Artificial Intelligence

9. Probabilistic Reasoning

On Computers that Think about What is Likely to be True

Jörg Hoffmann       Wolfgang Wahlster

SAARLAND UNIVERSITY
COMPUTER SCIENCE

Summer Term 2012
Agenda

1. Motivation
2. Basics of Probability Theory
3. Basic Probabilistic Inference
4. Probabilistic Inference Using Bayes’ Rule
5. Bayesian Networks
6. Inference in Bayesian Networks
7. Closing Remarks
Motivation

Sources of uncertainty in decision-making:

- Incomplete knowledge and partial observability.
- Non-deterministic actions.
- Unreliable sensors.
- Uncertainty about the domain behavior.

→ Drawing conclusions under uncertainty!
Decision-Making Under Uncertainty

**Example:** Giving a lecture.

- **Goal:** Be in HS002 at 12:15 to give a lecture.
- **Possible plans:**
  - $P_1$: Get up at 11:00, leave at 11:45, arrive at 12:00.
  - $P_2$: Get up at 11:45, leave at 12:05, arrive at 12:15.

**Better Example:** Which train to take to Frankfurt airport?
Uncertainty and Logics

**Example:** We want to build an expert dental diagnosis system, that deduces the **cause** (the disease) from the **symptoms**.

Can we encode this kind of reasoning in logics?

Say we have a toothache. How’s about: ("cavity" = "Loch im Zahn")

\[
\forall p [Symptom(p, toothache) \Rightarrow Disease(p, cavity)]
\]

Is this rule correct?

So what about:

\[
\forall p [Symptom(p, toothache) \Rightarrow \\
Disease(p, cavity) \lor Disease(p, gum\_disease) \lor \ldots]
\]
Uncertainty and Logics, ctd.

Perhaps a *causal* rule is better?

$$\forall p[Disease(p, cavity) \Rightarrow Symptom(p, toothache)]$$

→ Is this correct?

→ Does this allow to deduce a cause from a symptom?

→ Anyway, this still doesn’t allow to compare the plausibility of different causes.

→ Logics does not allow to weigh different alternatives, and it does not allow to express knowledge we’re not fully certain about (“all possible causes of toothache”).
Unreliable Sensors

**Example:** Robot localization.

Suppose we want to support localization using landmarks (e.g., Eiffel tower) to narrow down the area. (“If you see the Eiffel tower, then you’re in Paris”.)

**Difficulty:** Sensors can be imprecise.

- Even if a landmark is perceived, we cannot conclude with certainty that the robot is at that location.
- Even if no landmark is perceived, we cannot conclude with certainty that the robot is not at that location.

→ Only the **probability** of being in that area increases or decreases.
Beliefs and Probabilities

What do we model with probabilities?

- We (and other agents) often do not know for sure whether or not a fact/rule holds true.
- We typically believe to a certain degree that it is true.
- Probabilities measure the degree of belief in a fact, given our current knowledge.
- The agent is 90% convinced by its sensor information = in 9 out of 10 cases, the information is correct.
- Symptom(p, toothache) \Rightarrow Disease(p, cavity) with 80% (or 0.8) probability = in 8 out of 10 cases, toothache is caused by cavity.
- For any given p, in reality we do/do not have cavity, 1 or 0. The “probability”

→ Probabilities represent and measure the uncertainty that stems from lack of knowledge.
How do we obtain probabilities?

- Use **statistics** to **assess** (measure) such degrees of belief:
  - The agent is 90% convinced by its sensor information = in 9 out of 10 cases, the information is correct.
  - *Symptom*(p, toothache) => *Disease*(p, cavity) with 80% probability = in 8 out of 10 cases, toothache is caused by cavity.
    
    The “80%” refers to the fraction within the set of all p' that are indistinguishable from p as far as our knowledge is concerned.

What is probabilistic reasoning? Deduce probabilities from knowledge about other probabilities.

- Probabilistic reasoning determines, based on probabilities that are relatively easy to assess, probabilities that are very difficult to assess.
(Uncertainty and Rational Decisions)

Here, we’re only concerned with deducing the likelihood of facts, not with action choice. In general, selecting actions is of course important.

**Rational Agents:**

- We have a choice of **actions** (go to FRA early, go to FRA just in time).
- These can lead to different solutions with different **probabilities**.
- The **actions** have different **costs**.
- The **results** have different **utilities**

→ The **ideal rational agent** chooses the action with the **maximum expected utility**.

**Decision Theory = Utility Theory + Probability Theory.**
A particular kind of utility-based agent:

function DT-AGENT(percept) returns an action

persistent: belief_state, probabilistic beliefs about the current state of the world
action, the agent’s action

update belief_state based on action and percept
calculate outcome probabilities for actions,
given action descriptions and current belief_state
select action with highest expected utility
given probabilities of outcomes and utility information

return action
Question!

Do the two statements “I don’t know which side of the dice will appear” vs. “Each side will appear with the same probability” have the same meaning?

(A): Yes.  (B): No.
Questionnaire, ctd.

Question!

What are sources of uncertainty in your life?

(A): Not knowing whether a train will be late.
(B): Not knowing what the exam questions will be.
(C): Not knowing whether the road can safely be crossed.
(D): Not knowing the outcome of a dice throw.
Unconditional Probabilities

**Definition.** Given a random variable $X$, $P(X = x)$ denotes the unconditional probability, or prior probability, that $X$ has value $x$ in the absence of any other information.

**Example:** $P(Cavity = true) = 0.2$, where $Cavity$ is a random variable whose value is true iff some given person has a cavity.

→ We will refer to the fact $X = x$ as an event, or an outcome.

→ The notation uppercase “$X$” for a variable, and lowercase “$x$” for one of its values will be used frequently. (Follows Russel/Norvig.)
Random Variables

In general, random variables can have arbitrary domains. Here, we consider finite-domain random variables only, and Boolean random variables most of the time.

Example:

\[
\begin{align*}
P(Weather = \text{sunny}) &= 0.7 \\
P(Weather = \text{rain}) &= 0.2 \\
P(Weather = \text{cloudy}) &= 0.08 \\
P(Weather = \text{snow}) &= 0.02 \\
P(Headache = \text{true}) &= 0.1
\end{align*}
\]

→ By convention, we denote Boolean random variables with \( A, B \), and more general finite-domain random variables with \( X, Y \).

