

Artificial Intelligence

ENCS 434

Uncertainty & Probabilistic Reasoning

Logic

- Logics are characterized by what they commit to as "**primitives**".

Logic	What Exists in World	Knowledge States
Propositional	facts	true/false/unknown
First-Order	facts, objects, relations	true/false/unknown
Temporal	facts, objects, relations, times	true/false/unknown
Probability Theory	facts	degree of belief 0..1
Fuzzy	degree of truth	degree of belief 0..1

Probability

- $P(a)$ is the probability of proposition “a”
 - E.g., $P(\text{it will rain in London tomorrow})$
 - The proposition a is actually true or false in the real-world
 - $P(a)$ = “prior” or marginal or unconditional probability
 - Assumes no other information is available
- Axioms:
 - $0 \leq P(a) \leq 1$
 - $P(\text{NOT}(a)) = 1 - P(a)$
 - $P(\text{true}) = 1$
 - $P(\text{false}) = 0$
 - $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$

Probability and Logic

- Probability can be viewed as a generalization of propositional logic
- $P(a)$:
 - a is any sentence in propositional logic
 - Belief of agent in a is no longer restricted to *true, false, unknown*
 - $P(a)$ can range from 0 to 1
 - $P(a) = 0$, and $P(a) = 1$ are special cases
 - So logic can be viewed as a special case of probability

Sources of Uncertainty

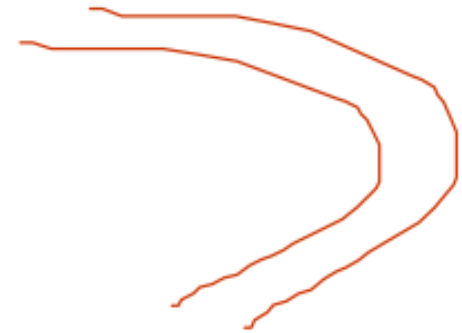
Being uncertain/hesitate/not sure/ in your decision may be due to:

- Information is incomplete.
- Information is not fully reliable.
- Representation language is inaccurate.
- Information comes from multiple sources and it is conflicting.

- e.g. Which mode of transportation is safer?
 - Car or Plane?
 - What is the probability of an accident?
- Probability theory enables us to make rational decisions.

Example of Uncertainty

- Assume a camera and vision system is used to estimate the curvature of the road ahead.
- **There's uncertainty about which way it curves**
 - Limited pixel resolution, noise in image
 - Algorithm for “road detection” is not perfect
- **This uncertainty can be represented with a simple probability model:**



$$P(\text{road curves to left}|E) = 0.6$$

$$P(\text{road goes straight}|E) = 0.3$$

$$P(\text{road curves to right}|E) = 0.1$$

- * Where the probability of an event is a measure of agent's belief in the event given the evidence **E**.

Rules with Uncertainty

- If we are sure that only **cavity causes toothache** then we can add the following rule into the KB of the ES.

- If **toothache** then **problem is cavity**

- **But not all patients have toothaches due to cavities**

So we can set up a rule like:

- If **toothache** \wedge \neg (**gum disease**) \wedge \neg (**filling**) \wedge ... then **problem = cavity**

- Another method would be:

- If **toothache** then **problem is cavity with 0.8 probability**

or $P(\text{cavity}|\text{toothache}) = 0.8$

- **The probability of cavity is 0.8 given toothache is all that is known.**

Making Decisions under Uncertainty

Suppose I believe the following are the possible decisions:

$$P(A_1 \text{ gets me there on time} \mid \dots) = 0.04$$

$$P(A_2 \text{ gets me there on time} \mid \dots) = 0.70$$

$$P(A_3 \text{ gets me there on time} \mid \dots) = 0.95$$

$$P(A_4 \text{ gets me there on time} \mid \dots) = 0.9999$$

■ Which action to choose?

Depends on my preferences for missing flight vs. time spent waiting,

....

- **Utility theory** is used to represent and infer preferences
- **Decision theory** = probability theory + utility theory

Uncertainty in the World Model

- **True uncertainty:** rules **are** probabilistic in nature
 - Rolling dice, flipping a coin?
- **Laziness:** too hard to determine exception less rules
 - Takes too much work to determine all of the relevant factors.
 - Too hard to use the enormous rules that result.
- **Theoretical ignorance:** don't know all the rules
 - Problem domain has no complete theory (**medical diagnosis**).
- **Practical ignorance:** do know all the rules **BUT**
 - Haven't collected all relevant information for a particular case

Handling Uncertain knowledge

- Probability provides a way of summarizing the uncertainty that comes from our laziness and/or ignorance.
- An assignment of probability to a proposition is analogous to saying whether or not a given logical sentence is entailed by the knowledge base.
- The agent's knowledge provides only a degree of belief in the relevant sentences.
- As the agent receives new percepts, its probability assessments are updated to reflect the new evidence.
- Before the evidence is obtained, we talk about prior or unconditional probability.
- After the evidence is obtained, we talk about posterior or conditional probability.

