V)$ *factorizes* according $\big(\{W_1, \ldots, W_p\}, \{\Psi_1, \ldots, \Psi_p\}\big)$ if $P(V)$ can be written as:

$$P(v) = k \cdot \prod_{i=1}^{p} \Psi_i(w_i)$$

where $k \in \mathbb{R}$, w_i is a realization of W_i that meets the values of v.

Example



$V = \{A, B, C\},\ W_1 = \{A, B\},\ W_2 = \{B, C\}$

$\mathrm{dom}(A) = \{a_1, a_2\}$
$\mathrm{dom}(B) = \{b_1, b_2\}$
$\mathrm{dom}(C) = \{c_1, c_2\}$

$P(a, b, c) = \frac{1}{8}$

$\Psi_1 : \{a_1, a_2\} \times \{b_1, b_2\} \to \mathbb{R}$
$\Psi_2 : \{b_1, b_2\} \times \{c_1, c_2\} \to \mathbb{R}$

$\Psi_1(a, b) = \frac{1}{4}$
$\Psi_2(b, c) = \frac{1}{2}$

$(\{W_1, W_2\}, \{\Psi_1, \Psi_2\})$ is a potential representation of P.

Factorization of a Belief Network

Let (V, E, P) be an belief network and $\{C_1, \dots, C_p\}$ the cliques of the join tree. For every node $v \in V$ choose a clique C such that v and all of its parents are contained in C, i. e. $\{v\} \cup c(v) \subseteq C$. The chosen clique is designated as $f(v)$.

To arrive at a factorization $(\{C_1, \dots, C_p\}, \{\Psi_1, \dots, \Psi_p\})$ of P the factor potentials are:

$$\Psi_i(c_i) = \prod_{v: f(v) = C_i} P(v \mid c(v))$$

Separator Sets and Residual Sets

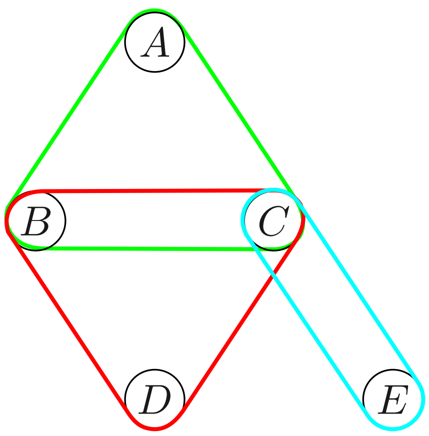
Let $\{C_1, \dots, C_p\}$ be a set of cliques w. r. t. V. The sets

$$S_i = C_i \cap (C_1 \cup \dots \cup C_{i-1}), \qquad i = 1, \dots, p, \qquad S_1 = \emptyset$$

are called *separator sets* with their corresponding *residual sets*

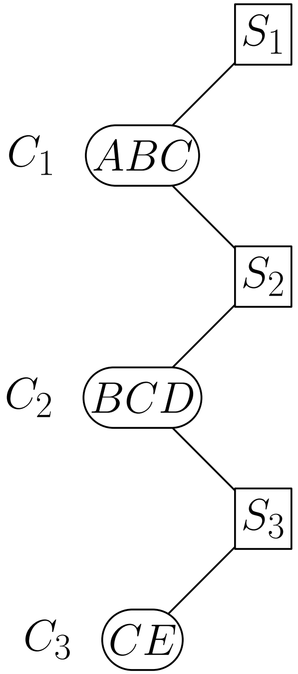
$$R_i = C_i \backslash S_i$$

Example



$$S_1 = \emptyset \qquad R_1 = \{A, B, C\}$$
$$S_2 = \{B, C\} \qquad R_2 = \{D\}$$
$$S_3 = \{C\} \qquad R_3 = \{E\}$$

$$f(A) = C_1$$
$$f(B) = C_1$$
$$f(C) = C_1$$
$$f(D) = C_2$$
$$f(E) = C_3$$



$$\Psi_1(C_1) = P(A, B, C \mid \emptyset) = P(A) \cdot P(C \mid A) \cdot P(B|A)$$
$$\Psi_2(C_2) = P(D \mid B, C)$$
$$\Psi_3(C_3) = P(E \mid C)$$

Propagation is accomplished by sending π- and λ-messages across the cliques in the tree. The emerging potentials are maintained by each clique.

Learning Graphical Models

A (simple) Learning Approach

What does lerning mean?

- **Given:** A database D with samples over a set of attributes V.
- **Desired:** A network over V for which D is maximal probable, i. e. that describes best the data.

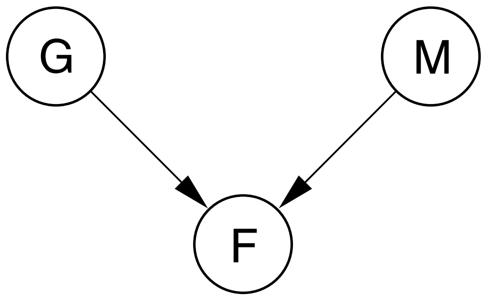
Alternative definition of a Bayesian network:

$$B = (B_S, B_P)$$

B_S	Structure:	The graph encoding the (in)dependencies
B_P	Parameters:	The entries of the potential tables, i. e. the conditional probabilities.

$A_1 = \mathsf{G}$	$Q_{11} = \phi$
$a_{11} = \mathsf{g}$	
$a_{12} = \overline{\mathsf{g}}$	

$A_2 = \mathsf{M}$	$Q_{21} = \phi$
$a_{12} = \mathsf{m}$	
$a_{22} = \overline{\mathsf{m}}$	



$A_3 = \mathsf{F}$	$Q_{31} = (\mathsf{g}, \mathsf{m})$	$Q_{32} = (\mathsf{g}, \overline{\mathsf{m}})$	$Q_{33} = (\overline{\mathsf{g}}, \mathsf{m})$	$Q_{34} = (\overline{\mathsf{g}}, \overline{\mathsf{m}})$
$a_{31} = \mathsf{f}$				
$a_{32} = \overline{\mathsf{f}}$				

- $V = \{\mathsf{G}, \mathsf{M}, \mathsf{F}\}$
- $\mathrm{dom}(\mathsf{G}) = \{\mathsf{g}, \overline{\mathsf{g}}\}$
- $\mathrm{dom}(\mathsf{M}) = \{\mathsf{m}, \overline{\mathsf{m}}\}$
- $\mathrm{dom}(\mathsf{F}) = \{\mathsf{f}, \overline{\mathsf{f}}\}$

- The potential tables' layout is determined by the graph structure.
- The parameters (i. e. the table entries) can be easily estimated from the database, e. g.:

$$\hat{P}(\mathsf{f} \mid \mathsf{g}, \mathsf{m}) = \frac{\#(\mathsf{F} = \mathsf{f}, \mathsf{G} = \mathsf{g}, \mathsf{M} = \mathsf{m})}{\#(\mathsf{G} = \mathsf{g}, \mathsf{M} = \mathsf{m})}$$

Likelihood of a database

Flu G Malaria M Fever F	$\overline{g}$ $\overline{m}$ $\overline{f}$	$\overline{g}$ $\overline{m}$ f	$\overline{g}$ m $\overline{f}$	$\overline{g}$ m f	g $\overline{m}$ $\overline{f}$	g $\overline{m}$ f	g m $\overline{f}$	g m f
#	34	6	2	8	16	24	0	10

Database D with 100 entries for 3 attributes.

$$P(D \mid B_S, B_P) = \prod_{h=1}^{100} P(c_h \mid B_S, B_P)$$

$$= \underbrace{\overbrace{P(\mathsf{g},\mathsf{m},\mathsf{f})}^{\text{Case 1}} \cdot \dots \cdot \overbrace{P(\mathsf{g},\mathsf{m},\mathsf{f})}^{\text{Case 10}}}_{\text{10 times}} \quad \cdots \quad \underbrace{\overbrace{P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f})}^{\text{Case 51}} \cdot \dots \cdot \overbrace{P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f})}^{\text{Case 58}}}_{\text{8 times}} \quad \cdots \quad \underbrace{\overbrace{P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}})}^{\text{Case 67}} \cdot \dots \cdot \overbrace{P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}})}^{\text{Case 100}}}_{\text{34 times}}$$

$$= P(\mathsf{g},\mathsf{m},\mathsf{f})^{10} \quad \cdots \quad P(\overline{\mathsf{g}},\mathsf{m},\mathsf{f})^{8} \quad \cdots \quad P(\overline{\mathsf{g}},\overline{\mathsf{m}},\overline{\mathsf{f}})^{34}$$

$$= P(\mathsf{f} \mid \mathsf{g},\mathsf{m})^{10} P(\mathsf{g})^{10} P(\mathsf{m})^{10} \quad \cdots \quad P(\mathsf{f} \mid \overline{\mathsf{g}},\mathsf{m})^{8} P(\overline{\mathsf{g}})^{8} P(\mathsf{m})^{8} \quad \cdots \quad P(\overline{\mathsf{f}} \mid \overline{\mathsf{g}},\overline{\mathsf{m}})^{34} P(\overline{\mathsf{g}})^{34} P(\overline{\mathsf{m}})^{34}$$

Likelihood of a database (2)

$$P(D \mid B_S, B_P) \quad = \quad \prod_{h=1}^{100} P(c_h \mid B_S, B_P)$$

$$= P(\mathsf{f} \mid \mathsf{g}, \mathsf{m})^{10} P(\bar{\mathsf{f}} \mid \mathsf{g}, \mathsf{m})^{0} P(\mathsf{f} \mid \mathsf{g}, \overline{\mathsf{m}})^{24} P(\bar{\mathsf{f}} \mid \mathsf{g}, \overline{\mathsf{m}})^{16}$$

$$\cdot\, P(\mathsf{f} \mid \overline{\mathsf{g}}, \mathsf{m})^{8} P(\bar{\mathsf{f}} \mid \overline{\mathsf{g}}, \mathsf{m})^{2} P(\mathsf{f} \mid \overline{\mathsf{g}}, \overline{\mathsf{m}})^{6} P(\bar{\mathsf{f}} \mid \overline{\mathsf{g}}, \overline{\mathsf{m}})^{34}$$

$$\cdot\, P(\mathsf{g})^{50} P(\overline{\mathsf{g}})^{50} P(\mathsf{m})^{20} P(\overline{\mathsf{m}})^{80}$$

The last equation shows the principle of reordering the factors:

- First, we sort by attributes (here: F, G then M).
- Within the same attributes, factors are grouped by the parent attributes' values combinations (here: for F: (g, m), $(\mathsf{g}, \overline{\mathsf{m}})$, $(\overline{\mathsf{g}}, \mathsf{m})$ and $(\overline{\mathsf{g}}, \overline{\mathsf{m}})$).
- Finally, it is sorted by attribute values (here: for F: first f, then $\bar{\mathsf{f}}$).

Likelihood of a database (3)

General likelihood of a database D:

$$P(D \mid B_S, B_P) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}}$$

General potential table:

A_i	Q_{i1}	$\cdots$	Q_{ij}	$\cdots$	Q_{iq_i}
a_{i1}	θ_{i11}	$\cdots$	θ_{ij1}	$\cdots$	θ_{iq_i1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_{ik}	θ_{i1k}	$\cdots$	θ_{ijk}	$\cdots$	θ_{iq_ik}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a_{ir_i}	θ_{i1r_i}	$\cdots$	θ_{ijr_i}	$\cdots$	$\theta_{iq_ir_i}$

A (simple) Learning Approach (2)

Back to our inital question: How to find the structure that yields the highest likelihood of the database D?

$$\widehat{B}_S = \operatorname*{arg\,max}_{B_S \in \mathcal{B}_V} P(D \mid B_S, B_P)$$

$\mathcal{B}_V$ designates the set of all directed, acyclic graphs with V as the set of nodes.

Flaws of this approch:

- Inserting edges cannot lower the likelihood, i. e. the result of a maximum likelihood approch will always be a fully connected graph.
- The set $\mathcal{B}_V$ grows over-exponentially in $|V|$.

$\Rightarrow$ Assumptions and heuristics needed!

Learning Approaches

(A) Test whether a candidate graph decomposes the distribution/relation

(B) Conditional independence tests

(C) Measure marginal independence strengths

Since the search space $\mathcal{B}_V$ is too large, we cannot exhaustively enumerate all candidate graphs.

$\Rightarrow$ Search algorithms needed, consisting of

- an evaluation measure (to measure the "fitness" of the current solution candidate)
- a search heuristic to traverse $\mathcal{B}_V$, e. g.:
    - random-guided search (e. g. generic algorithms)
    - greedy search (presented later)

Example for (A): Test for Decomposition

Given a solution candidate $B_S \in \mathcal{B}_V$, how good does it explain the database D?

- Compare the distribution defined by B_S with the given empirical distribution of D.
- If both are identical, a solution B_S has been found.

However, in most (real) cases, there is no exact decomposition, so we have to find the candidate B_S that approximates best the distribution of D.

$\Rightarrow$ Measure for the quality of approximation between distributions needed.

Kullback-Leibler cross entropy

Let $(\Omega, 2^\Omega, P)$ and $(\Omega, 2^\Omega, P^*)$ be two finite probability spaces. Then

$$I_{\mathrm{KLdiv}}(P, P^*) = \sum_{\omega \in \Omega} P(\omega) \cdot \log_2 \frac{P(\omega)}{P^*(\omega)}$$

is called the *Kullback-Leiber cross entropy* of P and P^*.

Remark:

$$I_{\mathrm{KLdiv}}(P, P^*) \geq 0; \quad I_{\mathrm{KLdiv}}(P, P^*) = 0 \;\Leftrightarrow\; P \equiv P^*$$

Where does this this equation come from?

Information Content

The information content of a message ω that occurs with probability $p(\omega)$ is defined as

$$\mathrm{Inf}(\omega) = -\log_2 p(\omega).$$

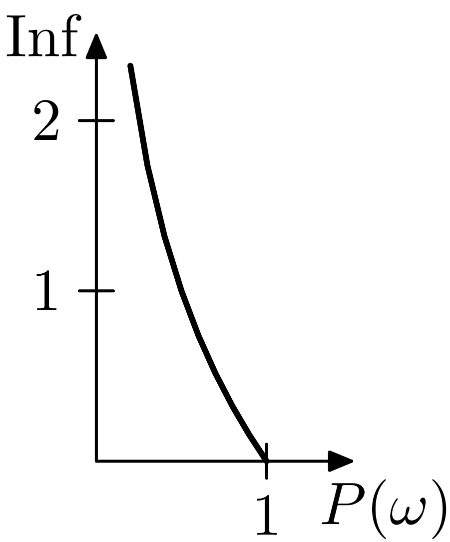
Intention:

- Neglect all subjective references to ω and let the information content be determined by $p(\omega)$ only.
- The information of a certain message ($p(\omega) = 1$) is zero.
- The less frequent a message occurs (i. e., the less probable it is), the more interesting is the fact of its occurrence:

$$p(\omega_1) < p(\omega_2) \quad \Rightarrow \quad \mathrm{Inf}(\omega_1) > \mathrm{Inf}(\omega_2)$$

- We only use one bit to encode the occurrence of a message with probability $\frac{1}{2}$.

The function Inf fulfills all these requirements.



- The set of all messages Ω can be considered a set of elementary events.
- Then Inf becomes a random variable, the expected value of which can be written as follows:

$$E(\mathrm{Inf}) = -\sum_{\omega \in \Omega} p(\omega) \cdot \log_2 p(\omega) \stackrel{\mathrm{Def}}{=} H(P)$$

Shannon Entropy

Let $(\Omega, 2^\Omega, P)$ be a probability space. Then,

$$H^{(\text{Shannon})}(P) = -\sum_{\omega\in\Omega} P(\omega)\log_2 P(\omega)$$

is called the *Shannon entropy* of P, where $0 \cdot \log_2 0 = 0$ is assumed.