→ For Boolean variable \( \text{Name} \), we write \( \text{name} \) for \( \text{Name} = \text{true} \) and \( \neg \text{name} \) for \( \text{Name} = \text{false} \).
Propositions

**Definition.** Given a set $V$ of random variables $X$ each with domain $D_X$, a *proposition* is a propositional formula over the atoms

$$\{X = c \mid X \in V, c \in D_v\}.$$  

A function $P$ that maps propositions into $[0, 1]$ is a *probability measure* if (i) $P(\top) = 1$ and (ii) for all propositions $A$, $P(a) = \sum_{I \models A} P(I)$ where $I$ is an interpretation of $\{X = c \mid X \in V, c \in D_v\}$.

→ Propositions can be viewed as Boolean random variables; we will denote them with $A, B$ as well.

→ Propositions represent sets of combined outcomes: the interpretations satisfying the formula.

**Example:** $P(\text{cavity} \land \text{toothache}) = 0.12$ is the probability that some given person has both a cavity and a toothache. (Note the use of cavity for $Cavity = \text{true}$; same for Toothache.)
**Definition.** The probability distribution for a random variable $X$, denoted $P(X)$, is the vector of probabilities for the (ordered) domain of $X$.

**Example:**

\[
P(\text{Headache}) = \langle 0.1, 0.9 \rangle
\]

\[
P(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02 \rangle
\]

define the probability distributions for the random variables *Headache* and *Weather*.

→ For a set of variables, the joint probability distribution. E.g., $P(\text{Headache, Weather})$ is a $4 \times 2$ table:

<table>
<thead>
<tr>
<th></th>
<th>Headache = true</th>
<th>Headache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Weather} = \text{sunny}$</td>
<td>$P(W = \text{sunny} \land \text{headache})$</td>
<td>$P(W = \text{sunny} \land \neg \text{headache})$</td>
</tr>
<tr>
<td>$\text{Weather} = \text{rain}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Weather} = \text{cloudy}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Weather} = \text{snow}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conditional Probabilities: Intuition

Does the probability of an outcome change when we gather additional knowledge?

**Example:** The probability of a cavity increases if we know the patient has a toothache.

→ Given additional information, we can no longer use the prior probabilities!

Given propositions $A$ and $B$, $P(a \mid b)$ denotes the conditional probability of $a$ ($A = \text{true}$) given that all we know is $b$ ($B = \text{true}$).

**Example:** $P(\text{cavity}) = 0.2$ vs. $P(\text{cavity} \mid \text{toothache}) = 0.6$. And $P(\text{cavity} \mid \text{toothache} \land \neg \text{cavity}) =$
Definition. Given propositions $A$ and $B$ where $P(b) \neq 0$, the conditional probability, or posterior probability, of $a$ given $b$, denoted $P(a \mid b)$, is defined as:

$$P(a \mid b) = \frac{P(a \land b)}{P(b)}.$$ 

→ The likelihood of having $a$ and $b$, within the set of outcomes where we have $b$.

Example: $P(cavity \land toothache) = 0.12$ and $P(toothache) = 0.2$ yield $P(cavity \mid toothache) =$
Definition. Given random variables \( X \) and \( Y \), the conditional probability distribution of \( X \) given \( Y \), denoted \( P(X \mid Y) \), is the table of all conditional probabilities of values of \( X \) given values of \( Y \).

→ For sets of variables: \( P(X_1, \ldots, X_n \mid Y_1, \ldots, Y_m) \).

Example: \( P(Weather \mid Headache) \) is a \( 4 \times 2 \) table:

<table>
<thead>
<tr>
<th></th>
<th>Headache = true</th>
<th>Headache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Weather = sunny )</td>
<td>( P(W = sunny \mid \text{headache}) )</td>
<td>( P(W = sunny \mid \neg \text{headache}) )</td>
</tr>
<tr>
<td>( Weather = rain )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Weather = cloudy )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Weather = snow )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Product Rule

**Proposition (Product Rule).** Given propositions $A$ and $B$, 
$$P(a \land b) = P(a \mid b)P(b).$$
(Direct from definition.)

→ If we know the values of $P(b)$ and $P(a \mid b)$, then we can compute $P(a \land b)$.

→ Similarly, $P(a \land b) = P(b \mid a)P(a)$.

**Notation.** $P(X, Y) = P(X \mid Y)P(Y)$ is a system of equations:

\[
\begin{align*}
P(W = \text{sunny} \land \text{headache}) &= P(W = \text{sunny} \mid \text{headache})P(\text{headache}) \\
P(W = \text{rain} \land \text{headache}) &= P(W = \text{rain} \mid \text{headache})P(\text{headache}) \\
\ldots &= \ldots \\
P(W = \text{snow} \land \neg\text{headache}) &= P(W = \text{snow} \mid \neg\text{headache})P(\neg\text{headache})
\end{align*}
\]

→ Similar for unconditional distributions, $P(X, Y) = P(X)P(Y)$. 

Jörg Hoffmann and Wolfgang Wahlster Artificial Intelligence 23/89
Independence

Definition. Events $a$ and $b$ are said to be independent if

$$P(a \land b) = P(a)P(b).$$

Proposition. Given independent events $a$ and $b$ where $P(b) \neq 0$, we have $P(a \mid b) = P(a)$.

Proof.

→ Similarly, if $P(a) \neq 0$, we have $P(b \mid a) = P(b)$.

Example. $P(Dice1 = 6 \land Dice2 = 6) = \frac{1}{36}$ However, $P(toothache) = 0.2$ and $P(cavity) = 0.2$, but

$P(cavity \land toothache) = 0.12 > 0.04!$ (Of the toothache cases, many have a cavity.)

Definition. Random variables $X$ and $Y$ are said to be independent if

$$P(X, Y) = P(X)P(Y).$$
Theorem (Kolmogorov’s axioms). A function $P$ that maps propositions into $[0, 1]$ is a probability measure if and only if (i) $P(\top) = 1$ and (ii) for all propositions $A, B$: $P(a \lor b) = P(a) + P(b) - P(a \land b)$.

→ In other words, we can equivalently replace our previous “(ii) for all propositions $A$, $P(a) = \sum_{I \models A} P(I)$” with the present condition.