Syntax

■ **Random Variables (RV):**

- Are capitalized (usually) e.g. Sky, RoadCurvature, Temperature
- Refer to attributes of the world whose "status" is unknown
- Have one and only one value at a time.
- Have a **domain** of **values** that are possible states of the world:

- **Boolean:** Domain = <true, false>

Cavity=true abbreviated as cavity

Cavity=false abbreviated as \neg cavity

- **Discrete:** Domain is countable (includes Boolean)

Values are **exhaustive and mutually exclusive**

e.g. Sky domain = <clear, partly_cloudy, overcast>

Sky=clear abbreviated as clear

Sky is not clear also abbreviated as \neg clear

Syntax: Events

- **Any collection of outcomes**
- **Simple event**
 - Outcome with 1 characteristic, Probability of tossing 1 coin
- **Compound event**
 - Collection of outcomes or simple events
 - 2 or more characteristics, tossing 2 coins
- **Joint event**
 - 2 events occurring simultaneously
 - Probability of being rich and happy
- **Experiment: Tossing 2 coins.**

<u>Event</u>	<u>Outcomes in Event</u>
Sample space	HH, HT, TH, TT
1 head & 1 tail	HT, TH
Heads on 1st coin	HH, HT
At least 1 head	HH, HT, TH
Heads on both	HH

Syntax: Atomic Events

- An **atomic event**: an assignment of particular values to all the variables (complete specification of the state of the domain).
- If the world consists of only **two Boolean variables Cavity and Toothache**, then **there are 4 distinct atomic events**:
 - $\text{Cavity} = \text{false} \wedge \text{Toothache} = \text{false}$
 - $\text{Cavity} = \text{false} \wedge \text{Toothache} = \text{true}$
 - $\text{Cavity} = \text{true} \wedge \text{Toothache} = \text{false}$
 - $\text{Cavity} = \text{true} \wedge \text{Toothache} = \text{true}$
- **Properties of atomic events**:
 - They're **mutually exclusive**:
 - At most one can be the case
 - Set of all possible atomic events is **exhaustive**:
 - At least one must be the case

Prior versus Conditional Probability

- **Prior probability (A):** Probability of A in absence of any other information
- **Conditional Probability (A|B):** Probability of A given that we already know B

$$P(\text{Cavity}) = 0.1$$

10% of all individuals have a Cavity

$$P(\text{Toothache}) = 0.05$$

5% have a Toothache

$$P(\text{Cavity}|\text{Toothache}) = 0.8$$

given that we know the individual has Toothache, there is 80% chance of him having Cavity

$$P(\text{Cavity}|\text{Toothache} \wedge \text{not Gumdisease}) = 0.9$$

additionally given that another diagnosis is already excluded, conditional probability increases

$$P(\text{Cavity}|\text{Toothache} \wedge \text{FalseTeeth}) = 0$$

adding information does not necessarily increase the probability

Assigning Probabilities

■ **A priori classical method**

- Objects have a tendency to behave in certain ways
- Coin has a propensity to come up heads with a probability .5. Some scientists say Coin has a propensity to come up heads with a probability .333.

■ **Empirical classical method**

- Probabilities come from experiments
- If 10 of 100 people tested have a cavity then $P(\text{cavity}) = .1$
- Probability means the fraction that would be observed in the limit of infinitely many samples

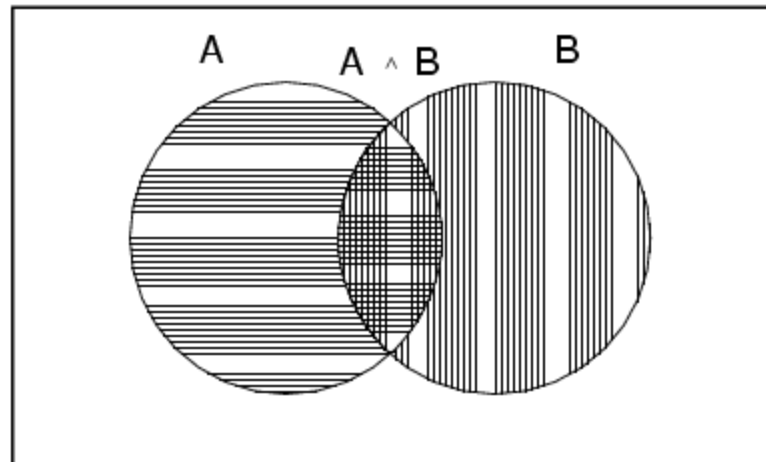
■ **Subjective method**

- Probabilities characterize an **agent's belief or point of view**
- Have no external physical significance

Axioms of probability

- For any propositions A, B
 - $0 \leq P(A) \leq 1$
 - $P(\text{true}) = 1$ and $P(\text{false}) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

True



Using the axioms of probability

- $P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$

(by axiom 3 with $b = \neg a$)

- $P(\text{true}) = P(a) + P(\neg a) - P(\text{false})$

(by logical equivalence)

- $1 = P(a) + P(\neg a)$ *(by axiom 2)*

- $P(\neg a) = 1 - P(a)$ *(by algebra).*

$$\sum_{i=1} P(D = d_i) = 1.$$

$$P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i).$$

Probability Distributions

Given A is a Random Variable taking values in $\langle a_1, a_2, \dots, a_n \rangle$

- $P(a)$ represents a **single probability** where $A=a$, e.g. $P(a) = 1/n$

e.g. if A is Sky, and the **domain of $A = \langle \text{clear, partly_cloudy, overcast} \rangle$**

then $P(a)$ means any one of $P(\text{clear})$, $P(\text{partly_cloudy})$, $P(\text{overcast})$

- **Probability Distribution: $\sum P(A_i)$**

- If A takes n values, then $P(A)$ is a set of n probabilities

- The set of values $\{P(a_1), P(a_2), \dots, P(a_n)\}$

- **Property: $\sum P(a_i) = P(A=a_1) + P(A=a_2) + \dots + P(A=a_n) = 1$**

Sum over all values in the domain of variable A is **1** if the domain is **exhaustive and mutually exclusive**.