- $H^{(\text{Shannon})}(P)$ is the expected value (in bits) of the information content that is related to the occurrence of the elementary events $\omega \in \Omega$.

$$H^{(\text{Shannon})}(P) = \sum_{\omega\in\Omega} \underbrace{P(\omega)}_{\text{Probability of } \omega} \cdot \underbrace{(-\log_2 P(\omega))}_{\substack{\text{Information content of } \omega \text{ (identi-} \\ \text{fication of outcome } \omega \text{ out of } \frac{1}{P(\omega)} \\ \text{outcomes).}}}$$

Excursus: Approximation Measure

- We could define $D(P, P^*)$ as the degree that P is approximated by P^* in the following way:

$$D(P, P^*) \;=\; H^{(\text{Shannon})}(P^*) \;-\; H^{(\text{Shannon})}(P)$$

- Assume two variables X and Y with the joint distribution $P(X, Y)$.

- Further let

$$P^*(X, Y) \;=\; P(X) \cdot P(Y)$$

be the joint distribution in the case of independence.

$$H^{(\text{Sh.})}(P) = -\sum_{(x,y)\,\in\,\Omega_X \times \Omega_Y} P(x, y) \log_2 P(x, y)$$

Back to: Kullback-Leibler

$$
\begin{aligned}
H^{(\mathrm{Sh.})}(P^*) &= -\sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x)P(y)\log_2(P(x)P(y)) \\
&= -\sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x)P(y)\log_2 P(x) - \sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x)P(y)\log_2 P(y) \\
&= -\sum_{x\,\in\,\Omega_X} P(x)\log_2 P(x) - \sum_{y\,\in\,\Omega_Y} P(y)\log_2 P(y) \\
&= -\sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x,y)\log_2 P(x) - \sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x,y)\log_2 P(y) \\
&= -\sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x,y)\log_2(P(x)P(y))
\end{aligned}
$$

Therefore:

$$
D(P,P^*) \;=\; I_{\mathrm{KLdiv}}(P,P^*) \qquad = \sum_{(x,y)\,\in\,\Omega_X\times\Omega_Y} P(x,y)\cdot\log_2\frac{P(x,y)}{P(x)P(y)}
$$

Example for (B): Conditional Independence Tests

- Find an independence map B_S of the given database distribution.
- Measure the degree of independence between attributes by using the Kullback-Leibler cross entropy.

To measure the strength of dependence of two attributes A and B, we simply compare the joint distribution $P(A, B)$ with the distribution in the case of independence $P(A) \cdot P(B)$.

Mutual (Shannon) Information

Let A and B be two attributes and P a strictly positive probability measure. Then

$$I_{\text{mut}}(A, B) = \sum_{a \in \text{dom}(A)} \sum_{b \in \text{dom}(B)} P(A = a, B = b) \log_2 \frac{P(A = a, B = b)}{P(A = a) \cdot P(B = b)}$$

is called the *mutual (Shannon) information* or *(Shannon) cross entropy* of A and B w. r. t. P.

Example for (B): Conditional Independence Tests

Note, I_{mut} is also referred to as *Shannon information gain*.

To measure the strength of conditional independence,
we generalize I_{mut}:

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

We can now use the equation above to estimate attribute (in)dependencies and use this information while constructing an independence map.

Example for (C): Marginal Dependencies

- **Given:** A belief network (V, E, P) where only V and $P(V)$ are known.
  $P(V)$ may be estimated from data.
- **Desired:** Belief tree (V, E^*, P^*) for which P is approximated best by P^*.

Steps to determine (V, E^*, P^*)

1. For tree $T = (V, E')$ determine (V, E', P_T) with

$$D(P, P_T) \;=\; \min\{D(P, P') \mid (V, E', P') \text{ is belief tree}\}$$

   (P_T is the projection of P on T)

2. Determine a belief tree (V, E^*, P^*) with

$$D(P, P^*) \;=\; \min\{D(P, P_T) \mid \ T \text{ is tree with node set } V\}$$

   (P_T is the *projection* of P on T with $\forall X \in V: \ P_T(X \mid c(X)) = P(X \mid c(X))$
   where $c(X)$ denotes the direct predecessor (parent) of X.)

Example for (C): Marginal Dependencies

Chow, Liu 1968

$D(P, P^*)$ is minimal w. r. t. to step 2 of enumeration on the previous slide if P^* is a projection of P on a MWST (maximum weight spanning tree), in which the weight of every edge $(X, Y) \in E^*$ is defined by

$$I(X, Y) \stackrel{\text{Def}}{=} \sum_{(x,y)\,\in\,\Omega_X \times \Omega_Y} P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)} \geq 0$$

If (V, E, P) is a belief tree, then the projection P_T on every MWST $T = (V, E')$ coincides with P.

Construction of a MWST

1. Determine $P(X, Y)$ for all $(X, Y) \in V \times V$, with $X \neq Y$
2. Calculate all $\frac{n(n-1)}{2}$ edge weights $I(X, Y)$
3. Assign two edges with the highest weights to the tree (V, E^*) under construction.
4. Assign to (V, E^*) an edge not yet assigned with highest weight without forming a loop.
5. Repeat step 4 until $n - 1$ egdes have been assigned (the MWST is then constructed).
6. Determine P^* with Chow-Liu theorem. This results in the desired belief tree (V, E^*, P^*).

Example for (C): Marginal Dependencies

Remarks:

- MWST construction requires $O(|V|^2)$ steps.
- P^* is a maximum likelihood estimation of P, if estimated from a given database
- Disadvantage: algorithm only efficient on tree-like structures.
  However, after extention polytrees are constructable as well.

K2 Algorithm

- Proposed by [Cooper and Herskovits 1992]
- Greedy algorithm (category (C))
- Uses the K2 metric to evaluate the quality of a candidate graph

$$\begin{aligned}\widehat{B}_S &= \arg\max_{B_S} P(B_S \mid D) = \arg\max_{B_S} \frac{P(B_S, D)}{P(D)} \\ &= \arg\max_{B_S} P(B_S, D)\end{aligned}$$

$\Rightarrow$ Find an equation for $P(B_S, D)$.

Model Averaging

We first consider $P(B_S, D)$ to be the marginalization of $P(B_S, B_P, D)$ over all possible parameters B_P.

$$
\begin{aligned}
P(B_S, D) &= \int_{B_P} P(B_S, B_P, D)\,\mathrm{d}B_P \\
&= \int_{B_P} P(D \mid B_S, B_P)\,P(B_S, B_P)\,\mathrm{d}B_P \\
&= \int_{B_P} P(D \mid B_S, B_P)\,f(B_P \mid B_S)P(B_S)\,\mathrm{d}B_P \\
&= \underbrace{P(B_S)}_{\text{A priori prob.}} \int_{B_P} \underbrace{P(D \mid B_S, B_P)}_{\text{Likelihood of } D}\ \underbrace{f(B_P \mid B_S)}_{\text{Parameter densities}}\ \mathrm{d}B_P
\end{aligned}
$$

K2 Algorithm

- The a priori distribution $P(B_S)$ can be used to bias the evaluation measure towards user-specific network structures.
- Substitute the likelihood definition:

$$P(B_S, D) \;=\; P(B_S) \int_{B_P} \left[\prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}} \right] f(B_P \mid B_S) \, \mathrm{d}B_P$$

K2 Algorithm

- The parameter densities $f(B_P \mid B_S)$ describe the probabilities of the parameters given a network structure. They are densities of second order (distribution over distributions)
- For fixed i and j, a vector $(\theta_{ij1}, \ldots, \theta_{ijr_i})$ represents a probability distribution, namely the j-th column of the i-th potential table.
- Assuming mutual independence between the potential tables, we arrive for $f(B_P \mid B_S)$ at the following:

$$f(B_P \mid B_S) \quad = \quad \prod_{i=1}^{n} \prod_{j=1}^{q_i} f(\theta_{ij1}, \ldots, \theta_{ijr_i})$$

Thus, we can further concretize the equation for $P(B_S, D)$:

$$
\begin{aligned}
P(B_S, D) \;&=\; P(B_S) \int_{\theta_{ijk}} \cdots \int \left[\prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}} \right] \cdot \left[\prod_{i=1}^{n} \prod_{j=1}^{q_i} f(\theta_{ij1}, \ldots, \theta_{ijr_i}) \right] \mathrm{d}\theta_{111}, \ldots, \mathrm{d}\theta_{nq_nr_n} \\
&=\; P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\theta_{ijk}} \cdots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}} \right] \cdot f(\theta_{ij1}, \ldots, \theta_{ijr_i}) \, \mathrm{d}\theta_{ij1}, \ldots, \mathrm{d}\theta_{ijr_i}
\end{aligned}
$$

K2 Algorithm

A last assumption: for fixed i and j the density $f(\theta_{ij1}, \ldots, \theta_{ijr_i})$ is uniform:

$$f(\theta_{ij1}, \ldots, \theta_{ijr_i}) \;=\; (r_i - 1)!$$

$$\begin{aligned}
P(B_S, D) \;&=\; P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \int_{\theta_{ijk}} \cdots \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}} \right] \cdot (r_i - 1)! \, \mathrm{d}\theta_{ij1}, \ldots, \mathrm{d}\theta_{ijr_i} \\
&=\; P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} (r_i - 1)! \quad \underbrace{\int_{\theta_{ijk}} \cdots \int \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}} \, \mathrm{d}\theta_{ij1}, \ldots, \mathrm{d}\theta_{ijr_i}}_{\text{Dirichlet's integral} \,=\, \frac{\prod_{k=1}^{r_i} \alpha_{ijk}!}{(\sum_{k=1}^{r_i} \alpha_{ijk} + r_i - 1)!}}
\end{aligned}$$

Thus, we finally arrive at an expression for $P(B_S, D)$ which we identify with the K2 metric of P_S given the data D:

$$P(B_S, D) \quad = \quad \mathrm{K2}(B_S \mid D) \quad = \quad P(B_S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}! \right]$$

$$\text{with} \quad N_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$$

Properties of the K2 metric

- **Global** — Refers to the outer product: the total value of the K2 metric is the product over all K2 values of attribute families.
- **Local** — The likelihood equation assumes that given a parents instantiation, the probabilities for the respective child attribute values are mutual independent. This is reflected in the product over all q_i different parent attributes' value combinations of attribute A_i.

We exploit the global property to write the K2 metric as follows:

$$\mathrm{K2}(B_S \mid D) \quad = \quad P(B_S) \prod_{i=1}^{n} \mathrm{K2}_{\mathrm{local}}(A_i \mid D)$$

with

$$\mathrm{K2}_{\mathrm{local}}(A_i \mid D) \quad = \quad \prod_{j=1}^{q_i} \left[\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}! \right]$$

K2 Algorithm

Prerequisites:

- Choose a topological order on the attributes $(A_1, \ldots, A_n)$
- Start out with a network that consists of n isolated nodes.
- Let q_i be the quality of the i-th attribute given parent attributes M:

$$q_i(M) = \mathrm{K2}_{\mathrm{local}}(A_i \mid D) \quad \text{with} \quad \mathrm{parents}(A_i) = M$$

K2 Algorithm

Execution:

1. Determine for the parentless node A_i the quality measure $q_i(\emptyset)$
2. Evaluate for every predecessor $\{A_1, \dots, A_{i-1}\}$ whether inserted as parent of A_i, the quality measure would increase. Let Y be the node that yields the highest quality.

$$Y = \operatorname*{arg\,max}_{1 \le l \le i-1} q_i(\{A_l\})$$

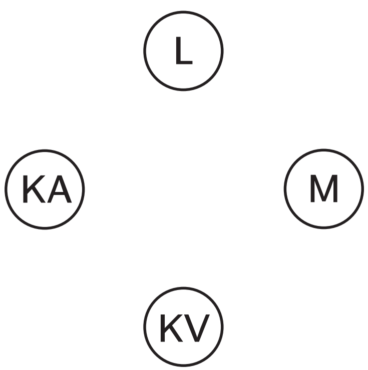
   This best quality measure be $g = q_i(\{Y\})$.

3. If g is better than $q_i(\emptyset)$, Y is inserted permanently as a parent node: $\text{parents}(A_i) = \{Y\}$
4. Repeat steps 2 und 3 to increase the parent set until no quality increase can be achieved or no nodes are left or a predefined maximum number of parent nodes per node is reached.

K2 Algorithm

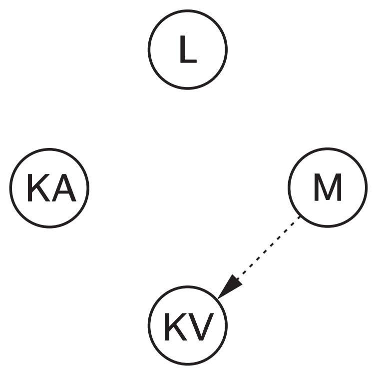
for $i \leftarrow 1 \dots n$ **do** // Initialization
  $\text{parents}(A_i) \leftarrow \emptyset$
end for
for $i \leftarrow n \dots 1$ **do** // Iteration
  repeat
    Select $Y \in \{A_1, \dots, A_{i-1}\} \setminus \text{parents}(A_i)$,
      which maximizes $g = q_i(\text{parents}(A_i) \cup \{Y\})$
    $\delta \leftarrow g - q_i(\text{parents}(A_i))$
    if $\delta > 0$ **then**
      $\text{parents}(A_i) \leftarrow \text{parents}(A_i) \cup \{Y\}$
    end if
  until $\delta \leq 0$ or $\text{parents}(A_i) = \{A_1, \dots, A_{i-1}\}$ or $|\text{parents}(A_i)| = n_{\max}$
end for

Demo of K2 Algorithm



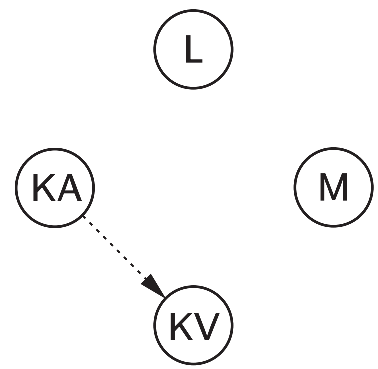
-43942.99

Step 1 – Edgeless graph



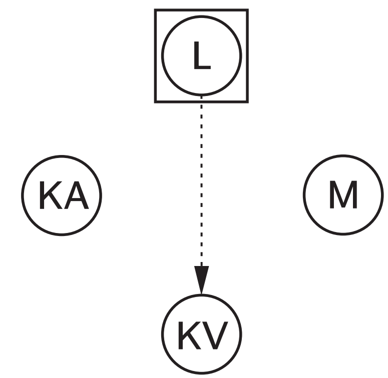
-43950.53

Step 2 – Insert M temporarily.



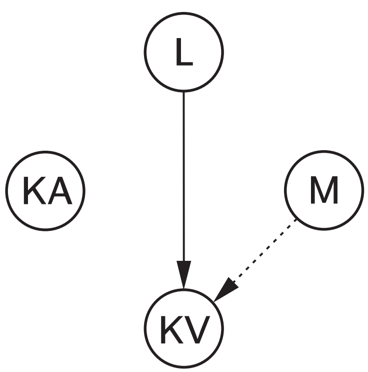
-43949.30

Step 3 – Insert KA temporarily.



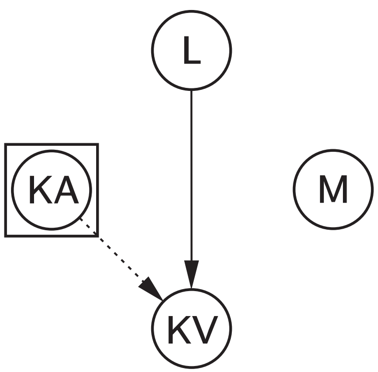
-43942.48

Step 4 – Node L maximizes K2 value and thus is added permantently.



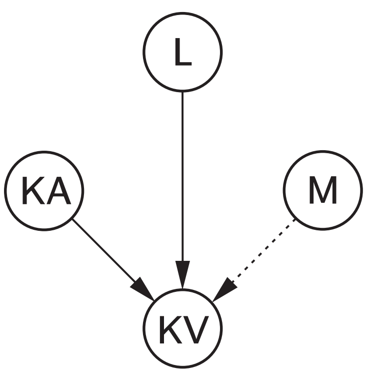
−43964.12

Step 5 – Insert M temporarily.



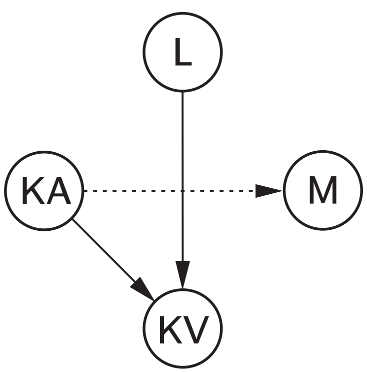
−43933.71

Step 6 – KA is added as second parent node of KV.