Example. How to derive that, for all propositions $A$, $P(\neg a) = 1 - P(a)$?

We have $P(a \lor \neg a) = \quad$ and $P(a \land \neg a) = \quad$, and thus $1 = P(a) + P(\neg a) - 0$. 
Why are the Axioms Reasonable?

If $P$ represents an objectively observable probability, the axioms clearly make sense.

But why should an agent respect these axioms, when modeling its subjective own belief?
→ The axioms limit the set of beliefs that an agent can maintain.

There’s been a looong discussion . . .

de Finetti, 1931: If an agent has a belief that violates Kolmogorov’s axioms, then there exists a combination of “bets” on propositions so that the agent always looses money.

→ If your beliefs are contradictory, then you will not be successful in the long run (and even the next minute if your opponent is clever).
Questionnaire

Question!

About “household animals”: Say \( P(\text{dog}) = 0.4, \neg \text{dog} \Leftrightarrow \text{cat}, \) and \( P(\text{likesLasagna} \mid \text{cat}) = 0.5. \) Then \( P(\text{likesLasagna} \land \text{cat}) = \)

(A): 0.4

(C): 0.475

(B): 0.5

(D): 0.3

Question!

Can we compute the value of \( P(\text{likesLasagna}), \) given the above informations?

(A): Yes.

(B): No.
The Full Joint Probability Distribution

Definition. Given random variables $X_1, \ldots, X_n$, an atomic event is an assignment of values to all variables. (→ An interpretation/a state.)

Example. If $A$ and $B$ are Boolean random variables, then we have 4 atomic events:

- $A = 1, B = 1$
- $A = 1, B = 0$
- $A = 0, B = 1$
- $A = 0, B = 0$

Definition. Given random variables $X_1, \ldots, X_n$ and a probability measure $P$, the full joint probability distribution, denoted $P(X_1, \ldots, X_n)$, assigns a probability to every atomic event.

Example.

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>$\neg$toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>$\neg$cavity</td>
<td>0.08</td>
<td>0.72</td>
</tr>
</tbody>
</table>

→ Since all atomic events are disjoint (their conjunction is $\bot$), the sum of all fields is
Example.

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>¬cavity</td>
<td>0.08</td>
<td>0.72</td>
</tr>
</tbody>
</table>

How to compute $P(cavity)$?

How to compute $P(cavity \lor toothache)$?

How to compute $P(cavity \mid toothache)$?

→ All relevant probabilities can be computed using the full joint probability distribution, by expressing propositions as disjunctions of atomic events.
Is it a good idea to work with the full joint probability distribution?

- Computational cost of dealing with this size.
- Difficult to assess all these probabilities.

→ Is there a compact way to represent the full joint probability distribution?

→ Is there an efficient method to work with that representation?

→ Not in general, but it works in many cases. We can work directly with conditional probabilities, and exploit *conditional independence*.

→ **Bayesian networks**, later in this chapter. (First we need more basics.)
Illustration: Exploiting (Full) Independence

Example.

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>¬cavity</td>
<td>0.08</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Adding variable Weather with values \{sunny, rain, cloudy, snow\}, the full joint probability distribution contains how many probabilities?

But do your teeth influence the weather, or vice versa?

→ Weather is independent of each of Cavity and Toothache: For all value combinations \(c, t\) of Cavity and Toothache, and for all values \(w\) of Weather, we have \(P(c \land t \land w) = P(c \land t)P(w)\).

→ \(P(Cavity, Toothache, Weather)\) can be reconstructed from separate tables for Cavity, Toothache and Weather.

→ Independence can be used to represent the full joint probability distribution more compactly.

Sometimes variables are independent only under particular conditions

→ conditional independence.
The Chain Rule

**Proposition (Chain Rule).** Given random variables $X_1, \ldots, X_n$, we have

$$P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1) \cdot P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \cdot \cdots \cdot P(X_2 \mid X_1) \cdot P(X_1).$$

**Proof.** By iterated application of the product rule.

→ Given the same set of variables, this works for any ordering.

→ We can recover the probability of atomic events from sequenced conditional probabilities for any ordering of the variables.

→ One of the four basic techniques in Bayesian networks. (We will enhance the chain rule with conditional independence.)
Marginalization

→ Extract a sub-distribution from a larger joint distribution:

**Proposition (Marginalization).** Given sets of random variables $X$ and $Y$, we have:

$$P(X) = \sum_{y \in Y} P(X, y) = \sum_{y \in Y} P(X \mid y)P(y).$$

where $\sum_{y \in Y}$ sums over all possible value combinations of $Y$.

**Example.**

$$P(Cavity) = \sum_{y \in \{Toothache\}} P(Cavity, y)$$

$$P(cavity) = P(cavity, toothache) + P(cavity, \neg toothache)$$

$$P(\neg cavity) = P(\neg cavity, toothache) + P(\neg cavity, \neg toothache)$$
Normalization

Example. The two cases cavity, $\neg$cavity sum up to 1:

\[
P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.12}{0.2} = 0.6
\]

\[
P(\neg\text{cavity} \mid \text{toothache}) = \frac{P(\neg\text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.08}{0.2} = 0.4
\]

$\rightarrow$ Ignore constant factor $1/P(\text{toothache})$, normalize afterwards!

Definition. Given a vector $\langle p_1, \ldots, p_k \rangle$ of numbers in $[0, 1]$, the **normalization constant** $\alpha$ is defined as $1/\sum_{i=1}^{k} p_i$.

$\rightarrow$ Here: $\alpha\langle0.12, 0.8\rangle = \text{Normalized vector}$

Proposition (Normalization). Given a random variable $X$ and an event $e$, we have $P(X \mid e) = \alpha P(X, e)$.
Example. Use $X = \text{Cavity}$ and $e = \text{toothache}$. Then:

\[
P(C\text{avity} \mid \text{toothache}) = \alpha P(C\text{avity}, \text{toothache})
\]

\[
P(c\text{avity} \mid \text{toothache}) = \alpha P(c\text{avity} \land \text{toothache}) = \alpha 0.12
\]

\[
P(\neg c\text{avity} \mid \text{toothache}) = \alpha P(\neg c\text{avity} \land \text{toothache}) = \alpha 0.08
\]

We get $\alpha$ for $\langle 0.12, 0.08 \rangle =$

$\rightarrow$ Instead of assessing $P(e)$, we can perform a case analysis over $P(X, e)$ (for all values of $X$).