Joint Distribution

- **Numerical measure of likelihood that joint event will occur**
- $P(a, b, \dots)$: **Joint probability** of $A=a \wedge B=b \wedge \dots$
- If we have k random variables X_1, \dots, X_k
- The **joint distribution** of these variables is a table in which each entry gives the probability of one combination of values of X_1, \dots, X_k
- **Example:**

	Toothache	\neg Toothache
Cavity	0.04	0.06
\neg Cavity	0.01	0.89

$P(\neg\text{Cavity} \wedge \text{Toothache})$

$P(\text{Cavity} \wedge \neg\text{Toothache})$

Joint Distribution Says It All

	Toothache	\neg Toothache
Cavity	0.04	0.06
\neg Cavity	0.01	0.89

- $P(\text{Toothache}) = P((\text{Toothache} \wedge \text{Cavity}) \vee (\text{Toothache} \wedge \neg \text{Cavity}))$
 $= P(\text{Toothache} \wedge \text{Cavity}) + P(\text{Toothache} \wedge \neg \text{Cavity})$
 $= 0.04 + 0.01 = 0.05$
- $P(\text{Toothache} \vee \text{Cavity})$
 $= P((\text{Toothache} \wedge \text{Cavity}) \vee (\text{Toothache} \wedge \neg \text{Cavity})$
 $\vee (\neg \text{Toothache} \wedge \text{Cavity}))$
 $= 0.04 + 0.01 + 0.06 = 0.11$

Conditional Probability

□ **Conditional Probabilities**

□ Specify the belief in a proposition that is conditioned on a proposition being true.

□ **P(a|e): Conditional Probability** of **A=a** given **E=e** evidence is all that is known true.

■ **P(A|B) : Probability of A given B**

□ $P(\text{Cavity}|\text{Toothache}) = 0.8$

■ $P(A|B) = P(A \wedge B) / P(B)$

■ $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$

■ $P(A \wedge B \wedge C) = P(A|B,C) P(B|C) P(C)$

■ **Example:**

■ **E** is a set of symptoms, such as, coughing, sneezing, headache, ...

■ **H** is a diseases, e.g., common cold, flu,

■ The diagnosis problem is to find an **H** (diseases) such that **P(H|E)** is maximum.

Conditional Probability


- Conditional probabilities behave exactly like standard probabilities:
- $0 \leq P(a|e) \leq 1$: between 0 and 1 inclusive.
- $P(a_1|e) + P(a_2|e) + \dots + P(a_n|e) = 1$: sum to 1 where a_1, \dots, a_n are all values in the domain of RV A .
- **Negation for conditional probabilities:** $P(\neg a|e) = 1 - P(a|e)$
- **P(conjunction of events | e):** $P(a \wedge b \wedge c | e)$ or as $P(a, b, c | e)$
 - The agent's belief in the sentence $a \wedge b \wedge c$ conditioned on e being true.
- **P(a | conjunction of evidences):** $P(a | e \wedge f \wedge g)$ or as $P(a | e, f, g)$
 - The agent's belief in the sentence a conditioned on $e \wedge f \wedge g$ being true.

Reasoning Under Uncertainty

Joint Probability Distribution(JPD)

- A **joint event** describes two occurrences at the same time.
 - e.g., **(A and B)** specifies that both propositions A and B are true in the world.
- A joint probability distribution over a set of random variables specifies a probability for each possible combinations of values for those variables.
 - e.g., a joint probability distribution for Boolean variables X and Y specifies a probability for four cases:
 - **P(X and Y), P(X and \neg Y), P(\neg X and Y), and finally P(\neg X and \neg Y)**
- The sum of the joint probabilities of all cases must be equal to 1
- The joint probability of two **events under absolute independence.**

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

-  The random variables A and B are called independent if occurrence of B does not influence on probability of A.

Full Joint Probability Distribution Table (FJPDT)

- **FJPDT**: Represents all the possible combination of an experiment repeated N times over the random variables A and B .

$$N = n11 + n12 + n21 + n22$$

- $n11$ represents the number of time $A \wedge B$ was observed, $n12$ represents the number of times $\neg A \wedge B$ was observed, $n21$ represents the number of time $A \wedge \neg B$, $n22$ represents the number of time $\neg A \wedge \neg B$.

	A	$\neg A$
B	$n11$	$n12$
$\neg B$	$n21$	$n22$

- $P(A) = n11 + n21 / N$ $P(B) = n11 + n12 / N$ $P(A \wedge B) = n11 / N$
- $P(A|B) = P(A \wedge B) / P(B) = n11 / n11 + n12$
- $P(B|A) = P(A \wedge B) / P(A) = n11 / n11 + n21$

Reasoning Under Uncertainty: Using FJPDT

👉 **What is $P(\text{green})$?**

$\frac{1}{2}$ since 50-50 % chance of green being picked over plum

$$P(a) = \sum P(e_i)$$

where e_i is an element of $\mathbf{e}(a)$

Who	What	Where	Probability
plum	rope	hall	1/8
plum	rope	study	1/8
plum	pipe	hall	1/8
plum	pipe	study	1/8
green	rope	hall	1/8
green	rope	study	1/8
green	pipe	hall	1/8
green	pipe	study	1/8

$$P(\text{green}) = ? \quad \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

$$P(\text{pipe}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

- * This is the **marginal probability** of pipe ignoring any information about the other events.
- * It can be just a prior probability.
- * This process is called **marginalization** or **summing out**.