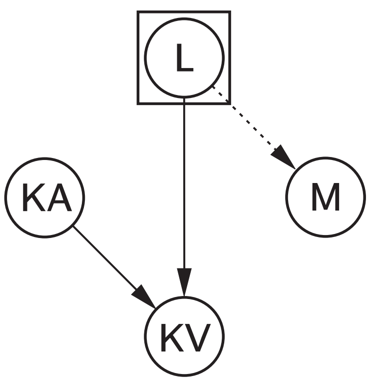
−43995.02

Step 7 – M does not increase the quality of the network if insertes as third parent node.



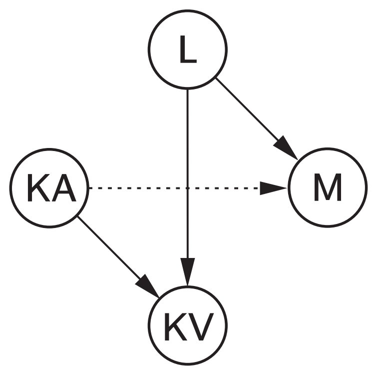
−43492.08

Step 8 – Insert KA temporarily.



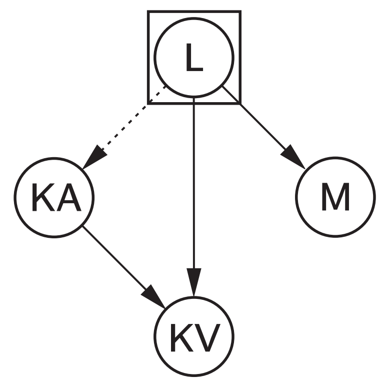
−40823.76

Step 9 – Node L becomes perent node of M.



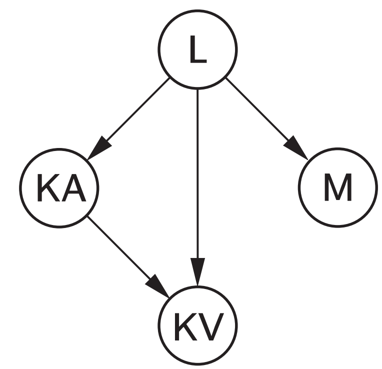
−40903.39

Step 10 – Adding KA does not increase overall network quaility.



−39190.67

Step 11 – Node L becomes parent node of KA.



Result

Decision Graphs / Influence Diagrams

Motivation

Up to now, we used Bayesian networks for

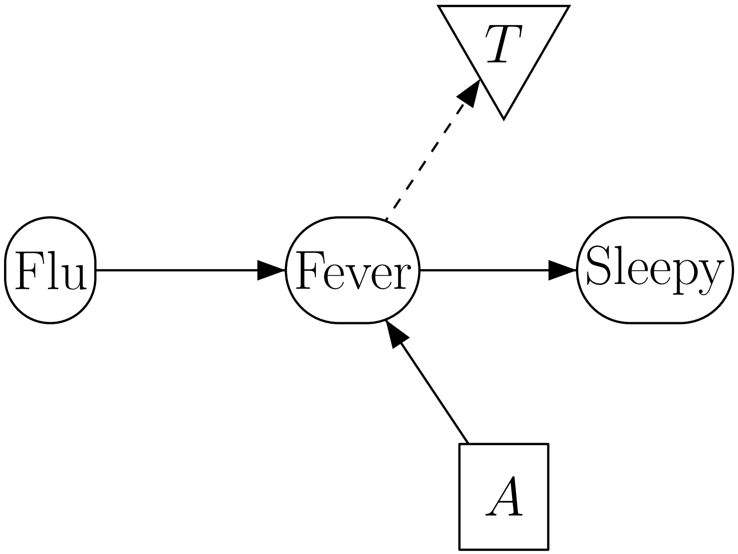
- modeling (in)dependence relations between random/chance variables
- quantifying the strength of these relations by assigning (conditional) probabilities
- update these probabilities after evidence observations

However, in practical, this is only a part of a more complex task: **decision making under uncertainty**.

If a set of actions solves a problem, we have to choose one particular action based on predefined criteria, e. g. costs and/or gains.

Therefore, we will now augment the current framework with special nodes that serve these purposes.

Example: Observations and Actions



T... Temperature

A... Aspirine

- Rectangular nodes: intervening actions/decisions
- Triangular nodes: test actions/observations
- Observations may change probabilities of nodes that are causes:

  Observing $T = 37°C$ decreases probability of Fever and Flu (and, of course, Sleepy).

- The impact of intervening actions can only follow the direction of the (causal) edges:

  Taking Aspirine (A) decreases the probability of Fever and Sleepy and may result in an alike observation for T. However, it cannot change the state for Flu since Aspirine only eases the pain and does not kill viruses.

Example: Utilities

Mildew Fungus Infestation (dt. Mehltau-Befall)

Before the harvest, a farmer checks the state of his crop and decides whether to apply a fungi treatment or not.

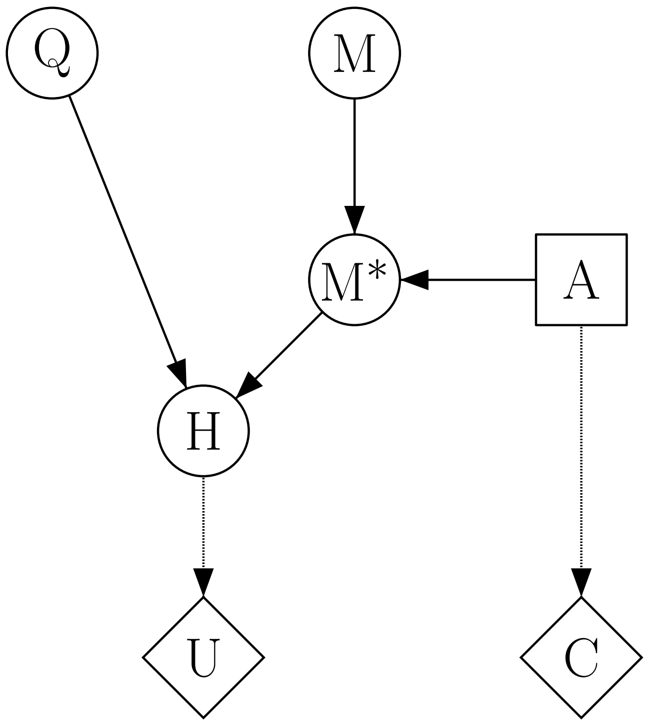
- Q — Quality of the crop
- M — Mildew infestation severity
- H — Harvest quality
- A — Action to be taken
- M* — Mildew infestation after action A
- U — Utility function of the harvest (i. e. the benefit)
- C — Utility functon of the action (i. e. the treatment costs)

edges leading to chance nodes

edges leading to decision nodes

edges leading to utility nodes

Example: Utilities (2)



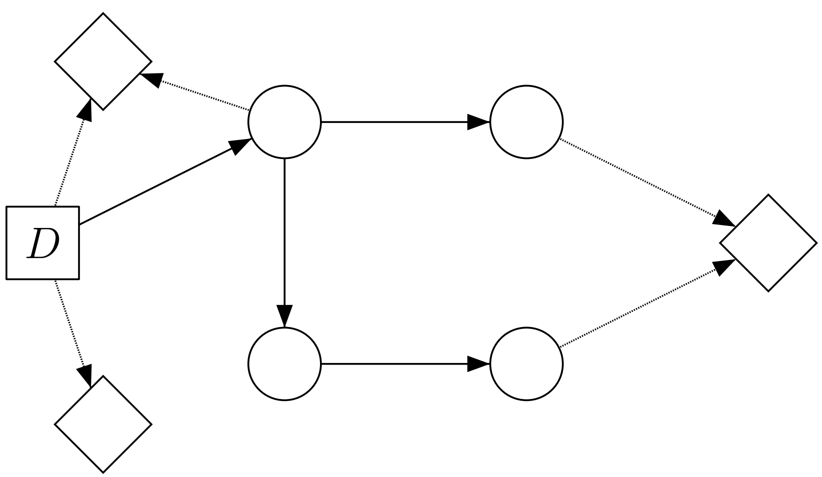
- Diamond-shaped nodes: utility functions (costs/benefits)
- Given the quality of the crops and the mildew state, which action maximizes the benefit?
- $C(\mathrm{A}) < 0$
- $U(\mathrm{H}) \geq 0$
- Expected total utility of action $\mathrm{A} = a$:

$$\mathrm{E}(U(a \mid q, m)) = C(a) + \sum_{h} U(h) \cdot P(h \mid a, q, m)$$

Single-Action Models

A single-action model consists of

- a Bayesian network representing the chance nodes
- one decision (action) node
- a set of utility nodes
- decision nodes can affect chance and utility nodes
- utility nodes can be affected by chance and decision nodes



Given n utility nodes $U_1, \ldots, U_n$ and assuming they all depend on only one respective chance node X_i, the total expected utility given a decision $D = d$ and (chance node) evidence e is defined as:
vskip-2mm

$$\mathrm{E}(U(d \mid e)) = \sum_{i=1}^{n} \sum_{x \in \mathrm{dom}(X_i)} U_1(x_1) \cdot P(x_1 \mid d, e)$$

The optimal decision d^* is then chosen:

$$d^* = \underset{d \in \mathrm{dom}(D)}{\arg\max}\ \mathrm{E}(U(d \mid e))$$

Influence Diagrams

An influence diagram consists of a directed acyclic graph over chance nodes, decision nodes and utility nodes that obey the following structural properties:

- there is a directed path comprising all decision nodes
- utility nodes cannot have children
- decision and chance nodes are discrete
- utility nodes do not have states
- chance nodes are assigned potential tables given their parents (including decision nodes)
- each utility node U gets assigned a real-valued utility function over its parents

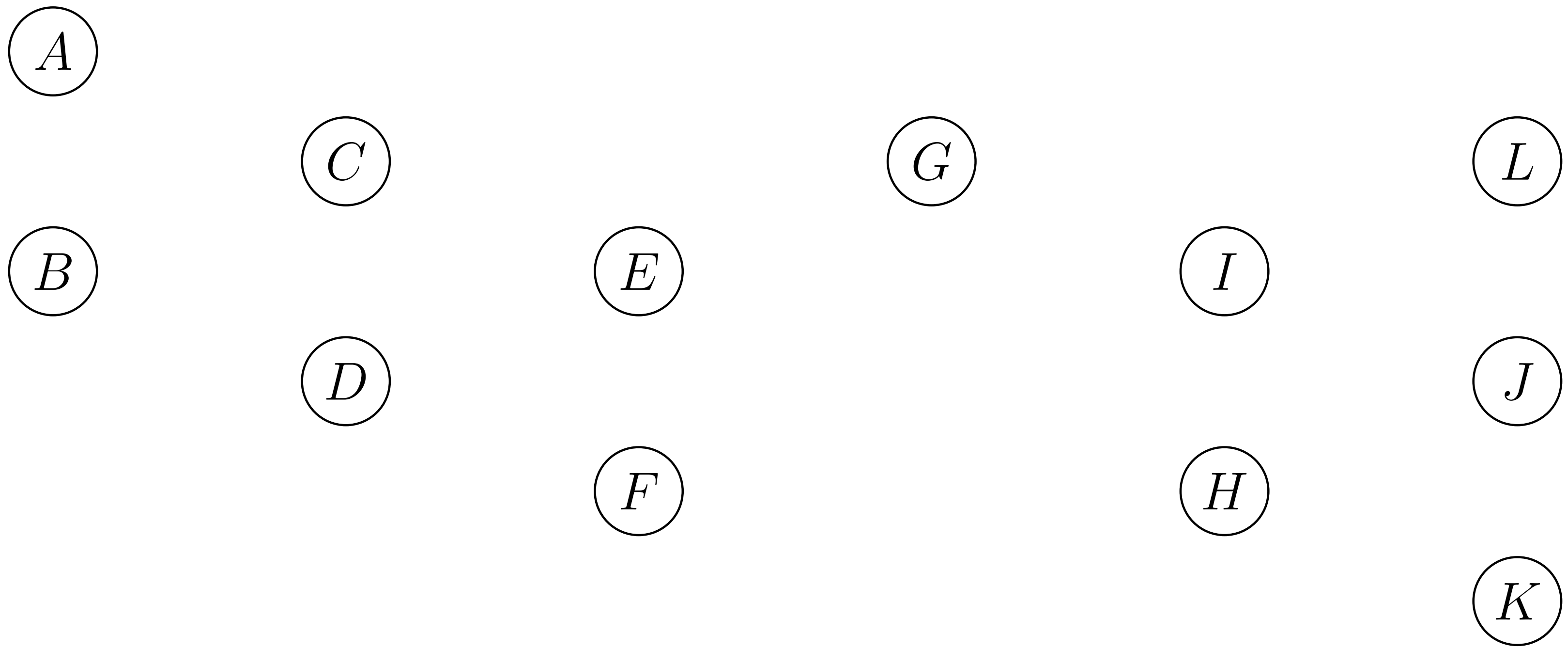
$$U : \bigtimes_{X \in \mathrm{parents}(U)} \mathrm{dom}(X) \to \mathbb{R}$$

Influence Diagrams (2)

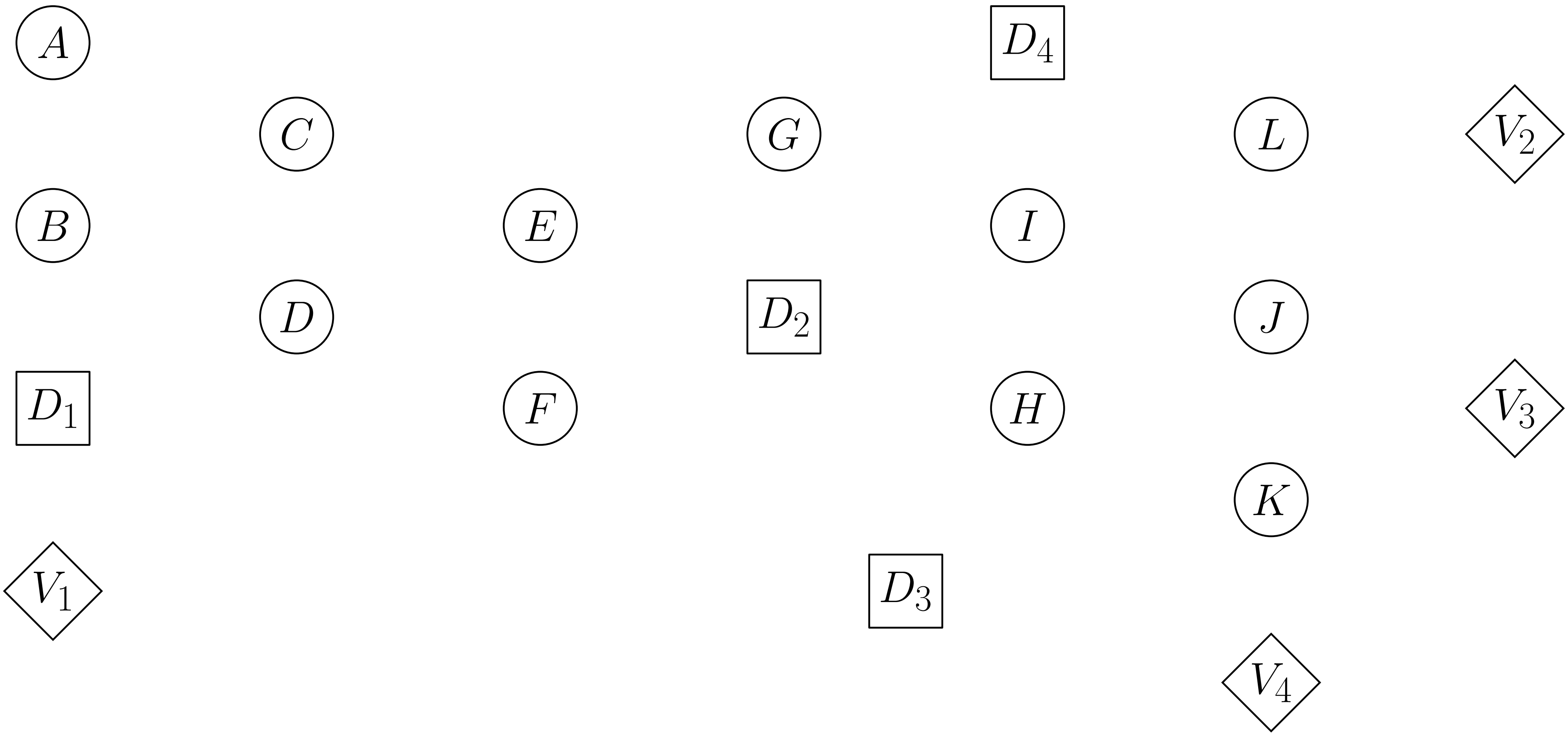
- Links into decision nodes carry no quantitative information, they only introduce a temporal ordering.
- The required path between the decision nodes induces a temporal partition of the chance nodes:

  If there are n decision nodes, then for $1 \leq i < n$ the set I_i represents all chance nodes that have to be observed after decision D_i but before decision D_{i+1}.
- I_0 is the set of chance nodes to be observed before any decision.
- I_n is the set of chance nodes that are not observed.