$\rightarrow$ Since the sum of all cases is 1, and the factor for a fixed event $e$ is constant, it suffices to know the cases’ relative weight.

Normalization+Marginalization. Given variable $X$, event $e$, and variable set $Y$: $P(X \mid e) = \alpha P(X, e) = \alpha \sum_{y \in Y} P(X, e, y)$.

$\rightarrow$ Another one of the four basic techniques in Bayesian networks.
About “household animals”: Say we have two random variables, Animal and LikesChappi where the latter is Boolean and the former has values \{dog, cat, other\}. Say we know \( P(\text{dog}) = 0.4 \) and \( P(\text{likesChappi} \mid \text{dog}) = 0.8 \). Can we compute \( P(\text{likesChappi} \land \text{dog}) \)?

(A): Yes. (B): No.
Questionnaire

Question!

**About “household animals”:** Say we have two random variables, *Animal* and *LikesChappi* where the latter is Boolean and the former has values \{*dog*, *cat*, *other*\}. Say we know

\[
P(Animal) = \langle 0.4, 0.4, 0.2 \rangle \quad \text{and} \quad P(likeschappi \mid Animal) = \langle 0.8, 0.1, 0.7 \rangle.
\]

Can we compute \(P(LikesChappi)\)?

(A): Yes. \quad (B): No.
About "household animals": Say we know
\[ P(\text{likeschappi} \land \text{dog}) = 0.32 \quad \text{and} \quad P(\neg\text{likeschappi} \land \text{dog}) = 0.08. \]
Can we compute \[ P(\text{likeschappi} \mid \text{dog})? \]
(A): Yes.  (B): No.
Proposition (**Bayes’ Rule**). Given propositions $A$ and $B$ where $P(a) \neq 0$ and $P(b) \neq 0$, we have:

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}.$$ 

**Proof.** By the product rule, we have $P(a \land b) = P(a \mid b)P(b)$ and $P(a \land b) = P(b \mid a)P(a)$. Thus from which the claim follows.

**Notation.** System of equations:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}.$$
Applying Bayes’ Rule

**Example.** Say we know that \( P(\text{toothache} \mid \text{cavity}) = 0.6, \)
\( P(\text{cavity}) = 0.2, \) and \( P(\text{toothache}) = 0.2. \) Can we compute
\( P(\text{cavity} \mid \text{toothache})? \)

Why don’t we simply assess \( P(\text{cavity} \mid \text{toothache}) \) directly?

\( P(\text{toothache} \mid \text{cavity}) \) is **causal**, \( P(\text{cavity} \mid \text{toothache}) \) is **diagnostic**.

- Typically, causal dependencies are easier to assess.
- Causal dependencies are robust over frequency of the causes.

Example: If there is a cavity epidemic then \( P(\text{cavity}) \) increases, and
with it \( P(\text{toothache}) \) and \( P(\text{cavity} \mid \text{toothache}) \); but
\( P(\text{toothache} \mid \text{cavity}) \) remains the same.

\( \rightarrow \) Bayes’ rule allows to perform diagnosis based on knowledge about
prior probabilities and causal dependencies.
Weighing the Alternatives

**Example.** The patient has toothache. We wish to diagnose whether it’s more likely to be a cavity or gum disease. We know as before that 

\[ P(\text{toothache} \mid \text{cavity}) = 0.6 \text{ and } P(\text{cavity}) = 0.2, \] 

and we know that 

\[ P(\text{toothache} \mid \text{gum disease}) = 0.8 \text{ and } P(\text{gum disease}) = 0.02. \] 

We do not know the value of \( P(\text{toothache}) \). Can we perform the diagnosis?

**Proposition.** *Given propositions \( A_1, A_2, \) and \( B \) where \( P(a_i) \neq 0 \) and \( P(b) \neq 0 \), we have:*

\[
\frac{P(a_1 \mid b)}{P(a_2 \mid b)} = \frac{P(b \mid a_1) P(a_1)}{P(b \mid a_2) P(a_2)}. 
\]

**Proof.** By Bayes’ rule,

\[
\frac{P(a_1 \mid b)}{P(a_2 \mid b)} = \frac{P(b \mid a_1) P(a_1)}{P(b \mid a_2) P(a_2)} \times \frac{P(b)}{P(b \mid a_2) P(a_2)} 
\]

\( \rightarrow P(b) \) can be cancelled! No need to assess it.
Weighing the Alternatives, ctd.

**Same Example.** We know $P(\text{toothache} \mid \text{cavity}) = 0.6$, $P(\text{cavity}) = 0.2$, $P(\text{toothache} \mid \text{gumdisease}) = 0.8$, $P(\text{gumdisease}) = 0.02$. Which is higher, $P(\text{cavity} \mid \text{toothache})$ or $P(\text{gumdisease} \mid \text{toothache})$?

**Reminder.**

$$\frac{P(a_1 \mid b)}{P(a_2 \mid b)} = \frac{P(b \mid a_1)P(a_1)}{P(b \mid a_2)P(a_2)}$$

→ We use $A_1 = \phantom{\text{toothache}}$, $A_2 = \phantom{\text{cavity}}$, $B = \phantom{\text{toothache}}$

→ Abbreviating the names, we get:

$$\frac{P(c \mid t)}{P(g \mid t)} = \frac{P(t \mid c)P(c)}{P(t \mid g)P(g)} = \frac{0.6 \times 0.2}{0.8 \times 0.02} = 7.5$$

So which diagnosis is more likely?
Normalization and Bayes’ Rule

Say we know $P(a)$ and $P(b \mid a)$, and want $P(a \mid b)$. An alternative to assessing $P(b)$ is to assess $P(b \mid \neg a)$: $(P(\neg a) = 1 - P(a)$ is known)

**Proposition.** Given propositions $A$ and $B$ where $P(a) \neq 0$ and $P(b) \neq 0$, we have:

$$P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b \mid a)P(a) + P(b \mid \neg a)P(\neg a)}.$$

**Proof.** Use Bayes’ rule and the fact that (\*) $P(a \mid b) + P(\neg a \mid b) = 1$:

(A) $P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}$; (B) $P(\neg a \mid b) = \frac{P(b \mid \neg a)P(\neg a)}{P(b)}$.