Reasoning Under Uncertainty: Using FJPDT

👉 **What is** $P(\text{green, pipe, hall})$?

$$1/8 = P(\text{green}) * P(\text{pipe}) * P(\text{hall}) = 1/2 * 1/2 * 1/2$$

Who	What	Where	Probability
plum	rope	hall	1/8
plum	rope	study	1/8
plum	pipe	hall	1/8
plum	pipe	study	1/8
green	rope	hall	1/8
green	rope	study	1/8
green	pipe	hall	1/8
green	pipe	study	1/8

Prior probability for each is $1/8$

- Each equally likely

- e.g. $P(\text{plum,rope,hall}) = 1/8$

▪ $P(\text{atomic_event}_i) = 1$

- Since each RV's domain is exhaustive & mutually exclusive

- e.g. $1 = 1/8 + 1/8 + 1/8 + 1/8 + 1/8 + 1/8 + 1/8 + 1/8$

Reasoning Under Uncertainty: Using FJPD

👉 How do you figure out more complex probabilities?

$$P(a) = \sum P(e_i)$$

where e_i is an element of $e(a)$

Who	What	Where	Probability
plum	rope	hall	1/8
plum	rope	study	1/8
plum	pipe	hall	1/8
plum	pipe	study	1/8
green	rope	hall	1/8
green	rope	study	1/8
green	pipe	hall	1/8
green	pipe	study	1/8

$$P(\text{green, pipe}) = ? \quad 1/8 + 1/8 = 1/4$$

$$P(\text{rope, hall}) = ? \quad 1/8 + 1/8 = 1/4$$

$$P(\text{rope} \vee \text{hall}) = ?$$

$$1/8 + 1/8 + 1/8 + 1/8 + 1/8 + 1/8 = 3/4$$

Independence RV

- **We used the random variables Who, What, Where because they are independent.**
- **How are these RVs independent?**

Picking the card for one RV doesn't affect the others.

E.g. Picking the murder from the deck of “Who” cards doesn't affect which weapon is chosen or location.

- 👉 **Absolute Independence: The random variables X and Y are called independent if occurrence of Y does not influence on probability of X .**

1. $P(X|Y) = P(X)$

2. $P(Y|X) = P(Y)$

3. $P(X,Y) = P(X) P(Y)$

Independence RV

- **Conditional Independence:**

RVs (X, Y) are dependent on another RV (Z) but are independent of each other.

1. $P(X|Y,Z) = P(X|Z)$

2. $P(Y|X,Z) = P(Y|Z)$

3. $P(X,Y|Z) = P(X|Z) P(Y|Z)$

- **Idea:**

sneezing (x) and itchy eyes (y) are both directly caused by hayfever (z) but neither sneezing nor itchy eyes has a direct effect on each other.

- **This lets us decompose the joint distribution:**

- $P(A \wedge B \wedge C) = P(A | C) P(B | C) P(C)$

Reasoning under Uncertainty: Using FJPDT

- Assume three Boolean RVs: Hayfever (HF), Sneeze (SN), ItchyEyes (IE) and fictional probabilities:

HF	SN	IE	Probability
false	false	false	0.5
false	false	true	0.09
false	true	false	0.1
false	true	true	0.1
true	false	false	0.01
true	false	true	0.06
true	true	false	0.04
true	true	true	0.1

$$P(a) = \sum P(e_i)$$

where e_i is an element of $e(a)$

$$P(\text{sn}) = 0.1 + 0.1 + 0.04 + 0.1 = 0.34$$

$$P(\text{hf}) = 0.01 + 0.06 + 0.04 + 0.1 = 0.21$$

$$P(\text{sn,ie}) = 0.1 + 0.1 = 0.20$$

$$P(\text{hf,sn}) = 0.04 + 0.1 = 0.14$$

Reasoning under Uncertainty: Using FJPD

- Assume three Boolean RVs: Hayfever (HF), Sneeze (SN), ItchyEyes (IE) and fictional probabilities:

HF	SN	IE	Probability
false	false	false	0.5
false	false	true	0.09
false	true	false	0.1
false	true	true	0.1
true	false	false	0.01
true	false	true	0.06
true	true	false	0.04
true	true	true	0.1

$$P(a|e) = P(a, e) / P(e)$$

$$\begin{aligned} P(\text{hf} | \text{sn}) &= P(\text{hf,sn}) / P(\text{sn}) \\ &= 0.14 / 0.34 = 0.41 \end{aligned}$$

$$\begin{aligned} P(\text{hf} | \text{ie}) &= P(\text{hf,ie}) / P(\text{ie}) \\ &= 0.16 / 0.35 = 0.46 \end{aligned}$$

Combining Multiple Evidence

- Using the Full Joint Prob. Dist. Table:

- $$P(v_1, \dots, v_k | v_{k+1}, \dots, v_n) = \frac{\sum P(V_1=v_1, \dots, V_n=v_n)}{\sum P(V_{k+1}=v_{k+1}, \dots, V_n=v_n)}$$

1. Sum of all entries in the table, where $V_1=v_1, \dots, V_n=v_n$
2. Divided by the sum of all entries in the table corresponding to the evidence, where $V_{k+1}=v_{k+1}, \dots, V_n=v_n$