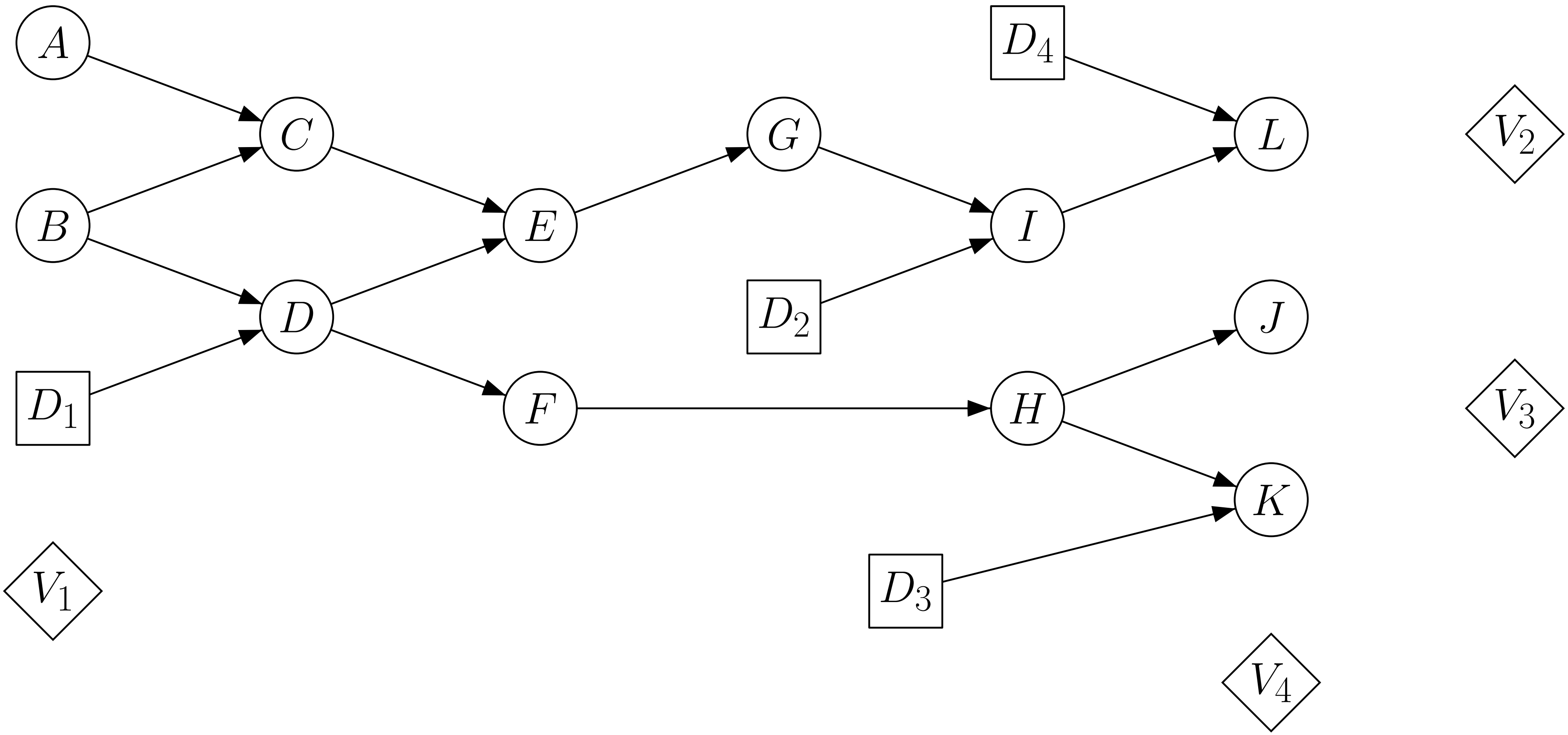
Influence Diagrams (3)



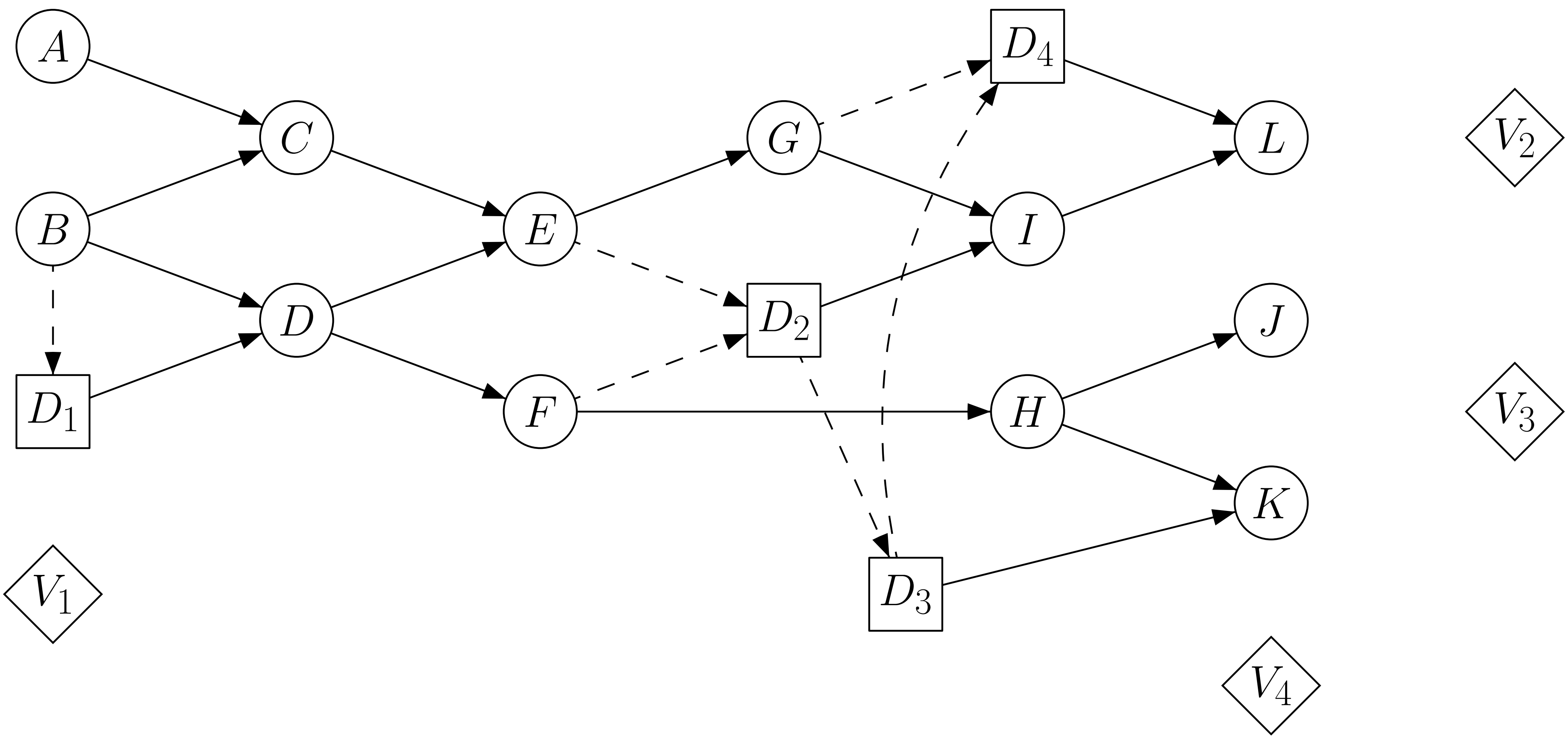
Influence Diagrams (3)



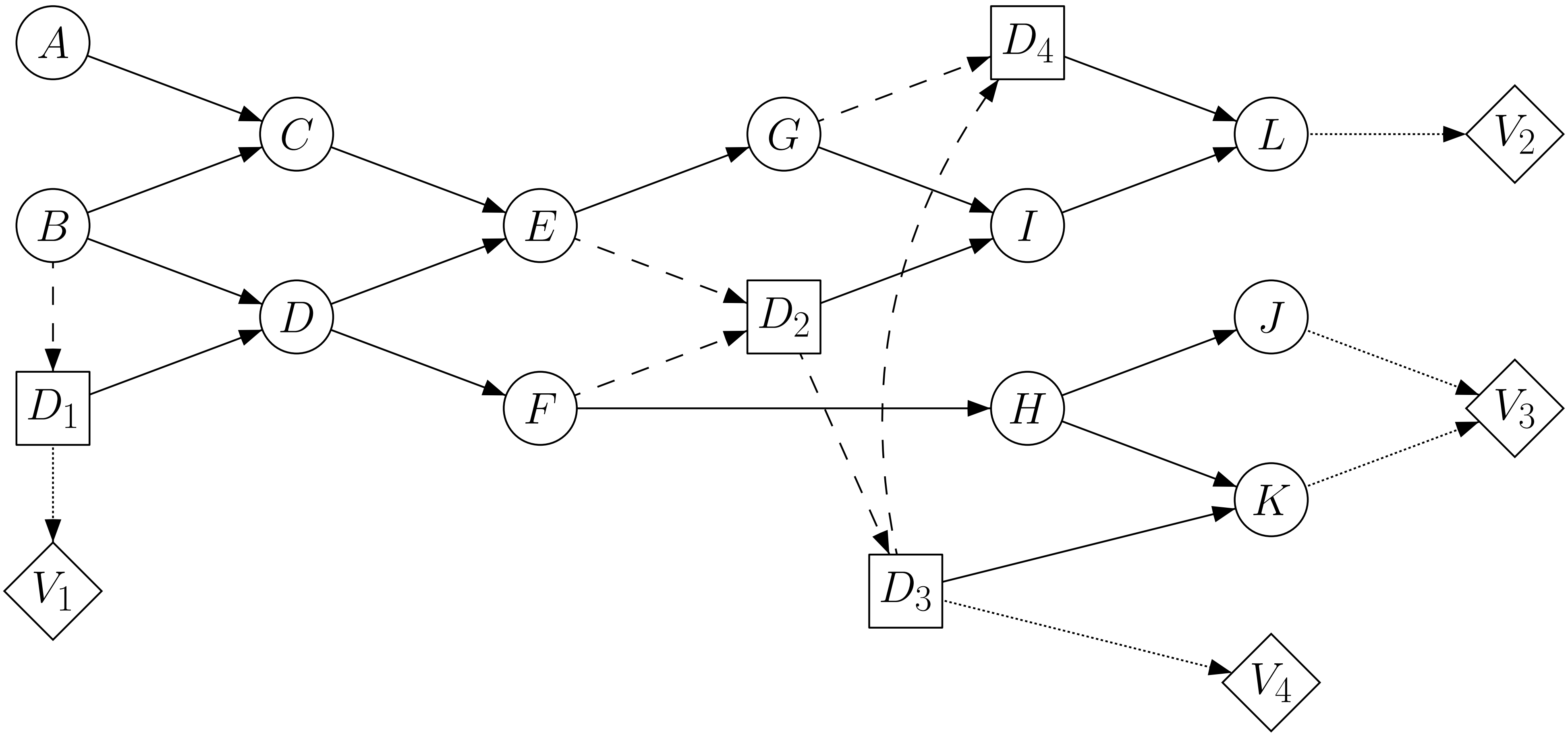
Influence Diagrams (3)

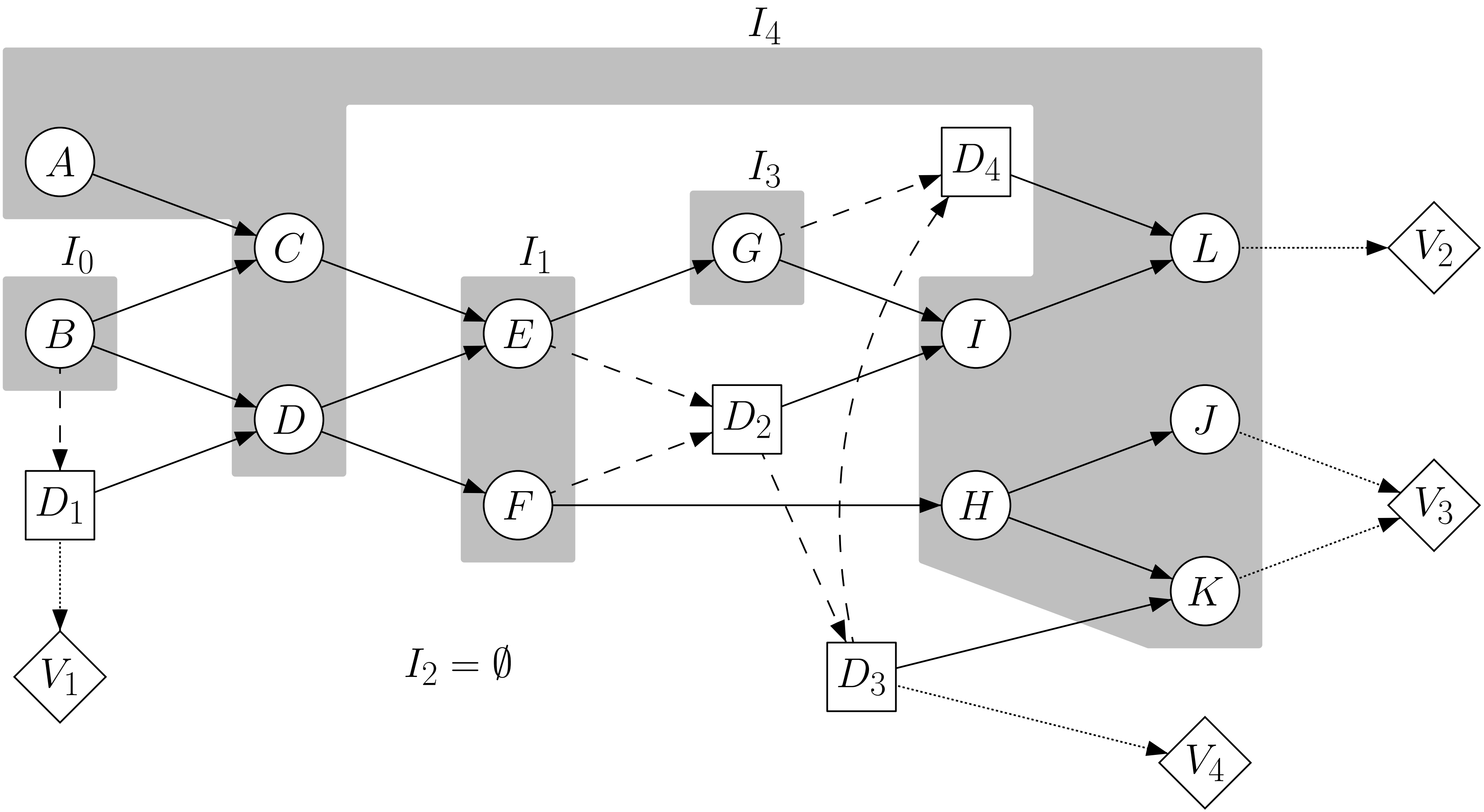


Influence Diagrams (3)