(A) + (B): (C) $P(a \mid b) + P(\neg a \mid b) = \frac{P(b \mid a)P(a)}{P(b)} + \frac{P(b \mid \neg a)P(\neg a)}{P(b)}$.

(C) \*P(b), with (\*): (D) $P(b) = P(b \mid a)P(a) + P(b \mid \neg a)P(\neg a)$.

And now?
What did this have to do with normalization?

**Reminder.** Normalization for variable $X$ and event $e$:

$$P(X \mid e) = \alpha P(X, e)$$

**Reminder.** Product rule:

$$P(X, e) = P(e \mid X) P(X)$$

Putting this together yields “normalization+product”:

$$P(X \mid e) = \alpha P(X, e) = \alpha P(e \mid X) P(X)$$

→ Applied to previous proposition, with $A$ as $X$ and $b$ as event $e$:

$$P(A \mid b) = \alpha P(b \mid A) P(A)$$

$$P(a \mid b) = \alpha P(b \mid a) P(a)$$

$$P(\neg a \mid b) = \alpha P(b \mid \neg a) P(\neg a)$$

where $\alpha = 1/[P(b \mid a) P(a) + P(b \mid \neg a) P(\neg a)]$. 
Normalization and Bayes’ Rule: Example

Your doctor tells you that you have tested positive for a serious but rare (1/10000) disease. This test is correct to 99%, i.e., 1% false positive and 1% false negative results. Should you be afraid?

\[ P(d \mid t) = \frac{P(t \mid d)P(d)}{P(t)} = \frac{P(t \mid d)P(d)}{P(t \mid d)P(d) + P(t \mid \neg d)P(\neg d)} \]

\[ P(d) = \quad ; \quad P(t \mid d) = \quad ; \quad P(t \mid \neg d) = \quad . \]

\[ P(d \mid t) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = \frac{0.000099}{0.000099 + 0.009999} \approx 0.01 \]

→ If the test imprecision is much greater than the frequency of the disease, then a positive result is not as threatening as you might think.
Multiple Evidence

Example. Say we know from medicinical studies that \( P(\text{cavity}) = 0.2 \), \( P(\text{toothache} \mid \text{cavity}) = 0.6 \), \( P(\text{toothache} \mid \neg \text{cavity}) = 0.1 \), \( P(\text{catch} \mid \text{cavity}) = 0.9 \), and \( P(\text{catch} \mid \neg \text{cavity}) = 0.2 \). Say we observe the symptoms \text{toothache} and \text{catch}, i.e., the dentist’s probe catches in the aching tooth. What is \( P(\text{cavity} \mid \text{toothache} \land \text{catch}) \)?

\[
\text{→ Bayes’ rule:} \quad P(\text{cavity} \mid \text{toothache} \land \text{catch}) = \frac{P(\text{toothache} \land \text{catch} \mid \text{cavity})P(\text{cavity})}{P(\text{toothache} \land \text{catch})}
\]

\[
\text{→ Normalization + product on } X = \text{Cavity and } e = \text{toothache} \land \text{catch}:
\]

\[
P(\text{Cavity} \mid \text{toothache}, \text{catch}) = \alpha P(\text{toothache}, \text{catch} \mid \text{Cavity})P(\text{Cavity})
\]
\[
P(\text{cavity} \mid \text{toothache} \land \text{catch}) = \alpha P(\text{toothache} \land \text{catch} \mid \text{cavity})P(\text{cavity})
\]
\[
P(\neg \text{cavity} \mid \text{toothache} \land \text{catch}) = \alpha P(\text{toothache} \land \text{catch} \mid \neg \text{cavity})P(\neg \text{cavity})
\]
Multiple Evidence, ctd.

\[ P(Cavity \mid toothache, catch) = \alpha P(toothache, catch \mid Cavity) P(Cavity) \]

**Question!**

**So, is everything fine?**

(A): Yes.  
(B): No.

**Are Toothache and Catch independent?**

\[ P(Toothache, Catch \mid cavity) = P(Toothache \mid cavity) P(Catch \mid cavity) \]

\[ P(Toothache, Catch \mid \neg cavity) = P(Toothache \mid \neg cavity) P(Catch \mid \neg cavity) \]

(For cavity: this may cause both, but they don’t influence each other. For \neg cavity: catch and/or toothache would each be caused by something else.)
Conditional Independence

**Definition.** Given sets of random variables \( Z_1, Z_2, Z \), we say that \( Z_1 \) and \( Z_2 \) are *conditionally independent* given \( Z \) if:

\[
P(Z_1, Z_2 \mid Z) = P(Z_1 \mid Z)P(Z_2 \mid Z).
\]

We alternatively say that \( Z_1 \) is conditionally independent of \( Z_2 \) given \( Z \).

**Example.** Normalization+product: \( P(Cavity \mid toothache, catch) = \alpha P(toothache, catch \mid Cavity)P(Cavity) \).

\[
\Rightarrow \text{Using } \{Toothache\} \text{ as } Z_1, \{Catch\} \text{ as } Z_2, \text{ and } \{Cavity\} \text{ as } Z, \text{ from the equation system defining conditional independence, for values toothache and catch, we get } \alpha P(toothache, catch \mid Cavity)P(Cavity) =
\]

\[
\Rightarrow \text{For } Cavity = \text{true we get } P(cavity \mid toothache \land catch) \approx 0.96. \quad \text{For } Cavity = \text{false we get } \text{So } \alpha \approx 8.92 \text{ and }
\]
Conditional Independence, ctd.

**Proposition.** If $Z_1$ and $Z_2$ are conditionally independent given $Z$, then $P(Z_1 \mid Z_2, Z) = P(Z_1 \mid Z)$.

**Example.** Using \{Toothache\} as $Z_1$, \{Catch\} as $Z_2$, and \{Cavity\} as $Z$: $P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity})$.

→ Similar for \{Catch\} as $Z_1$, \{Toothache\} as $Z_2$, and \{Cavity\} as $Z$?

→ But there may be dependencies within $Z_1$ or $Z_2$ (e.g., $Z_2 = \{\text{Toothache}, \text{Sleeplessness}\}$).
Exploiting Conditional Independence: Overview

1. **Graph captures variable dependencies.** (Variables $X_1, \ldots, X_n$.)