Combining Multiple Evidence

- Assume three Boolean RVs: Hayfever (HF), Sneeze (SN), ItchyEyes (IE) and fictional probabilities:

HF	SN	IE	Probability
false	false	false	0.5
false	false	true	0.09
false	true	false	0.1
false	true	true	0.1
true	false	false	0.01
true	false	true	0.06
true	true	false	0.04
true	true	true	0.1

$P(a|b, c) = P(a,b,c) / \sum P(b,c)$
as described in prior slide

$$\begin{aligned} P(hf | sn, ie) &= P(hf,sn,ie) / \sum P(sn,ie) \\ &= 0.1 / (0.1+0.1) \\ &= 0.5 \end{aligned}$$

Evaluating FJPDT

- **Advantage**

- All combinations are available
- Any joint or unconditional probability can be computed

- **Disadvantage**

- **Combinatorial Explosion!** For **N** variables, need 2^N individual probabilities.
- Difficult to get probabilities for all combinations

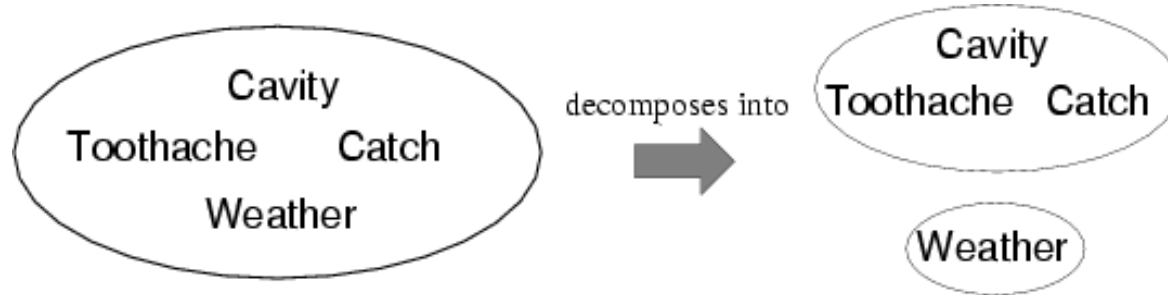
- As FJPDT is large, is there an alternative technique?

Yes - use **Bayes' Rule** to calculate probabilities and represent independence assertions using **Bayesian networks**.

Independence

- A and B are independent iff

$$\mathbf{P}(A | B) = \mathbf{P}(A) \text{ or } \mathbf{P}(B | A) = \mathbf{P}(B) \text{ or } \mathbf{P}(A, B) = \mathbf{P}(A) \mathbf{P}(B)$$



$$\begin{aligned} &\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather}) \\ &= \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Weather}) \end{aligned}$$

- **Absolute independence powerful but rare**
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional independence

- $\mathbf{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
(1) $\mathbf{P}(\textit{catch} \mid \textit{toothache}, \textit{cavity}) = \mathbf{P}(\textit{catch} \mid \textit{cavity})$
- The same independence holds if I haven't got a cavity:
(2) $\mathbf{P}(\textit{catch} \mid \textit{toothache}, \neg\textit{cavity}) = \mathbf{P}(\textit{catch} \mid \neg\textit{cavity})$
- *Catch* is **conditionally independent** of *Toothache* given *Cavity*:
 $\mathbf{P}(\textit{Catch} \mid \textit{Toothache}, \textit{Cavity}) = \mathbf{P}(\textit{Catch} \mid \textit{Cavity})$
- Equivalent statements:
 $\mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) = \mathbf{P}(\textit{Toothache} \mid \textit{Cavity})$
 $\mathbf{P}(\textit{Toothache}, \textit{Catch} \mid \textit{Cavity}) = \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity})$

Conditional independence

- Write out full joint distribution using chain rule:

$$\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch}, \textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity})$$

I.e., $2 + 2 + 1 = 5$ independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n .
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayesian networks

Bayes' Rule

- From the definition of conditional probability we have

- $p(a|b) = p(a,b) / p(b)$ (1)

- From the same definition we also have

- $p(b|a) = p(a,b) / p(a)$ (2)

- So, $p(a,b) = p(a|b) p(b)$ (from (1))
 $= p(b|a) p(a)$ (from (2))

- Dividing both sides by $p(b)$, we get:

$$p(a|b) = \frac{p(b|a) p(a)}{p(b)}$$

(This is Bayes' rule)

Why is Bayes' Rule useful?

- In practice an agent must reason as follows

effects -> causes

e.g., symptoms -> diseases

- But normally we build models in the “forward” causal direction

causes -> effects

e.g., diseases -> symptoms

- Bayes rule allows us to work “backward” using the output of the forward model to infer causes (inputs)

- Very useful in applications involving diagnosis

- Say we know: $p(d) = p(\text{disease}) = 0.001$, $p(s) = p(\text{symptom}) = 0.01$
and $p(\text{symptom } s | d) = 0.9$

- If someone has the symptom what is the probability they have the disease?