Influence Diagrams (3)

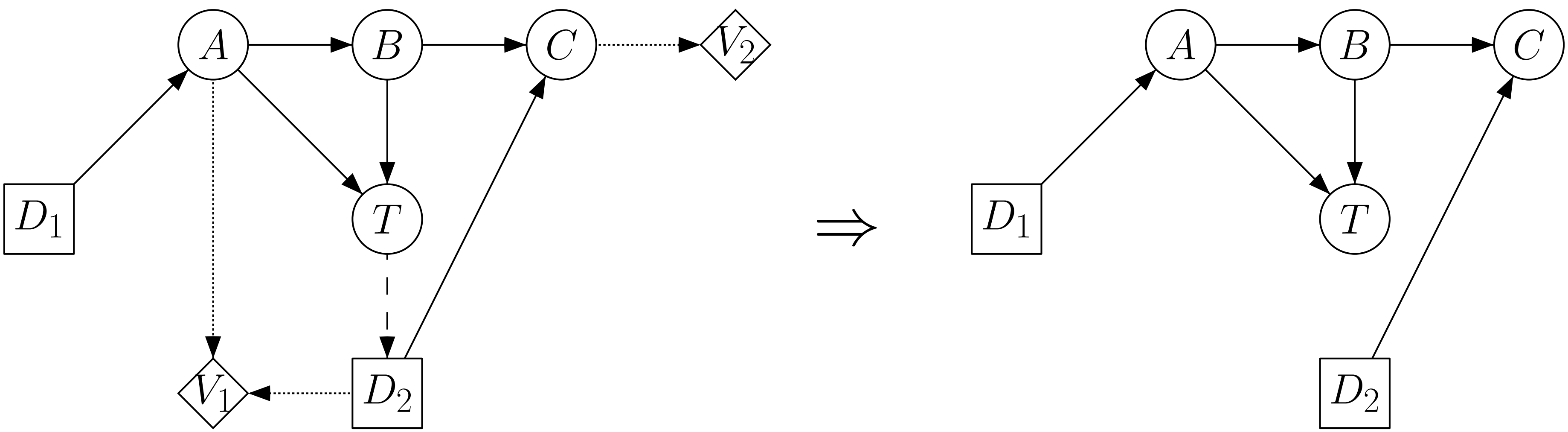



I_4
A
I_3
D_4
I_0
C
I_1
G
L
V_2
B
E
I
D
D_2
J
D_1
F
H
V_3
K
V_1
$I_2 = \emptyset$
D_3
V_4

d-Separation in Influence Diagrams

To be able to use the d-separation, we need to preprocess the graphical structure of an influence diagram as follows:

- remove all utility nodes (and the edges towards them)
- remove edges that point to decision nodes



For example: $C \perp\!\!\!\perp T \mid B$ or $\{A, T\} \perp\!\!\!\perp D_2 \mid \emptyset$.

Chain Rule

The semantics of an influence diagram disallow some probabilities:

- $P(D)$ for a decision node D has no meaning
- $P(A \mid D)$ has no meaning unless a decision $d \in \mathrm{dom}(D)$ has been chosen

Given an influence diagram G with U_C being the set of chance nodes and U_D being the set of decision nodes, we can factorize P as follows:

$$P(U_C \mid U_D) = \prod_{X \in U_C} P(X \mid \mathrm{parents}(X))$$

Solutions to Influence Diagrams

- Given: an influence diagram
- Desired: a strategy which decision(s) to make

Policy

A *policy* for decision D_i is a mapping σ_i, which for any configuration of the past of D_i yields a decision for D_i, i. e.

$$\sigma_i(I_0, D_1, I_1, \ldots, D_{i-1}, I_{i-1}) \in \mathrm{dom}(D_i)$$

Strategy

A *strategy* for an influence diagram is a set of policies, one for each decision node.

Solution

A *solution* to an influence diagram is a strategy maximizing the expected utility.

Solutions to Influence Diagrams (2)

Assume, we are given an influence diagram G over $U = U_C \cup U_D$ and U_V.

- U_C ... set of chance nodes
- U_D ... set of decision nodes and
- $U_V = \{V_i\}$ ... set of utility nodes

Further, we know the following temporal order:

$$I_0 \prec D_1 \prec I_1 \prec \cdots \prec D_n \prec I_n$$

The total utility V be defined as the sum of all utility nodes: $V = \sum_i V_i$

Solutions to Influence Diagrams (3)

- An optimal policy for D_i is

$$\sigma_i(I_0, D_1, \ldots, I_{i-1}) = \arg\max_{d_i} \sum_{I_i} \max_{d_{i+1}} \cdots \max_{d_n} \sum_{I_n} P(U_C \mid U_D) \cdot V$$

  where $d_x \in \mathrm{dom}(D_x)$.

- The expected utility from following policy σ_i (and acting optimally in the future) is

$$\rho_i(I_0, D_1, \ldots, I_{i-1}) = \frac{\max_{d_i} \sum_{I_i} \max_{d_{i+1}} \cdots \max_{d_n} \sum_{I_n} P(U_C \mid U_D) \cdot V}{P(I_0, \ldots, I_{i-1} \mid D_1, \ldots, D_{i-1})}$$

  where $d_x \in \mathrm{dom}(D_x)$.

Solutions to Influence Diagrams (4)

- An optimal strategy yields the maximum expected utility of

$$\mathrm{MEU}(G) = \sum_{I_0} \max_{d_1} \sum_{I_1} \max_{d_2} \cdots \max_{d_n} \sum_{I_n} P(U_C \mid U_D) \cdot V$$

- $\sum\limits_{I_i}$ means (sum-)marginalizing over all nodes in I_i

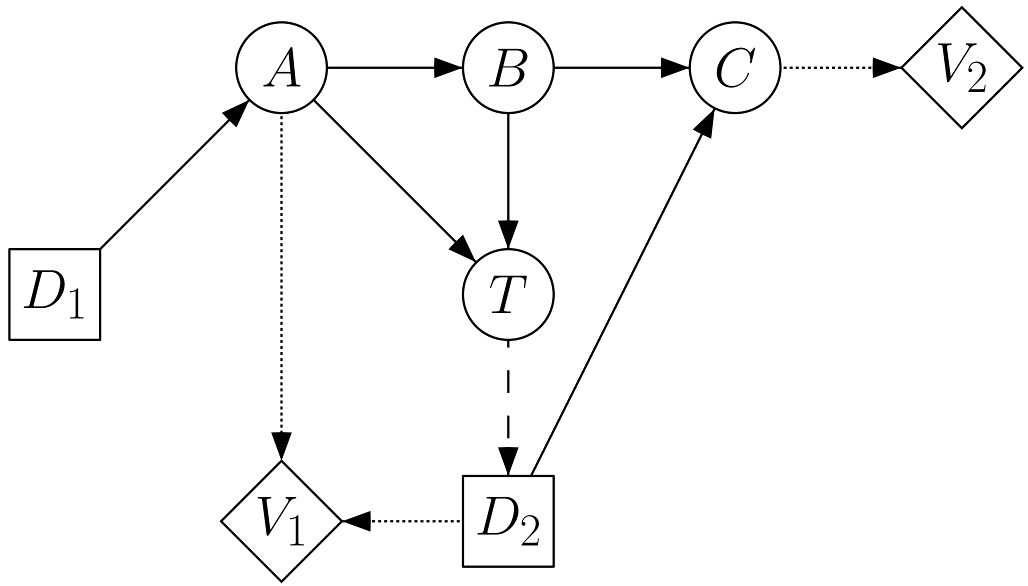
- $\max\limits_{d_i}$ means taking the maximum over all $d_i \in \mathrm{dom}(D_i)$ and thus (max-)marginalizing over D_i

- Everytime I_i is marginalized out, the result is used to determine a policy for D_i.

- Marginalization in reverse temporal order

- $\Rightarrow$ use simplification techniques from the Bayesian network realm to simplify the joint probability distribution $P(U_C \mid U_D)$

Example



$P(A \mid D_1)$	$d_1^{(1)}$	$d_1^{(2)}$
y	0.2	0.8
n	0.8	0.2

$P(B \mid A)$	y	n
y	0.8	0.2
n	0.2	0.8

$P(T \mid A, B)$	y, y	y, n	n, y	n, n
y	0.9	0.5	0.5	0.1
n	0.1	0.5	0.5	0.9

$V_2(C)$	
y	10
n	0

$V_1(A, D_2)$	$d_2^{(1)}$	$d_2^{(2)}$
y	3	0
n	0	2

Utility functions

$P(C \mid B, D_2)$	y, $d_2^{(1)}$	y, $d_2^{(2)}$	n, $d_2^{(1)}$	n, $d_2^{(2)}$
y	0.9	0.5	0.5	0.9
n	0.1	0.5	0.5	0.1

Chance potentials

Example (2)

For D_2 we can read from the graph:

$$I_0 = \emptyset \qquad I_1 = \{T\} \qquad I_2 = \{A, B, C\}$$

Thus, σ_2 can be solved to the following strategy:

$\sigma_2(\emptyset, D_1, \{T\})$	$d_1^{(1)}$	$d_1^{(2)}$
y	$d_2^{(1)}$	$d_2^{(1)}$
n	$d_2^{(2)}$	$d_2^{(2)}$

$\rho_2(\emptyset, D_1, \{T\})$	$d_1^{(1)}$	$d_1^{(2)}$
y	9.51	11.29
n	10.34	8.97

Finally, $\sigma_1 = d_1^{(2)}$ and $\mathrm{MEU}(G) = 10.58$.

Frameworks of Imprecision and Uncertainty

Problems with Probability Theory

Representation of Ignorance (dt. Unwissen)

- We are given a die with faces $1, \ldots, 6$
  What is the certainty of showing up face i ?
  - Conduct a statistical survey (roll the die 10000 times) and estimate the relative frequency: $P(\{i\}) = \frac{1}{6}$
  - Use subjective probabilities (which is often the normal case): We do not know anything (especially and explicitly we do not have any reason to assign unequal probabilities), so the most plausible distribution is a uniform one.

  $\Rightarrow$ Problem: Uniform distribution because of ignorance or extensive statistical tests

- Experts analyze aircraft shapes: 3 aircraft types A, B, C
  "It is type A or B with 90% certainty. About C, I don't have any clue and I do not want to commit myself. No preferences for A or B."

$\Rightarrow$ Problem: Propositions hard to handle with Bayesian theory

Modeling Imprecise Data

"$A \subseteq X$ being an imprecise date" means: the true value x_0 lies in A but there are no preferences on A.

Ω	set of possible elementary events
$\Theta = \{\xi\}$	set of observers
$\lambda(\xi)$	importance of observer ξ

Some elementary event from Ω occurs and every observer $\xi \in O$ shall announce which elementary events she personally considers possible. This set is denoted by $\Gamma(\xi) \subseteq \Omega$. $\Gamma(\xi)$ is then an imprecise date.

$\lambda : 2^{\Theta} \rightarrow [0, 1]$	probability measure (interpreted as importance measure)
$(\Theta, 2^{\Theta}, \lambda)$	probability space
$\Gamma : \Theta \rightarrow 2^{\Omega}$	set-valued mapping

Imprecise Data (2)

Let $A \subseteq \Omega$:

a) $\Gamma^*(A) \stackrel{\text{Def}}{=} \{\xi \in \Theta \mid \Gamma(\xi) \cap A \neq \emptyset\}$

b) $\Gamma_*(A) \stackrel{\text{Def}}{=} \{\xi \in \Theta \mid \Gamma(\xi) \neq \emptyset \text{ and } \Gamma(\xi) \subseteq A\}$

Remarks:

a) If $\xi \in \Gamma^*(A)$, then it is *plausible* for ξ that the occurred elementary event lies in A.

b) If $\xi \in \Gamma_*(A)$, then it is *certain* for ξ that the event lies in A.

c) $\{\xi \mid \Gamma(\xi) \neq \emptyset\} \;=\; \Gamma^*(\Omega) \;=\; \Gamma_*(\Omega)$

Let $\lambda(\Gamma^*(\Omega)) > 0$. Then we call

$$P^*(A) = \frac{\lambda(\Gamma^*(A))}{\lambda(\Gamma^*(\Omega))} \quad \text{the upper, and} \quad P_*(A) = \frac{\lambda(\Gamma_*(A))}{\lambda(\Gamma_*(\Omega))} \quad \text{the lower}$$

probability w. r. t. λ and Γ.

Example

$$\Theta = \{a, b, c, d\} \qquad \lambda\colon\ a \mapsto {}^1\!/_6 \qquad \Gamma\colon\ a \mapsto \{1\}$$
$$\Omega = \{1, 2, 3\} \qquad b \mapsto {}^1\!/_6 \qquad b \mapsto \{2\}$$
$$\Gamma^*(\Omega) = \{a, b, d\} \qquad c \mapsto {}^2\!/_6 \qquad c \mapsto \emptyset$$
$$\lambda(\Gamma^*(\Omega)) = {}^4\!/_6 \qquad d \mapsto {}^2\!/_6 \qquad d \mapsto \{2, 3\}$$

A	$\Gamma^*(A)$	$\Gamma_*(A)$	$P^*(A)$	$P_*(A)$
$\emptyset$	$\emptyset$	$\emptyset$	0	0
$\{1\}$	$\{a\}$	$\{a\}$	$\frac{1}{4}$	$\frac{1}{4}$
$\{2\}$	$\{b, d\}$	$\{b\}$	$\frac{3}{4}$	$\frac{1}{4}$
$\{3\}$	$\{d\}$	$\emptyset$	$\frac{1}{2}$	0
$\{1, 2\}$	$\{a, b, d\}$	$\{a, b\}$	1	$\frac{1}{2}$
$\{1, 3\}$	$\{a, d\}$	$\{a\}$	$\frac{3}{4}$	$\frac{1}{4}$
$\{2, 3\}$	$\{b, d\}$	$\{b, d\}$	$\frac{3}{4}$	$\frac{3}{4}$
$\{1, 2, 3\}$	$\{a, b, d\}$	$\{a, b, d\}$	1	1

One can consider $P^*(A)$ and $P_*(A)$ as upper and lower probability bounds.

Imprecise Data (3)

Some properties of probability bounds:

a) $P^*\colon\ 2^\Omega \to [0,1]$

b) $0 \leq P_* \leq P^* \leq 1, \qquad P_*(\emptyset) = P^*(\emptyset) = 0, \qquad P_*(\Omega) = P^*(\Omega) = 1$

c) $A \subseteq B \quad \Rightarrow \quad P^*(A) \leq P^*(B) \quad \text{and} \quad P_*(A) \leq P_*(B)$

d) $A \cap B = \emptyset \quad \not\Rightarrow \quad P^*(A) + P^*(B) = P^*(A \cup B)$

e) $P_*(A \cup B) \geq P_*(A) + P_*(B) - P_*(A \cap B)$

f) $P^*(A \cup B) \leq P^*(A) + P^*(B) - P^*(A \cap B)$

g) $P_*(A) = 1 - P^*(\Omega \backslash A)$

One can prove the following generalized equation:

$$P_*(\bigcup_{i=1}^{n} A_i) \quad \geq \sum_{\emptyset \neq I : I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \cdot P_*(\bigcap_{i \in I} A_i)$$

These set functions also play an important role in theoretical physics (capacities, Choquet, 1955). Shafer did generalize these thoughts and developed a theory of belief functions.

Belief Revision

How is new knowledge incoporated?

Every observer announces the location of the ship in form of a subset of all possible ship locations. Given these set-valued mappings, we can derive upper and lower probabilities with the help of the observer importance measure. Let us assume the ship is certainly at sea.