2. **Normalization + Marginalization.**
   \[
   P(X \mid e) = \alpha P(X, e) = \alpha \sum_{y \in Y} P(X, e, y).
   \]
   \[\rightarrow \text{A sum over atomic events!}\]

3. **Chain rule.** Order $X_1, \ldots, X_n$ consistent with dependency graph.
   \[
   P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1)P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \ldots P(X_1).
   \]

4. **Exploit conditional independence.** Instead of $P(X_i \mid X_{i-1} \ldots, X_1)$, with proposition on previous slide we can use $P(X_i \mid \text{Parents}(X_i))$.

   \[\rightarrow \text{Bayesian networks!}\]
Questionnaire

**Question!**

**About “household animals”:** Say \( P(\text{dog}) = 0.4, \)
\( P(\text{likeschappi} \mid \text{dog}) = 0.8, \) **and** \( P(\text{likeschappi}) = 0.5. \) **Can we compute** \( P(\text{dog} \mid \text{likeschappi})? \)

(A): Yes. \hspace{1cm} (B): No.
About “household animals”: Say we have the random variables and dependencies above, where \( \text{LikesChappi} \) and \( \text{LoudNoise} \) are Boolean, and \( \text{Animal} \) has values \{dog, cat, other\}. Say we know
\[
\begin{align*}
P(\text{Animal}) &= \langle 0.4, 0.4, 0.2 \rangle \quad \text{and} \\
P(\text{LoudNoise} \mid \text{Animal}) &= \langle 0.7, 0.2, 0.01 \rangle \quad \text{and} \\
P(\text{LikesChappi} \mid \text{Animal}) &= \langle 0.8, 0.1, 0.7 \rangle. \end{align*}
\]
Can we compute \( P(\text{dog} \mid \text{LoudNoise}, \text{LikesChappi}) \)?

(A): Yes.  

(B): No.
\[ \mathbf{P}(\text{likeschappi} \mid \text{Animal}) = \langle 0.8, 0.1, 0.7 \rangle, \quad \mathbf{P}(\text{loudnoise} \mid \text{Animal}) = \langle 0.7, 0.2, 0.01 \rangle, \]
\[ \mathbf{P}(\text{Animal}) = \langle 0.4, 0.4, 0.2 \rangle. \]
Summary

- **Uncertainty** is unavoidable in complex, dynamic worlds in which agents do not have perfect knowledge.

- **Probabilities** express the degree of belief of an agent, given its knowledge, into an event.

- **Conditional** probabilities express the likelihood of an event given observed evidence.

- **Bayes’ rule** allows us to derive, from probabilities that are easy to assess, probabilities that aren’t easy to assess.

- Given multiple evidence, we can exploit **conditional independence**.

  → Bayesian networks (up next) do this, in a comprehensive manner.
What is a Bayesian network? (Short: BN)

“A Bayesian network is a methodology for representing the full joint probability distribution. In some cases, that representation is compact.”

“A Bayesian network is a graph whose nodes are random variables $X_i$ and whose edges $(X_j, X_i)$ denote a direct influence of $X_j$ on $X_i$. Each node $X_i$ is associated with a conditional probability table (CPT), specifying $P(X_i \mid \text{Parents}(X_i))$.”

“A Bayesian network is a graphical way to depict conditional independence relations within a set of random variables.”

→ A Bayesian network (BN) represents the structure of a given domain. Probabilistic inference exploits that structure for improved efficiency.

→ BN inference: Determine the distribution of a query variable $X$ given observed evidence $e$: $P(X \mid e)$. 
**Example.** I got very valuable stuff at home. So I bought an alarm. Unfortunately, the alarm just rings at home, doesn’t call me on my mobile. I’ve got two neighbors, Mary and John who’ll call me if they hear the alarm. The problem is that, sometimes, the alarm is caused by an earthquake. Also, John might confuse the alarm with his telephone, and Maria might miss the alarm altogether because she typically listens to loud music.

**Question.** Given that both John and Mary call me, what is the probability of a burglary?
Cooking Recipe. (1) Design the random variables $X_1, \ldots, X_n$; (2) Identify their dependencies; (3) Insert the conditional probability tables $P(X_i \mid Parents(X_i))$.

Example. Let’s cook:

(1) Random variables:

(2) Dependencies:

(3) Conditional probability tables: Assess the probabilities, see next slide.
John, Mary, and the Alarm: The BN

(→ In each $P(X_i \mid Parents(X_i))$, we show only $P(X_i = true \mid Parents(X_i))$, not $P(X_i = false \mid parents(X_i))$ which is $1 - P(X_i = true \mid parents(X_i))$.)
Definition (Bayesian Network). Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$, a Bayesian network is an acyclic directed graph $(\{X_1, \ldots, X_n\}, E)$. We denote $\text{Parents}(X_i) := \{X_j \mid (X_j, X_i) \in E\}$. Each $X_i$ is associated with a function $\text{CPT}(X_i) : D_i \times (\times_{X_j \in \text{Parents}(X_i)} D_j) \mapsto [0, 1]$.

→ Also called belief networks, probabilistic networks, or causal networks in the literature.

→ This is a purely syntactical definition. What about the semantics?
The Meaning of BNs: Illustration

- **Alarm** depends on **Burglary** and **Earthquake**.
- **MaryCalls** only depends on **Alarm**.
  \[ P(MaryCalls \mid Alarm, Burglary) = P(MaryCalls \mid Alarm) \]

→ Bayesian networks are sets of independence assumptions.
Each node $X$ in a BN is conditionally independent of its non-descendants given its parents $\text{Parents}(X)$.
Given the value of *Alarm*, *MaryCalls* is independent of
The Meaning of BNs: Formal

Definition. Given a Bayesian network

\[ BN = (\{X_1, \ldots, X_n\}, E) \]

we identify BN with the following two assumptions:

\textbf{(A)} For \( 1 \leq i \leq n \), \( X_i \) is conditionally independent of \( \text{NonDescendants}(X_i) \) given \( \text{Parents}(X_i) \), where \( \text{NonDescendants}(X_i) := \{X_j \mid (X_i, X_j) \notin E^*\} \) with \( E^* \) denoting the transitive closure of \( E \).