- We need to find $p(d|s)$ from the information above

$$p(d|s) = \frac{p(s|d) p(d)}{p(s)} = \frac{0.9 \times 0.001}{0.01} = 0.09$$

Examples: Bayes' Rule

- **Bayes' Rule:**

$$P(b|a) = P(a|b)P(b)/P(a)$$

- **For Example:**

a=happy, b=sun

$$P(\text{sun}|\text{happy}) = ?$$

$$P(\text{happy}|\text{sun}) = 0.95$$

$$P(\text{sun}) = 0.5$$

$$P(\text{happy}) = 0.75$$

$$(0.95 * 0.5)/0.75 = \mathbf{0.63}$$

a= sneeze, b= fall

$$P(\text{fall}|\text{sneeze}) = ?$$

$$P(\text{sneeze}|\text{fall}) = 0.85$$

$$P(\text{fall}) = 0.25$$

$$P(\text{sneeze}) = 0.3$$

$$(0.85 * 0.25)/0.3 = \mathbf{0.71}$$

Bayes' Rule: Example

- **Using Bayes' Rule with causal knowledge:**

- **Diagnostic reasoning:** want to determine likelihood of a cause given an effect, which is difficult to obtain from a general population.

- e.g. symptom is s =stiffNeck, disease is m =meningitis

$P(s|m) = 1/2$ the casual knowledge

$P(m) = 1/50000, P(s) = 1/20$ prior probabilities

$P(m|s) = ?$ desired diagnostic knowledge

$$(1/2 * 1/50000) / (1/20) = 1/5000$$

- **Doctor can now use $P(m|s)$ to guide diagnosis.**

Combining Multiple Evidence: Using Bayes' Rule

👉 How do you update conditional probability of **Y** given two pieces of evidence **A** and **B**?

- General Bayes' Rule for multi-valued RVs:

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

let $X=A,B$:

$$\begin{aligned} P(Y|A,B) &= P(A,B|Y) P(Y) / P(A,B) \\ &= P(Y) P(B|A,Y) P(A|Y) / (P(B|A) P(A)) \\ &= P(Y) * (P(A|Y)/P(A)) * (P(B|A,Y)/P(B|A)) \end{aligned}$$

- $P(Y|A,B) = P(Y) * (P(A|Y)/P(A)) * (P(B|Y)/P(B))$ (Bayes' Rule Multi-Evidence)

* This equation used to define a **naïve Bayes classifier**.

Combining Multiple Evidence: Using Bayes' Rule

■ Example:

□ What is the likelihood that a patient has sclerosis colangitis **تصلب الأنسجة**?

□ Doctor naïvely assumes jaundice and fibrosis are independent.

□ Doctor's initial belief: $P(sc) = 1/1,000,000$

□ Examination reveals jaundice: $P(j) = 1/10,000$

$$P(j|sc) = 1/5$$

□ Doctor's belief after exam.: $P(sc|j) = P(sc)P(j|sc)/P(j)$
 $= 2/1000$

□ Test ids fibrosis of bile ducts: $P(f|sc) = 8/10$

$$P(f) = 1/100$$

□ Doctor's belief now is: $P(sc|j,f) = 16/100$

$$P(sc|j,f) = P(sc) * (P(j|sc)/P(j)) * (P(f|sc)/P(f))$$

$$P(Y|A,B) = P(Y) * (P(A|Y)/P(A)) * (P(B|Y)/P(B))$$

Bayesian networks

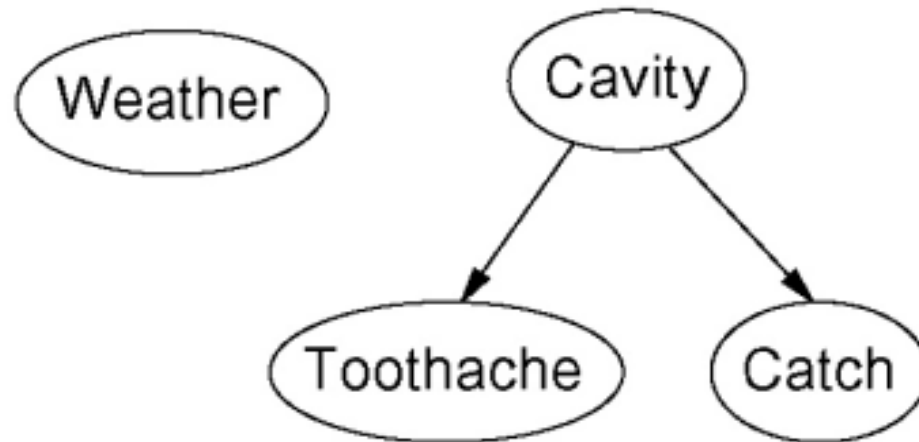
- ❑ A directed, acyclic graph (DAG)
- ❑ A set of nodes, one per variable (discrete or continuous)
- ❑ A set of directed links (arrows) connects pairs of nodes. X is a parent of Y if there is an arrow (direct influence) from node X to node Y .
- ❑ Each node X_i has a conditional probability distribution that quantifies the effect of the parents on the node.
- ❑ Combinations of the topology and the conditional distributions specify (implicitly) the full joint distribution for all the variables.

$$P(X_i \mid Parents(X_i))$$

Bayesian networks

Example 1 : The Teeth Disease Bayesian

Topology of network encodes conditional independence assertions:



Weather is independent of the other variables

Toothache and *Catch* are conditionally independent given *Cavity*

Examples of 3-way Bayesian Networks

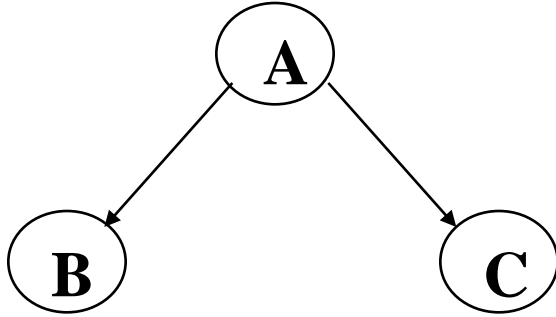
A

B

C

Marginal Independence:
 $p(A,B,C) = p(A) p(B) p(C)$

Examples of 3-way Bayesian Networks



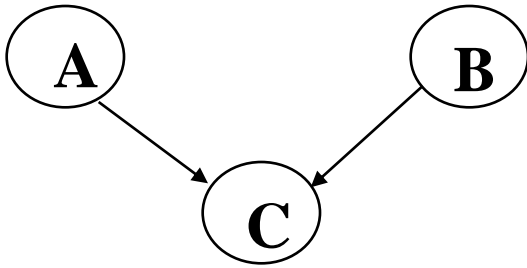
Conditionally independent effects:

$$p(A,B,C) = p(B|A)p(C|A)p(A)$$

**B and C are conditionally independent
Given A**

**e.g., A is a disease, and we model
B and C as conditionally independent
symptoms given A**

Examples of 3-way Bayesian Networks



Independent Causes:

$$p(A,B,C) = p(C|A,B)p(A)p(B)$$

“Explaining away” effect:

**Given C, observing A makes B less likely
e.g., earthquake/burglary/alarm example**

**A and B are (marginally) independent
but become dependent once C is known**

Examples of 3-way Bayesian Networks



Markov dependence:

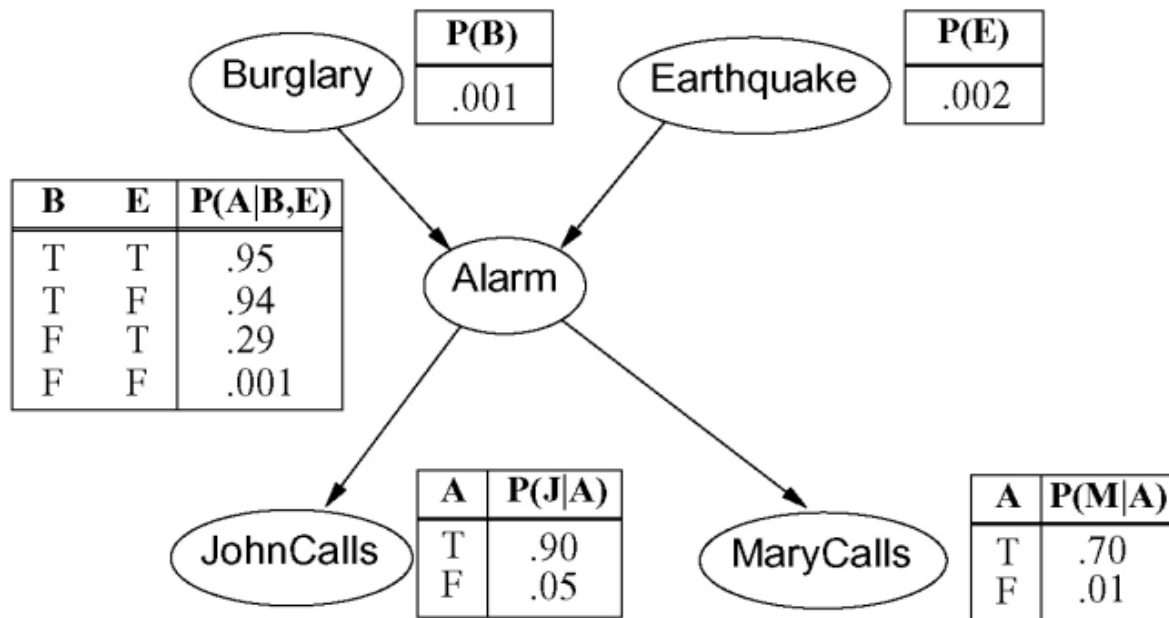
$$p(\mathbf{A}, \mathbf{B}, \mathbf{C}) = p(\mathbf{C}|\mathbf{B}) p(\mathbf{B}|\mathbf{A})p(\mathbf{A})$$

Example: Burglar alarm system

- I have a burglar alarm installed at home
 - It is fairly reliable at detecting a burglary, but also responds on occasion to minor earth quakes.
- I also have two neighbors, John and Mary
 - They have promised to call me at work when they hear the alarm
 - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
 - Mary likes rather loud music and sometimes misses the alarm altogether.
- Bayesian networks variables:
 - *Burglar, Earthquake, Alarm, JohnCalls, MaryCalls*

Example: Burglar alarm system

- Network topology reflects “causal” knowledge:
 - ❑ A burglar can set the alarm off
 - ❑ An earthquake can set the alarm off
 - ❑ The alarm can cause Mary to call
 - ❑ The alarm can cause John to call

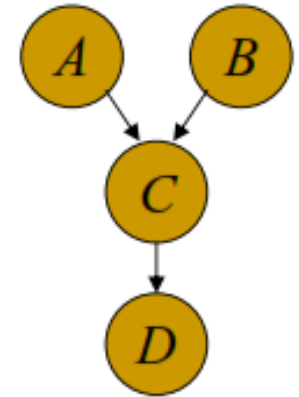


conditional probability table (CPT):
each row contains the conditional probability of each node value for a conditioning case (a possible combination of values for the parent nodes).