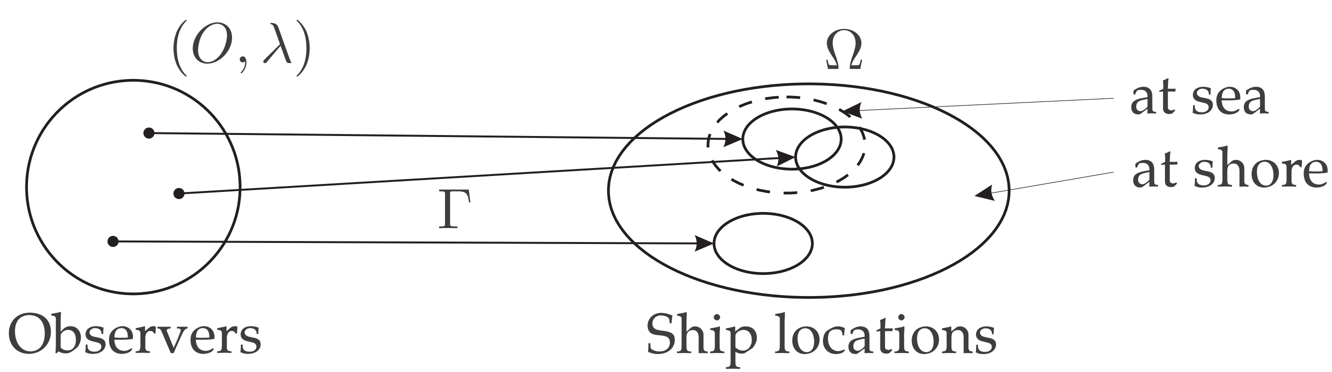
How do the upper/lower probabilities change?

Example

a) *Geometric Conditioning*
(observers that give partial or full wrong information are discarded)

$$P_*(A \mid B) = \frac{\lambda(\{\xi \in \Theta \mid \Gamma(\xi) \subseteq A \text{ and } \Gamma(\xi) \subseteq B\})}{\lambda(\{\xi \in \Theta \mid \Gamma(\xi) \subseteq B\})} = \frac{P_*(A \cap B)}{P_*(B)}$$

$$P^*(A \mid B) = \frac{\lambda(\{\xi \in \Theta \mid \Gamma(\xi) \subseteq B \text{ and } \Gamma(\xi) \cap A \neq \emptyset\})}{\lambda(\{\xi \in \Theta \mid \Gamma(\xi) \subseteq B\})} = \frac{P^*(A \cup \overline{B}) - P^*(\overline{B})}{1 - P^*(\overline{B})}$$

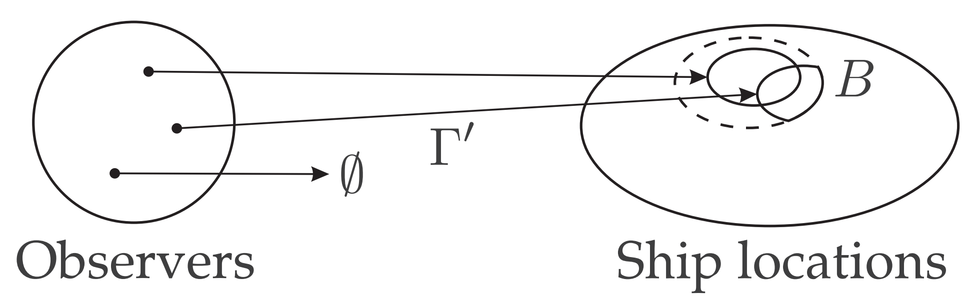


Belief Revision (2)

b) *Data Revision*
(the observed data is modified such that they fit the certain information)

$$(P_*)_B(A) = \frac{P_*(A \cup \overline{B}) - P_*(\overline{B})}{1 - P_*(B)}$$

$$(P^*)_B(A) = \frac{P^*(A \cap B)}{P^*(B)}$$



These two concepts have different semantics. There are several more belief revision concepts.

Imprecise Probabilities

Let x_0 be the true value but assume there is no information about $P(A)$ to decide whether $x_0 \in A$. There are only probability boundaries.

Let $\mathcal{L}$ be a set of probability measures. Then we call

$$(P_\mathcal{L})_* : 2^\Omega \to [0,1]\,, A \mapsto \inf\{P(A) \mid P \in \mathcal{L}\} \qquad \text{the lower and}$$
$$(P_\mathcal{L})^* : 2^\Omega \to [0,1]\,, A \mapsto \sup\{P(A) \mid P \in \mathcal{L}\} \qquad \text{the upper}$$

probability of A w. r. t. $\mathcal{L}$.

a) $(P_\mathcal{L})_*(\emptyset) = (P_\mathcal{L})^*(\emptyset) = 0; \quad (P_\mathcal{L})_*(\Omega) = (P_\mathcal{L})^*(\Omega) = 1$

b) $0 \le (P_\mathcal{L})_*(A) \le (P_\mathcal{L})^*(A) \le 1$

c) $(P_\mathcal{L})^*(A) = 1 - (P_\mathcal{L})_*(\overline{A})$

d) $(P_\mathcal{L})_*(A) + (P_\mathcal{L})_*(B) \le (P_\mathcal{L})_*(A \cup B)$

e) $(P_\mathcal{L})_*(A \cap B) + (P_\mathcal{L})_*(A \cup B) \not\geq (P_\mathcal{L})_*(A) + (P_\mathcal{L})_*(B)$

Belief Revision

Let $B \subseteq \Omega$ and $\mathcal{L}$ a class of probabilities. The we call

$$A \subseteq \Omega : (P_{\mathcal{L}})_*(A \mid B) = \inf\{P(A \mid B) \mid P \in \mathcal{L} \wedge P(B) > 0\} \qquad \text{the lower and}$$
$$A \subseteq \Omega : (P_{\mathcal{L}})^*(A \mid B) = \sup\{P(A \mid B) \mid P \in \mathcal{L} \wedge P(B) > 0\} \qquad \text{the upper}$$

conditional probability of A given B.

A class $\mathcal{L}$ of probability measures on $\Omega = \{\omega_1, \ldots, \omega_n\}$ is of type 1, iff there exist functions R_1 and R_2 from 2^Ω into $[0,1]$ with:

$$\mathcal{L} = \{P \mid \forall A \subseteq \Omega : R_1(A) \leq P(A) \leq R_2(A)\}$$

Belief Revision (2)

Intuition: P is determined by $P(\{\omega_i\})$, $i = 1, \ldots, n$ which corresponds to a point in $\mathbb{R}^n$ with coordinates $\big(P(\{\omega_1\}), \ldots, P(\{\omega_n\})\big)$.

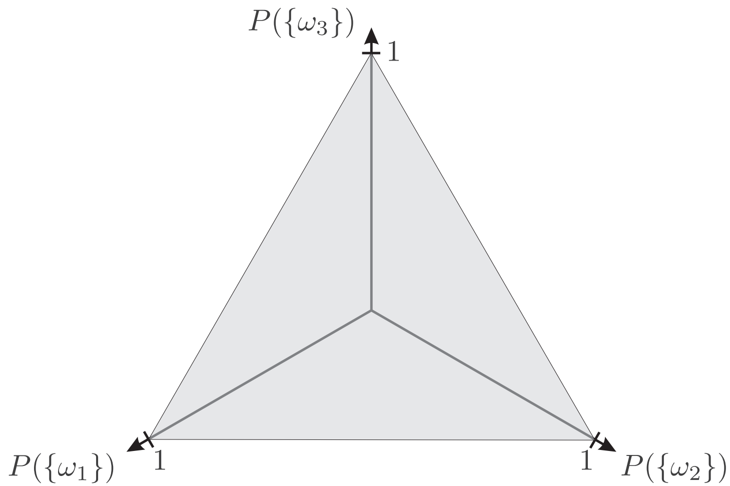
If $\mathcal{L}$ is type 1, it holds true that:

$$\begin{aligned}\mathcal{L} \Leftrightarrow \Big\{(r_1, \ldots, r_n) \in \mathbb{R}^n \mid \exists P\colon \forall A \subseteq \Omega\colon \\ (P_{\mathcal{L}})_*(A) \leq P(A) \leq (P_{\mathcal{L}})^*(A) \\ \text{and } \; r_i = P(\{\omega_i\}),\; i = 1, \ldots, n\Big\}\end{aligned}$$

Example

$$\Omega = \{\omega_1, \omega_2, \omega_3\}$$

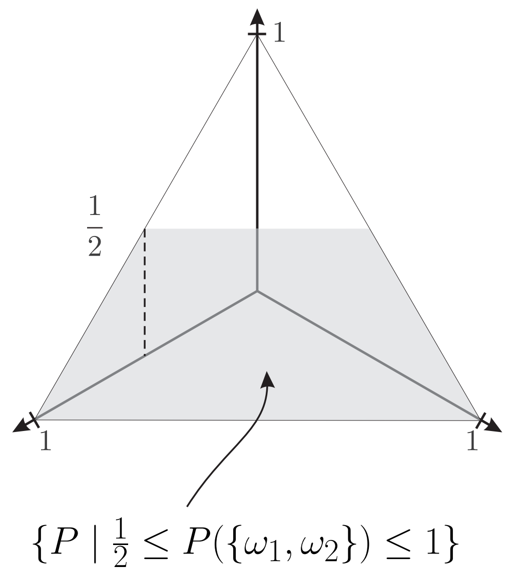
$$\mathcal{L} = \{P \mid \tfrac{1}{2} \leq P(\{\omega_1, \omega_2\}) \leq 1, \quad \tfrac{1}{2} \leq P(\{\omega_2, \omega_3\}) \leq 1, \quad \tfrac{1}{2} \leq P(\{\omega_1, \omega_3\}) \leq 1\}$$

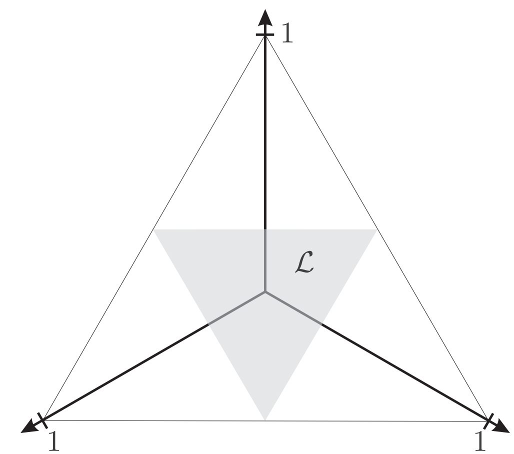


general restriction:

$0 \leq P(\{\omega_i\}) \leq 1$

$P(\{\omega_1\}) + P(\{\omega_2\}) + P(\{\omega_3\}) = 1$





Let $A_1 = \{\omega_1, \omega_2\}$, $A_2 = \{\omega_2, \omega_3\}$, $A_3 = \{\omega_1, \omega_3\}$

$$\begin{aligned} &P_*(A_1) + P_*(A_2) + P_*(A_3) - P_*(A_1 \cap A_2) - P_*(A_2 \cap A_3) - P_*(A_1 \cap A_3) + P_*(A_1 \cap A_2 \cap A_3) \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} - 0 - 0 - 0 + 0 = \frac{3}{2} > 1 = P(A_1 \cup A_2 \cup A_3) \end{aligned}$$

Belief Revision (3)

If $\mathcal{L}$ is type 1 and $(P_\mathcal{L})^*(A \cup B) \geq (P_\mathcal{L})^*(A) + (P_\mathcal{L})^*(B) - (P_\mathcal{L})^*(A \cap B)$, then

$$(P_\mathcal{L})^*(A \mid B) = \frac{(P_\mathcal{L})^*(A \cap B)}{(P_\mathcal{L})^*(A \cap B) + (P_\mathcal{L})_*(B \cap \overline{A})}$$

and

$$(P_\mathcal{L})_*(A \mid B) = \frac{(P_\mathcal{L})_*(A \cap B)}{(P_\mathcal{L})_*(A \cap B) + (P_\mathcal{L})^*(B \cap \overline{A})}$$

Let $\mathcal{L}$ be a class of type 1. $\mathcal{L}$ is of type 2, iff

$$(P_\mathcal{L})_*(A_1 \cup \cdots \cup A_n) \geq \sum_{I:\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \cdot (P_\mathcal{L})_*(\bigcap_{i \in I} A_i)$$

Motivation

(Θ, Q)	Sensors	
Ω	possible results, $\Gamma : \Theta \to 2^\Omega$	
Γ, Q	induce a probability m on 2^Ω	
$m :$	$A \mapsto Q(\{\theta \in \Theta \mid \Gamma(\theta) = A\})$	mass distribution
$\mathrm{Bel} :$	$A \mapsto \sum_{B:B\subseteq A} m(B)$	Belief (lower probability)
$\mathrm{Pl} :$	$A \mapsto \sum_{B:B\cap A\neq\emptyset} m(B)$	Plausibility (upper probability)

- Random sets: Dempster (1968)
- Belief functions: Shafer (1974)
  Development of a completely new uncertainty calculus

Belief Functions (2)

The function $\text{Bel} : 2^\Omega \to [0, 1]$ is called *belief function*, if it possesses the following properties:

- $\text{Bel}(\emptyset) = 0$
- $\text{Bel}(\Omega) = 1$
- $\forall n \in \mathbb{N}\colon\ \forall A_1, \ldots, A_n \in 2^\Omega :$
  $\text{Bel}(A_1 \cup \cdots \cup A_n) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} \cdot \text{Bel}(\bigcap_{i \in I} A_i)$

If Bel is a belief function then for $m : 2^\Omega \to \mathbb{R}$ with $m(A) = \sum_{B : B \subseteq A} (-1)^{|A \setminus B|} \cdot \text{Bel}(B)$ the following properties hold:

- $0 \leq m(A) \leq 1$
- $m(\emptyset) = 0$
- $\sum_{A \subseteq \Omega} m(A) = 1$

Belief Functions (3)

Let $|\Omega| < \infty$ and $f, g : 2^\Omega \to [0,1]$.

$$\forall A \subseteq \Omega\colon \left(f(A) = \sum_{B: B \subseteq A} g(B)\right)$$

$$\Leftrightarrow$$

$$\forall A \subseteq \Omega\colon \left(g(A) = \sum_{B: B \subseteq A} (-1)^{|B|} \cdot f(B)\right)$$

(g is called the *Möbius transformed* of f)

The mapping $m : 2^\Omega \to [0,1]$ is called a *mass distribution*, if the following properties hold:

- $m(\emptyset) = 0$
- $\sum_{A \subseteq \Omega} m(A) = 1$

Example

A	$\emptyset$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,3\}$	$\{1,2,3\}$
$m(A)$	0	$^1/_4$	$^1/_4$	0	0	0	$^2/_4$	0
$\text{Bel}(A)$	0	$^1/_4$	$^1/_4$	0	$^2/_4$	$^1/_4$	$^3/_4$	1

Belief $\widehat{=}$ lower probability with modified semantic

$$\text{Bel}(\{1,3\}) = m(\emptyset) + m(\{1\}) + m(\{3\}) + m(\{1,3\})$$
$$m(\{1,3\}) = \text{Bel}(\{1,3\}) - \text{Bel}(\{1\}) - \text{Bel}(\{3\})$$

$m(A)$	measure of the trust/belief that exactly A occurs
$\text{Bel}_m(A)$	measure of total belief that A occurs
$\text{Pl}_m(A)$	measure of not being able to disprove A (plausibility)

$$\text{Pl}_m(A) = \sum_{B: A\cap B\neq\emptyset} m(B) = 1 - \text{Bel}(\overline{A})$$

Given one of m, Bel or Pl, the other two can be efficiently computed.

Knowledge Representation

$m(\Omega) = 1$, $m(A) = 0$ else	total ignorance
$m(\{\omega_0\}) = 1$, $m(A) = 0$ else	value (ω_0) known
$m(\{\omega_i\}) = p_i, \sum_{i=1}^{n} p_i = 1$	Bayesian analysis

Further intermediate steps can be modeled.

Belief Revision

- Data Revision:
  - Mass of A flows onto $A \cap B$.
  - Masses are normalized to 1 ($\emptyset$-mass is destroyed)
- Geometric Conditioning:
  - Masses that do not lie completely inside B, flow off
  - Normalize

There is a mass flow from t to s (written: $s \sqsubseteq t$) iff for every $A \subseteq \Omega$ there exist functions $h_A : 2^\Omega \longrightarrow [0, 1]$ such that the following properties hold:

- $\sum_{B:B\subseteq\Omega} h_A(B) = t(A)$ for all A
- $h(A(B) \neq 0 \;\Rightarrow\; B \subseteq A$ for all A, B
- $s(B) = \frac{\sum_{A:A\subseteq\Omega} h_A(B)}{1-\sum_{A:A\subseteq\Omega} h_A(\emptyset)}$

Example

A	$s(A)$	$t(A)$	$u(A)$
$\emptyset$	0	0	0
$\{1\}$	0	0	0.1
$\{2\}$	0.4	0.4	0
$\{3\}$	0.1	0	0
$\{1,2\}$	0.2	0.5	0.1
$\{1,3\}$	0	0	0.4
$\{2,3\}$	0.3	0.1	0.4
Ω	0	0	0

The following relations hold:
$s \sqsubseteq t, t \sqsubseteq s, s \sqsubseteq u, t \sqsubseteq u, t \sqsubseteq t, u \not\sqsubseteq s$

Combination of Random Sets

Let $(\Omega, 2^\Omega)$ be a space of events. Further be $(O_1, 2^{O_1}, \lambda_1)$ and $(O_2, 2^{O_2}, \lambda_2)$ spaces of independent observers.