\textbf{(B)} For \( 1 \leq i \leq n \), all values \( x_i \) of \( X_i \), and all value combinations \( \text{parents}(X_i) \) of \( \text{Parents}(X_i) \), we have

\[
P(x_i \mid \text{parents}(X_i)) = CPT(x_i, \text{parents}(X_i)).
\]
Recovering the Full Joint Probability Distribution

“A Bayesian network is a methodology for representing the full joint probability distribution. In some cases, that representation is compact.”

→ How do we recover the full joint probability distribution \( P(X_1, \ldots, X_n) \) from \( BN = (\{X_1, \ldots, X_n\}, E) \)?

**Chain rule.** For any ordering \( X_1, \ldots, X_n \), we have:

\[
P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1)P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \ldots P(X_1)
\]

Choose \( X_1, \ldots, X_n \) consistent with \( BN: X_j \in Parents(X_i) \implies j < i \).

**Exploit conditional independence.** With \( BN \) assumption (A), instead of \( P(X_i \mid X_{i-1} \ldots, X_1) \) we can use \( P(X_i \mid Parents(X_i)) \):

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i))
\]

The distributions \( P(X_i \mid Parents(X_i)) \) are given by \( BN \) assumption (B).

→ Same for atomic events \( P(x_1, \ldots, x_n) \).
Recovering a Probability for John, Mary, and the Alarm

\[
P(j, m, a, \neg b, \neg e) =
\]
\[
= 0.000062
\]
Compactness of Bayesian Networks

**Definition.** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$, the *size* of a $BN = (\{X_1, \ldots, X_n\}, E)$ is defined as

$$size(BN) := \sum_{i=1}^{n} |D_i| \times \prod_{\text{Parents}(X_i)} |D_j|.$$  
(The total number of entries in the CPTs.)

→ Smaller BN ⇒ assess less probabilities, more efficient inference.

- Explicit full joint probability distribution has size
- If $|\text{Parents}(X_i) \leq k|$ for every $X_i$, then $size(BN) \leq$
  
  → For $|D_{\text{max}}| = 2$, $n = 20$, $k = 4$ we have $2^{20} = 1048576$ probabilities, but a Bayesian network of size

- In the worst case, $size(BN) =$

→ BNs are compact if each variable is directly influenced only by few other variables.
Constructing a Bayesian Network

1. Initialize $BN := (\{X_1, \ldots, X_n\}, E)$ where $E = \emptyset$.
2. Fix any order of the variables, $X_1, \ldots, X_n$.
3. for $i := 1, \ldots, n$ do
   1. Choose, from $X_1, \ldots, X_{i-1}$, a minimal set $Parents(X_i)$ of parents so that $P(X_i \mid X_{i-1} \ldots, X_1) = P(X_i \mid Parents(X_i))$.
   2. For each $X_j \in Parents(X_i)$, insert $(X_j, X_i)$ into $E$.
   3. Associate $X_i$ with $CPT(X_i)$ corresponding to $P(X_i \mid Parents(X_i))$.

Which variables we need to include into $Parents(X_i)$ depends on what “$X_1, \ldots, X_{i-1}$” is . . .

→ The size of the resulting $BN$ depends on the chosen order $X_1, \ldots, X_n$!

→ The size of a Bayesian network is not a fixed property of the domain. It depends also on the skill of the designer.
John and Mary Depend on the Variable Order!

Example. MaryCalls, JohnCalls, Alarm, Burglary, Earthquake.
Example. MaryCalls, JohnCalls, Earthquake, Burglary, Alarm.
These BNs link from symptoms to causes \( P(Cavity \mid Toothache) \).

→ Get dependencies between independent causes, and often between conditionally independent symptoms.

→ Conditional probabilities \( P(Symptom \mid Cause) \) often easier to assess (like \( P(Toothache \mid Cavity) \)).

→ We should order causes before symptoms.
**Questionnaire**

**Question!**

About “household animals”: Say $BN$ is the Bayesian network above. Which statements are correct?

(A): LoudNoise is independent of LikesChappi
(B): Animal is independent of LikesChappi
(C): Animal is conditionally independent of LikesChappi
(D): LikesChappi is conditionally independent of LoudNoise given Animal
Questionnaire

Question!
What is the Bayesian network we get by constructing according to the ordering $X_1 = \text{LoudNoise}, X_2 = \text{Animal}, X_3 = \text{LikesChappi}$?

Question!
What is the Bayesian network we get by constructing according to the ordering $X_1 = \text{LoudNoise}, X_2 = \text{LikesChappi}, X_3 = \text{Animal}$?
"Instantiate evidence variables and draw conclusions on query variables."

What is $P(\text{Burglary} \mid \text{johncalls})$?

What is $P(\text{Burglary} \mid \text{johncalls, marycalls})$?
Definition (Probabilistic Inference Task). Given random variables \( X_1, \ldots, X_n \), a probabilistic inference task consists of a set \( \mathbf{X} \subseteq \{X_1, \ldots, X_n\} \) of query variables, a set \( \mathbf{E} \subseteq \{X_1, \ldots, X_n\} \) of evidence variables, and an event \( e \) that assigns values to \( E \).

- We assume that a \( BN \) for \( X_1, \ldots, X_n \) is given.
- We wish to compute the posterior probability distribution \( P(X \mid e) \).
- We will refer to \( \mathbf{Y} := \{X_1, \ldots, X_n\} \backslash (\mathbf{X} \cup \mathbf{E}) \) as the hidden variables.

→ In the remainder, for simplicity, \( \mathbf{X} = \{X\} \) is a singleton.

Example.

\[ P(\text{Burglary} \mid \text{johncalls, marycalls}) = (0.284, 0.716) \]
Motivation
Probability Theory
Basic Inference
Bayes' Rule
Bayesian Networks
Inference in BNs
Closing Remarks

Inference by Enumeration: The Principle

Given evidence $e$, want to know $P(X \mid e)$. Hidden variables: $Y$.

1. **Bayesian network $BN$ captures variable dependencies.**

2. **Normalization+Marginalization.**
   
   $P(X \mid e) = \alpha P(X, e) = \alpha \sum_{y \in Y} P(X, e, y)$.
   