We call $(O_1 \times O_2, \lambda_1 \cdot \lambda_2)$ the product space of observers and

$$\Gamma : O_1 \times O_2 \to 2^\Omega, \Gamma(x_1, x_2) = \Gamma_1(x_1) \cap \Gamma_2(x_2)$$

the combined observer function.

We obtain with

$$(P_L)_*(A) = \frac{(\lambda_1 \cdot \lambda_2)(\{(x_1, x_2) \mid \Gamma(x_1, x_2) \neq \emptyset \wedge \Gamma(x_1, x_2) \sqsubseteq A\})}{(\lambda_1 \cdot \lambda_2)(\{(x_1, x_2 \mid \Gamma(x_1, x_2) \neq \emptyset)\})}$$

the lower probability of A that respects both observations.

Example

$$\Omega = \{1,2,3\} \qquad \begin{aligned} \lambda_1\colon\ \{a\} &\mapsto {}^1\!/_3 \\ \{b\} &\mapsto {}^2\!/_3 \end{aligned} \qquad \begin{aligned} \lambda_2\colon\ \{c\} &\mapsto {}^1\!/_2 \\ \lambda_2\colon\ \{d\} &\mapsto {}^1\!/_2 \end{aligned}$$

$$\begin{aligned} O_1 &= \{a,b\} \\ O_2 &= \{c,d\} \end{aligned} \qquad \Gamma_1\colon \begin{array}{ccc} a & \mapsto & \{1,2\} \\ b & \mapsto & \{2,3\} \end{array} \qquad \Gamma_2\colon \begin{array}{ccc} c & \mapsto & \{1\} \\ d & \mapsto & \{2,3\} \end{array}$$

Combination:

$$O_1 \times O_2 \;=\; \{\overline{ac}, \overline{bc}, \overline{ad}, \overline{bd}\}$$

$$\begin{aligned} \lambda\colon\ \{\overline{ac}\} &\mapsto {}^1\!/_6 \\ \{\overline{ad}\} &\mapsto {}^1\!/_6 \\ \{\overline{bc}\} &\mapsto {}^2\!/_6 \\ \{\overline{bd}\} &\mapsto {}^2\!/_6 \end{aligned} \qquad \begin{aligned} \Gamma\colon\ \overline{ac} &\mapsto \{1\} \\ \overline{ad} &\mapsto \{2\} \\ \overline{bc} &\mapsto \emptyset \\ \overline{bd} &\mapsto \{2,3\} \end{aligned}$$

$$\begin{aligned} \Gamma_*(\Omega) &= \{(x_1,x_2) \mid \Gamma(x_1,x_2) \neq \emptyset\} \\ &= \{\overline{ac}, \overline{ad}, \overline{bd}\} \end{aligned}$$

$$\lambda(\Gamma_*(\Omega)) = {}^4\!/_6$$

Example (2)

A	$m_1(A)$	$(P_*)_{\Gamma_1}(A)$	$m_2(A)$	$(P_*)_{\Gamma_2}(A)$	$m(A)$	$(P_*)_{\Gamma}(A)$
$\emptyset$	0	0	0	0	0	0
$\{1\}$	0	0	$1/2$	$1/2$	$1/4 = \frac{1/6}{4/6}$	$1/4$
$\{2\}$	0	0	0	0	$1/4$	$1/4$
$\{3\}$	0	0	0	0	0	0
$\{1,2\}$	$1/3$	$1/3$	0	$1/2$	0	$1/2$
$\{1,3\}$	0	0	0	$1/2$	0	$1/4$
$\{2,3\}$	$2/3$	$2/3$	$1/2$	$1/2$	$1/2$	$3/4$
$\{1,2,3\}$	0	1	0	1	0	1

Combinations of Mass Distributions

Motivation: Combination of m_1 and m_2

$m_1(A_i) \cdot m_2(B_j)$:	Mass attached to $A_i \cap B_j$, if only A_i or B_j are concerned
$\sum_{i,j:A_i \cap B_j = A} m_1(A_i) \cdot m_2(B_j)$:	Mass attached to A (after combination)

This consideration only leads to a mass distribution,
if $\sum_{i,j:A_i \cap B_j = \emptyset} m_1(A_i) \cdot m_2(B_j) = 0$.

If this sum is > 0 normalization takes place.

Combination Rule

If m_1 and m_2 are mass distributions over Ω with belief functions Bel_1 and Bel_2 and does further hold $\sum_{i,j:A_i\cap B_j=\emptyset} m_1(A_i)\cdot m_2(B_j) < 1$, then the function $m : 2^\Omega \to [0,1]\,, m(\emptyset) = 0$

$$m(A) = \frac{\sum_{B,C:B\cap C=A} m_1(B)\cdot m_2(C)}{1-\sum_{B,C:B\cap C=\emptyset} m_1(B)\cdot m_2(C)}$$

is a mass distribution. The belief function of m is denoted as $\mathrm{comb}(\mathrm{Bel}_1, \mathrm{Bel}_2)$ or $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$. The above formula is called the combination rule.

Example

$$m_1(\{1,2\}) = {}^1\!/_3 \qquad m_2(\{1\}) = {}^1\!/_2$$
$$m_1(\{2,3\}) = {}^2\!/_3 \qquad m_2(\{2,3\}) = {}^1\!/_2$$

$$\begin{aligned} m = m_1 \oplus m_2 :& \\ \{1\} &\mapsto \frac{{}^1\!/_6}{{}^4\!/_6} = {}^1\!/_4 \\ \{2\} &\mapsto \frac{{}^1\!/_6}{{}^4\!/_6} = {}^1\!/_4 \\ \emptyset &\mapsto 0 \\ \{2,3\} &\mapsto \frac{{}^2\!/_6}{{}^4\!/_6} = {}^1\!/_2 \end{aligned}$$

Combination Rule (2)

Remarks:

a) The result from the combination rule and the analysis of random sets is identical

b) There are more efficient ways of combination

c) $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2 = \mathrm{Bel}_2 \oplus \mathrm{Bel}_1$

d) $\oplus$ is associative

e) $\mathrm{Bel}_1 \oplus \mathrm{Bel}_1 \neq \mathrm{Bel}_1$ (in general)

f) $\mathrm{Bel}_2 : 2^\Omega \to [0,1]\,, m_2(B) = 1$

$$\mathrm{Bel}_2(A) = \begin{cases} 1 & \text{if} B \subseteq A \\ 0 & \text{otherwise} \end{cases}$$

The combination of Bel_1 and Bel_2 yields the data revision of m_1 with B.

Fuzzy Sets

Classical description of concepts/properties:

Example: concept "two-digit number"

a) as a set: $\{10, 11, \ldots, 99\} = M$

b) as predicate $\text{two-digit}(x) = \begin{cases} \textsf{true} & \text{if} \quad 10 \leq x \leq 99 \\ \textsf{false} & \text{else} \end{cases}$

Connection between a) and b):

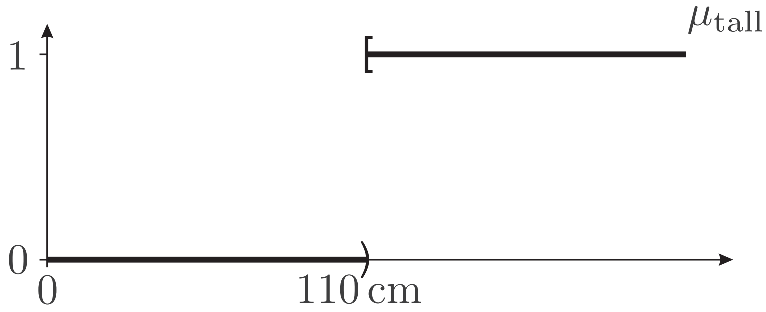
$$M = \{x \in \mathbb{N} \mid \text{two-digit}(x)\}; \qquad \text{two-digit}(x) \Leftrightarrow x \in M$$

Both concepts are not suited for defining concepts like:

- "large"
- "old"
- "heavy"

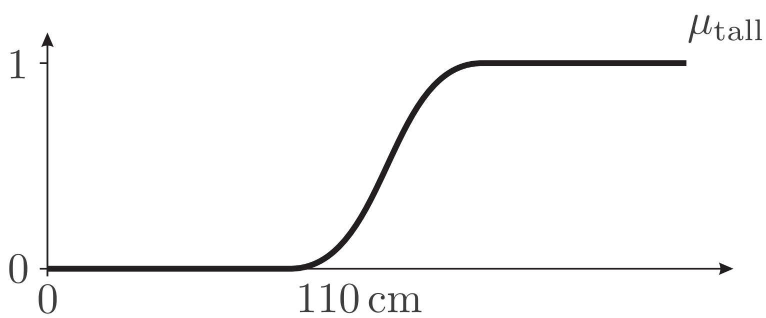
Example

"Set" of sizes (in cm) at which a child would be regarded "tall".



characteristic function of the concept "tall"
($\widehat{=} \{x \mid x \geq 110\,\text{cm}\}$)

The saltus at 110 cm from 0 to 1 is not intuitive. Therefore:



membership degree function

Fuzzy Sets

A *fuzzy set* over a basic set X is a mapping

$$\mu_X : X \to [0,1]$$

$\mu_X(x) \in [0,1]$ is the degree of membership of x to the fuzzy set μ_X.

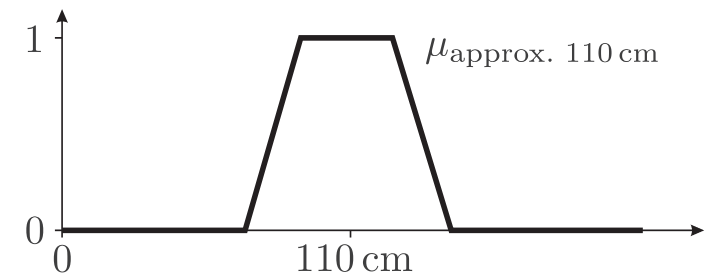
Operations on Fuzzy Sets

Combination of concepts like "tall", "approx. 110 cm", ...

a) The child is "tall" **and** "approx. 110 cm (tall)"

b) The child is "tall" **or** "approx. 110 cm (tall)"

c) The child is **not** "tall"



a)	$\widehat{=}$	Intersection:	classical:	$x \in A \cap B$	$\Leftrightarrow$	$x \in A \wedge x \in B$
b)	$\widehat{=}$	Union:	classical:	$x \in A \cup B$	$\Leftrightarrow$	$x \in A \vee x \in B$
c)	$\widehat{=}$	Complement:	classical:	$x \in \overline{A}$	$\Leftrightarrow$	$\neg(x \in A)$

Postulate:

$$\mu_{\text{tall}\wedge\text{approx. 110 cm}}(x) = \mu_{\text{tall}}(x) \top \mu_{\text{approx. 110 cm}}(x)$$

I. e., we need a mapping $\top : [0,1]^2 \longrightarrow [0,1]$

Generalized Conjunction, t-Norm

A *t-norm* is a mapping $\top : [0,1]^2 \to [0,1]$ with

(T1) $\top(a, 1) = a$

(T2) $a \leq a' \Rightarrow \top(a, b) \leq \top(a', b)$

(T3) $\top(a, b) = \top(b, a)$

(T4) $\top(\top(a, b), c) = \top(a, \top(b, c))$

Examples:

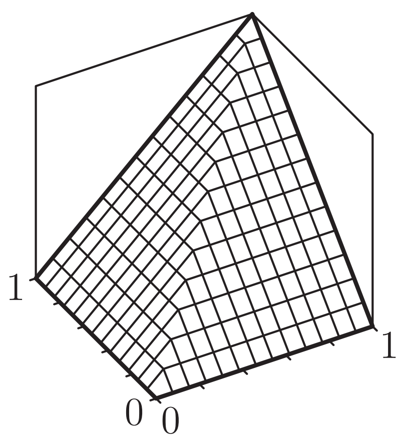
$$\underbrace{\min\{a, b\}}_{\text{largest t-norm, the only idempotent t-norm (i.e., } \top(a,a)=a)},\ a \cdot b,\ \max\{a + b - 1, 0\}$$

$$0 \leq \top(0, 0) \overset{(\mathrm{T2})}{\leq} \top(1, 0) \overset{(\mathrm{T3})}{=} \top(0, 1) \overset{(\mathrm{T1})}{=} 0; \quad \top(1, 1) \overset{(\mathrm{T1})}{=} 1$$

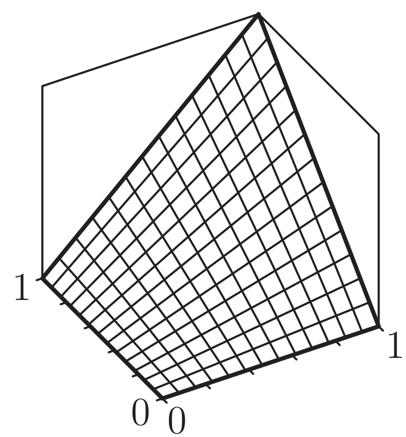
Reasonable claim: $\mu_{\text{tall}}(x) \top \mu_{\text{tall}}(x) = \mu_{\text{tall}}(x) \ \Rightarrow \ \top$ idempotent

t-Norms / Fuzzy Conjunctions

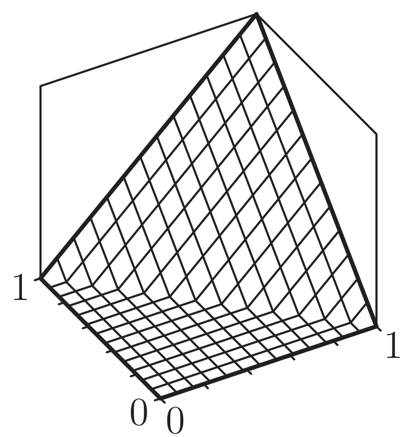
standard conjunction:	$\top_{\min}(a,b)$	$= \min\{a,b\}$
algebraic product:	$\top_{\text{prod}}(a,b)$	$= a \cdot b$
Łukasiewicz:	$\top_{\text{Łuka}}(a,b)$	$= \max\{0, a+b-1\}$
drastic product:	$\top_{-1}(a,b)$	$= \begin{cases} a, & \text{if } b = 1, \\ b, & \text{if } a = 1, \\ 0, & \text{otherwise.} \end{cases}$



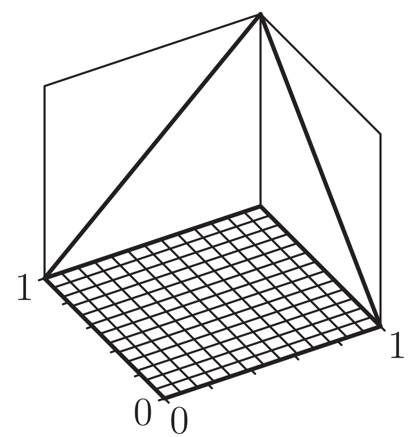
$\top_{\min}$



$\top_{\text{prod}}$



$\top_{\text{Łuka}}$



$\top_{-1}$

Example

$X = \{c_1, c_2, c_3\}$	Set of computers
μ_{cheap}	Fuzzy set of cheap computers
μ_{fast}	Fuzzy set of fast computers
$\mu_{\text{goodvalue}}$	$\mu_{\text{cheap}} \top \mu_{\text{fast}}$

Computer	Price	Speed	μ_{cheap}	μ_{fast}	$\mu_{\text{goodvalue}}$ $(\top = \top_{\text{min}})$	$(\top = \top_{\text{prod}})$
c_1	2000	20	1.0	0.4	0.4	0.40
c_2	2500	40	0.6	0.8	0.6	0.48
c_3	2500	50	0.6	0.9	0.6	0.54

Generalized Disjunction, t-Conorm

A *t-conorm* is a mapping $\bot : [0,1]^2 \to [0,1]$ with

(S1) $\bot(a, 0) = a$

(S2) $a \leq a' \Rightarrow \bot(a, b) \leq \bot(a', b)$

(S3) $\bot(a, b) = \bot(b, a)$

(S4) $\bot(\bot(a, b), c) = \bot(a, \bot(b, c))$

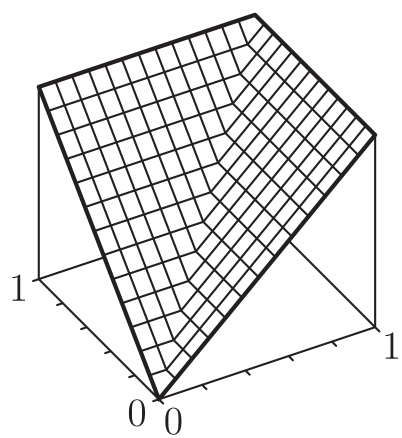
Examples:

$$\max\{a, b\},\ \ a + b - a \cdot b,\ \ \min\{a + b, 1\}$$

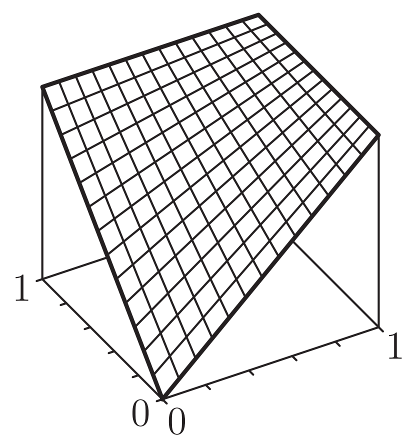
smallest t-conorm, the only idempotent t-conorm (i. e., $\bot(a, a) = a$)

t-Conorms / Fuzzy Disjunctions

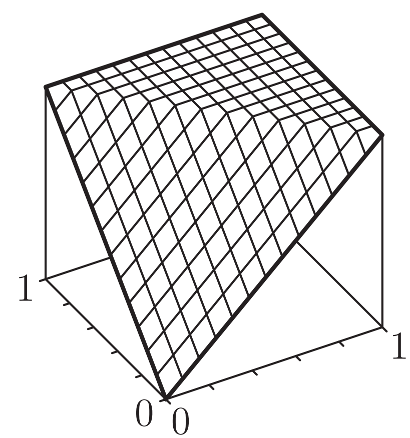
$$\begin{array}{lll}
\text{standard disjunction:} & \bot_{\max}(a,b) & = \max\{a,b\} \\
\text{algebraic sum:} & \bot_{\text{sum}}(a,b) & = a + b - a \cdot b \\
\text{Łukasiewicz:} & \bot_{\text{Łuka}}(a,b) & = \min\{1, a + b\} \\
\text{drastic sum:} & \bot_{-1}(a,b) & = \begin{cases} a, & \text{if } b = 0, \\ b, & \text{if } a = 0, \\ 1, & \text{otherwise.} \end{cases}
\end{array}$$



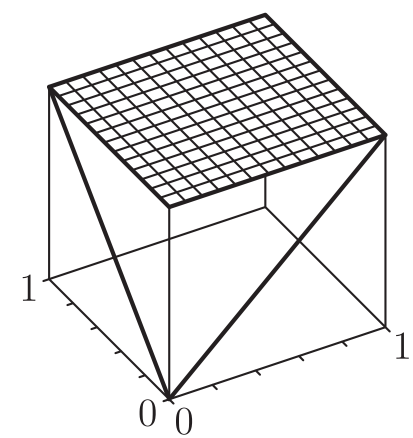
$\bot_{\max}$



$\bot_{\text{sum}}$



$\bot_{\text{Łuka}}$



$\bot_{-1}$

Generalized Negation

A *negation operator* is a mapping $\sim\colon [0,1] \to [0,1]$ with

(N1) $\sim 0 = 1$

(N2) $a \leq b \quad \Rightarrow \quad \sim b \leq \sim a$

(N3) $\sim(\sim a) = a$

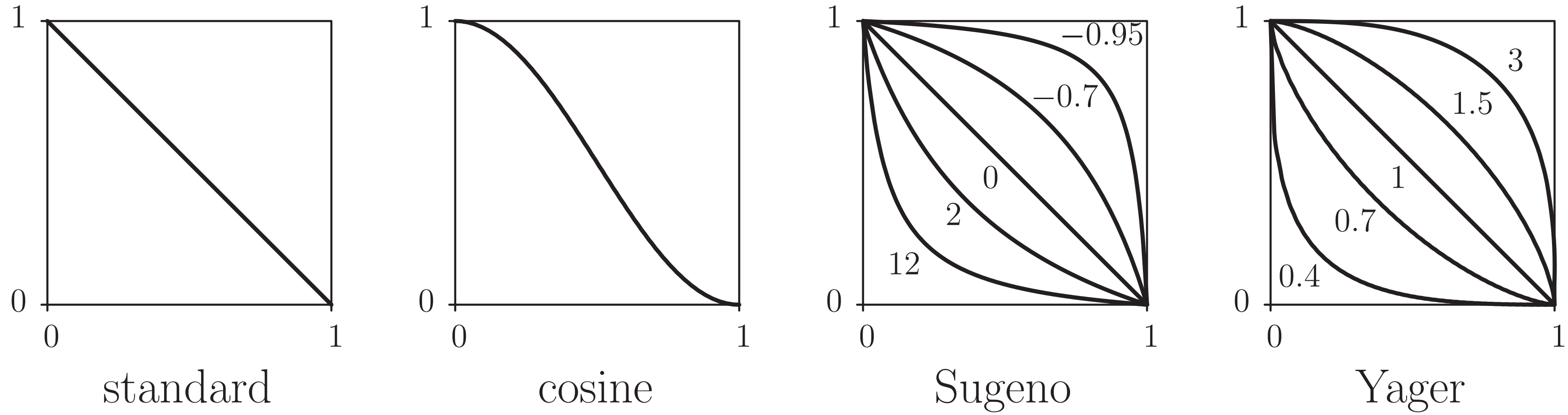
From (N1) and (N3) follows: $\quad \sim 1 = 0$

Relation between t-norms and t-conorms:

$\top$ t-norm $\quad \Leftrightarrow \quad \bot_{\sim}$ t-conorm: $\bot_{\sim}(a,b) = \ \sim(\top(\sim a, \sim b)) \ \left(a \vee b \mathrel{\widehat{=}} \neg(\neg a \wedge \neg b)\right)$

$\bot$ t-conorm $\quad \Leftrightarrow \quad \top_{\sim}$ t-norm: $\top_{\sim}(a,b) = \ \sim(\bot(\sim a, \sim b)) \ \left(a \wedge b \mathrel{\widehat{=}} \neg(\neg a \vee \neg b)\right)$

standard negation:	$\sim a$	$= 1 - a$
threshold negation:	$\sim(a;\theta)$	$= \begin{cases} 1, & \text{if } x \leq \theta, \\ 0, & \text{otherwise.} \end{cases}$
cosine negation:	$\sim a$	$= \frac{1}{2}(1 + \cos \pi a)$
Sugeno negation:	$\sim(a;\lambda)$	$= \dfrac{1 - a}{1 + \lambda a}$
Yager negation:	$\sim(a;\lambda)$	$= (1 - a^{\lambda})^{\frac{1}{\lambda}}$



Reasoning with Uncertainty Module (RUM)

Motivation:

$$\text{modus ponens (mp):} \quad \frac{A \to B, A}{B}, \qquad \text{modus tollens (mt):} \quad \frac{A \to B, \neg B}{\neg A}$$

Generalization of mp and mt on $[0,1]$-valued propositions, e. g.:

$$\mu_{\text{tall}}(x) \xrightarrow{0.8} \mu_{\text{heavy}}(x), \mu_{\text{tall}}(x) \geq 0.9 \quad \Rightarrow \quad \mu_{\text{heavy}} \geq ?$$

Modus Ponens: $[\![\cdot]\!]$ fulfillment degree

- Given: $[\![A \to B]\!] \geq \gamma$; $[\![A]\!] \geq \alpha$
- Desired: $[\![B]\!] \geq \beta = \beta(\gamma, \alpha)$
- $[\![B]\!] \geq [\![A \wedge (A \to B)]\!] = \top([\![A]\!], [\![A \to B]\!]) \geq \top(\alpha, \gamma) = \beta$

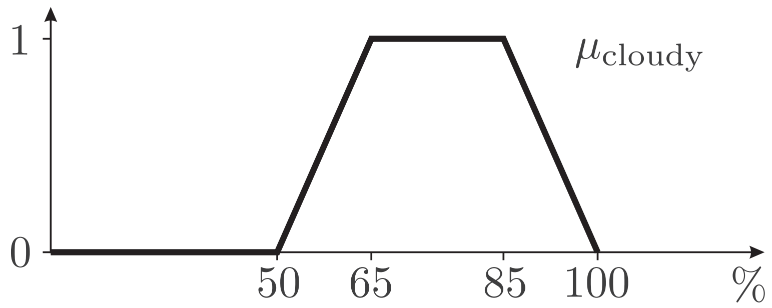
Modus Tollens:

- Given: $[\![B]\!] \leq \beta$,$[\![A \to B]\!] \geq \gamma$
- Desired: $[\![A]\!] \leq \alpha = \alpha(\beta, \gamma)$
- $[\![\neg A]\!] \geq [\![\neg B \wedge (A \to B)]\!] = \top(\sim(/B/), [\![A \to B]\!]) \geq \top(\sim(\beta), \gamma)$

$\Rightarrow$ $[\![A]\!] = [\![\neg\neg A]\!] = \sim([\![\neg A]\!]) \leq \sim(\top(\sim(\beta), \gamma)) = \bot(\beta, \sim(\gamma))$

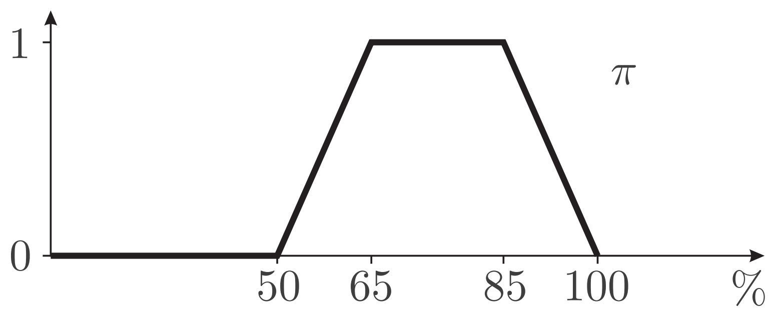
Possibility Theory

a) The vague concept "cloudy" is modeled by the fuzzy set μ_{cloudy}:



Vagueness

b) There exists a true but unknown value x_0. Every x is assigned a degree to which extent $x = x_0$ is considered possible.



Uncertainty

Possibility Theory (2)

$\pi(x)$ is a possibility degree

$\pi(x) = 0$	$x = x_0$ impossible
$\pi(x) = 1$	$x = x_0$ without restriction possible
$\pi(x) \in (0, 1)$	$x = x_0$ gradually possible

A *possibility distribution* π over Ω is a function $\pi : \Omega \rightarrow [0, 1]$ for which the condition

$$\exists \omega \in \Omega : \pi(\omega) = 1$$

holds.

Possibility and Necessity

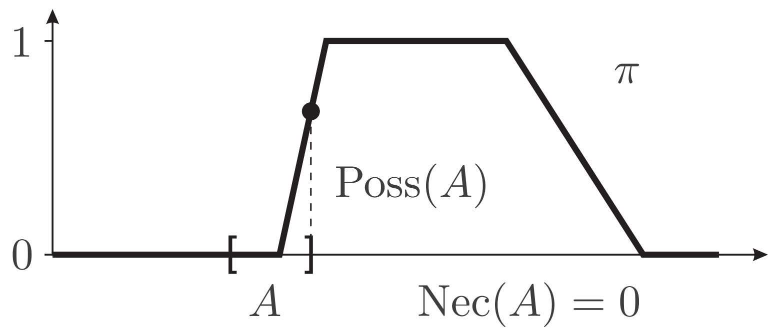
Let π be a possibility distribution over Ω.

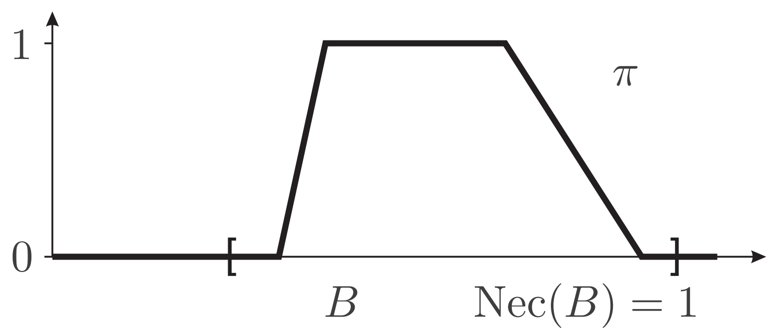
- The *possibility measure* Poss induced by π is defined as

$$\text{Poss} : 2^{\Omega} \to [0,1]\,, \qquad A \mapsto \sup\{\pi(x) \mid x \in A\}$$

- The *necessity measure* Nec induced by π is defined as

$$\text{Nec} : 2^{\Omega} \to [0,1]\,, \qquad A \mapsto 1 - \text{Poss}(\overline{A})$$





Possibility and Necessity (2)

The functions Poss and Nec fulfill the following properties:

$$\begin{array}{lll} \text{Poss}(\emptyset) = 0, & \text{Poss}(\Omega) = 1, & \text{Poss}(A \cup B) = \max\{\text{Poss}(A), \text{Poss}(B)\} \\ \text{Nec}(\emptyset) = 0, & \text{Nec}(\Omega) = 1, & \text{Nec}(A \cap B) = \min\{\text{Nec}(A), \text{Nec}(B)\} \end{array}$$

In general:

$$\begin{aligned} \text{Poss}(A \cap B) &\neq \min\{\text{Poss}(A), \text{Poss}(B)\} \\ \text{Nec}(A \cup B) &\neq \max\{\text{Nec}(A), \text{Nec}(B)\} \quad \text{but} \\ \text{Nec}(A \cup B) &\geq \max\{\text{Nec}(A), \text{Nec}(B)\} \end{aligned}$$

$\text{Nec}(A) = 0$ and $\text{Poss}(A) = 1$ represent complete ignorance.

Possibility and Necessity (3)

A mass distribution

$$m : 2^{\Omega} \to [0, 1]$$

with

$$\sum_{A: A \subseteq \Omega} m(A) = 1, \; m(\emptyset) = 0$$

is called *consonant*, if all sets A with $m(A) > 0$ (the so-called focal elements) form an *inclusion chain*, i. e. there exists for all such sets an enumeration such that:

$$A_1 \subseteq A_2 \subseteq \cdots \subseteq A_m$$

Possibility and Necessity (4)

If m is consonant, then the corresponding belief function

$$\mathrm{Bel}_m : 2^\Omega \to [0,1]\,; \quad A \mapsto \sum_{B:B\subseteq A} m(B)$$

has the properties of a necessity measure:

$$\mathrm{Bel}_m(\emptyset) = 0, \quad \mathrm{Bel}_m(\Omega) = 1, \quad \mathrm{Bel}_m(A \cap B) = \min\{\mathrm{Bel}_m(A), \mathrm{Bel}_m(B)\}$$

If m is consonant, then the corresponding plausibility function

$$\mathrm{Pl}_m : 2^\Omega \to [0,1]\,; \quad A \mapsto \sum_{B:B\cap A\neq\emptyset} m(B)$$

has the properties of a possibility measure.

Homepages

- Otto-von-Guericke-University of Magdeburg
  `http://www.uni-magdeburg.de/`

- School of Computer Science
  `http://www.cs.uni-magdeburg.de/`

- Computational Intelligence Group
  `http://fuzzy.cs.uni-magdeburg.de/`