   $\rightarrow$ Recover the summed-up probabilities $P(X, e, y)$ from $BN$!

3. **Chain rule.** Order $X_1, \ldots, X_n$ consistent with $BN$.
   
   $P(X_1, \ldots, X_n) = P(X_n \mid X_{n-1}, \ldots, X_1)P(X_{n-1} \mid X_{n-2}, \ldots, X_1) \ldots P(X_1)$.

4. **Exploit conditional independence.** Instead of $P(X_i \mid X_{i-1} \ldots, X_1)$,
   
   use $P(X_i \mid Parents(X_i))$.
   
   $\rightarrow$ Given a Bayesian network $BN$, probabilistic inference tasks can be solved as sums of products of conditional probabilities from $BN$.

$\rightarrow$ Sum over all value combinations of hidden variables.
Inference by Enumeration: John and Mary

- Consider $P(Burglary \mid johncalls, marycalls)$

- Hidden variables:
  - $P(B \mid j, m)$

Consider $Burglary = true$, i.e., $b$; similar for $\neg b$.

- By conditional independence:
  - $P(b \mid j, m) = \alpha(0.00059224, 0.0014919) \approx (0.284, 0.716)$
The Evaluation of $P(b \mid j, m)$, as a Search Tree

Inference by enumeration explores a tree whose nodes along each path are ordered according to $X_1, \ldots, X_n$. Branching is done over the hidden variables.
Algorithm for Inference by Enumeration

**function** ENUMERATION-ASK($X$, $e$, $bn$) **returns** a distribution over $X$

**inputs:** $X$, the query variable
$e$, observed values for variables $E$
$bn$, a Bayes net with variables $\{X\} \cup E \cup Y$ /* $Y$ = hidden variables */

$Q(X) \leftarrow$ a distribution over $X$, initially empty

for each value $x_i$ of $X$ do
  $Q(x_i) \leftarrow$ ENUMERATE-ALL($bn$.VARS, $e_{x_i}$)
  where $e_{x_i}$ is $e$ extended with $X = x_i$

return NORMALIZE($Q(X)$)

**function** ENUMERATE-ALL($vars$, $e$) **returns** a real number

if EMPTY?($vars$) then return 1.0

$Y \leftarrow$ FIRST($vars$)

if $Y$ has value $y$ in $e$
  then return $P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $e$)
else return $\sum_y P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $e_y$)
  where $e_y$ is $e$ extended with $Y = y$
Inference by Enumeration: Properties

→ The enumeration algorithm evaluates the tree in a depth-first manner.
  
  • Space complexity:
  
  • Time complexity:

→ In general, exact probabilistic inference is computationally hard (\#P, harder than NP).

→ The variable elimination algorithm improves on enumeration by avoiding repeated computation, and avoiding irrelevant computation.

→ In some special cases, variable elimination runs in polynomial time.
Variable Elimination in a Nutshell

**Avoiding repeated computation:** Evaluate expressions from right to left, storing all intermediate results. Query $P(B \mid j, m)$:

→ At start, $CPT$s of $BN$ yield factors (probability tables):

$$P(B \mid j, m) = \alpha \left( P(B) \sum_e P(e) \sum_a P(a \mid B, e) P(j \mid a) P(m \mid a) \right)$$

→ Then the computation is performed in terms of factor product and summing out variables from factors:

$$P(B \mid j, m) = \alpha f_1(B) \times \sum_e f_2(E) \times \sum_a f_3(A, B, E) \times f_4(A) \times f_5(A)$$

**Avoiding irrelevant computation.** Repeatedly remove leaf nodes outside $X \cup E$. Query $P(JohnCalls \mid burglary)$:

$$P(J \mid b) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(J \mid a) \sum_m P(m \mid a)$$

The rightmost sum equals 1 and can be dropped.
Good News. A graph is called **singly connected**, or a **polytree**, if there is at most one undirected path between any two nodes in the graph. *On polytree Bayesian networks, variable elimination runs in polynomial time.*

→ Is our $BN$ for Mary & John a polytree?

Bad News. For **multiply connected** Bayesian networks, in general probabilistic inference is **#P-hard**.

Life goes on . . . In the hard cases, if need be we can throw exactitude to the winds and **approximate**. *Example: Sampling techniques.*
**Questionnaire**

**About “household animals”:** Say $BN$ is the Bayesian network above. How can we compute $P(\text{dog} \mid \text{loudnoise})$?

(A): \[ P(\text{d} \mid \text{ln}) = \alpha \sum_{\text{lc}} P(\text{lc} \mid \text{d}) P(\text{ln} \mid \text{d}) \]

(B): \[ P(\text{d} \mid \text{ln}) = \alpha \sum_{\text{lc}} P(\text{lc} \mid \text{d}) P(\text{ln} \mid \text{d}) P(\text{d}) \]

(C): \[ P(\text{d} \mid \text{ln}) = \alpha P(\text{ln} \mid \text{d}) P(\text{d}) \sum_{\text{lc}} P(\text{lc} \mid \text{d}) \]

(D): \[ P(\text{d} \mid \text{ln}) = \alpha P(\text{ln} \mid \text{d}) P(\text{d}) \]
Other Approaches

Default Reasoning:

- **Motivation:** Do *you* make probability calculations in your head?
- **Method:** Instead “believe there is no burglary until proved otherwise”.
- **Downside:** Firm proofs often do not exist in practice.

Fuzzy Logic:

- **Motivation:** Capture vagueness (“John is tall”).
- **Method:** $Tall(John) \in [0, 1]$ is a “rule of thumb”, $\lor$ is max, $\land$ is min.
- **Downside:** Unintuitive behavior (e.g., $Tall(John) \land \neg Tall(John) = 0.4$), success only in applications that do not actually require reasoning.
Summary

- Bayesian networks (sometimes) allow a **compact representation** of the full joint probability distribution.
- Bayesian networks represent **conditional independence** relations in a domain.
- **Probabilistic inference** in Bayesian networks requires to compute the probability distribution of a set of **query variables**, given a set of **evidence variables**.
- Exact probabilistic inference is $\#P$-hard in general, but is easier for restricted **network structure**.
- **Variable elimination** is efficient for **poly-tree BNs**.
- Approximate probabilistic inference methods exist.
- Bayesian networks have nowadays largely surplanted all other approaches.