Computing Joint Probabilities: Using a Bayesian Network

1. Use product rule
2. Simplify using independence

For Example:



Compute $P(a,b,c,d) = P(d,c,b,a)$

order RVs in the joint probability bottom up D,C,B,A

$$= P(d|c,b,a) P(c,b,a)$$

Product Rule $P(d,c,b,a)$

$$= P(d|c) P(c,b,a)$$

Conditional Independ. of D given C

$$= P(d|c) P(c|b,a) P(b,a)$$

Product Rule $P(c,b,a)$

$$= P(d|c) P(c|b,a) P(b|a) P(a)$$

Product Rule $P(b,a)$

$$= P(d|c) P(c|b,a) P(b) P(a)$$

Independence of B and A
given no evidence

Computing Joint Probabilities: Using a Bayesian Network

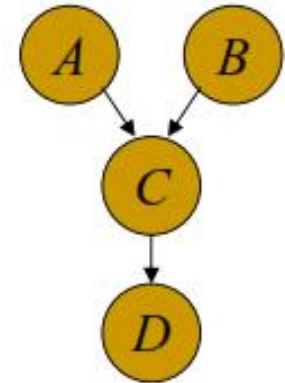
- **How is any joint probability computed?**
- **Answer:** Sum the relevant joint probabilities:
- **e.g. Compute:**

$$P(a,b) = P(a,b,c,d) + P(a,b,c,\neg d) + P(a,b,\neg c,d) + P(a,b,\neg c,\neg d)$$

- **e.g. Compute:**

$$P(c) = P(a,b,c,d) + P(a,\neg b,c,d) + P(\neg a,b,c,d) + P(\neg a,\neg b,c,d) + \\ P(a,b,c,\neg d) + P(a,\neg b,c,\neg d) + P(\neg a,b,c,\neg d) + P(\neg a,\neg b,c,\neg d)$$

- A BN can answer any query (i.e. probability) about the domain by summing the **relevant joint probabilities**.
- * Enumerating the relevant joint probabilities can require many computations!



Example: Computing JPD using BN

Lemma:

$$P(A | C) = P(A | B \wedge C)P(B | C) + P(A | \sim B \wedge C)P(\sim B | C)$$

Proof:

LHS $P(A | C) = \frac{P(A \wedge C)}{P(C)}$ by definition of cond prob

RHS $P(A | B \wedge C)P(B | C) + P(A | \sim B \wedge C)P(\sim B | C)$

by def. of
Cond. prob.

$$= \frac{P(A \wedge B \wedge C)}{P(B \wedge C)} \cdot \frac{P(B \wedge C)}{P(C)} + \frac{P(A \wedge \sim B \wedge C)}{P(\sim B \wedge C)} \cdot \frac{P(\sim B \wedge C)}{P(C)}$$

$$= \frac{P(A \wedge B \wedge C) + P(A \wedge \sim B \wedge C)}{P(C)}$$

$$= \frac{P(A \wedge C)}{P(C)}$$

An Example

- Let us compute $P(\text{Alarm}|\text{Burglary})$:

$$P(A | B) = P(A | E \wedge B)P(E | B) +$$

$$P(A | \sim E \wedge B)P(\sim E | B) \quad \text{by lemma}$$

$$P(A | B) = P(A | E \wedge B)P(E) +$$

$$P(A | \sim E \wedge B)P(\sim E) \quad \text{by abs. independence}$$

$$P(A | B) = (.95)(.002) + (.94)(.998) = .94$$

An Example

- **Let us compute $P(\text{JohnCalls}|\text{Burglary})$:**

$$P(J | B) = P(J | A \wedge B)P(A | B) + \\ P(J | \sim A \wedge B)P(\sim A | B) \quad \text{by lemma}$$

$$P(J | B) = P(J | A)P(A | B) + \\ P(J | \sim A)P(\sim A | B) \quad \text{by cond. independence}$$

$$P(J | B) = (.9)(.94) + (.05)(.06) = .85$$

This is an example of causal inference: from causes to effects.

An Example

- Let us compute $P(\text{MaryCalls}|\text{Burglary})$:

$$P(M | B) = P(M | A \wedge B)P(A | B) + P(M | \sim A \wedge B)P(\sim A | B) \quad \text{by lemma}$$

$$P(M | B) = P(M | A)P(A | B) + P(M | \sim A)P(\sim A | B) \quad \text{by cond. independence}$$

$$P(M | B) = (.7)(.94) + (.01)(.06) = .66$$

This is an example of causal inference: from causes to effects.

Probabilistic Reasoning: using a Bayesian Network

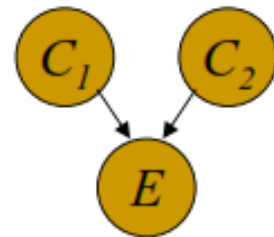
- **It is easy if the query involves nodes that are directly connected to each other. examples assumed to use Boolean RVs**
- **Simple causal inference: $P(E|C)$**
 - Conditional prob. dist. of effect **E** given cause **C** as evidence
 - Reasoning in same direction as arc, e.g. causes to effects
- **Simple diagnostic inference: $P(C|E)$**
 - Conditional prob. dist. of cause **C** given effect **E** as evidence
 - Reasoning in direction opposite of arc, e.g. effects to causes

Probabilistic Reasoning: Causal (Top-Down) Inference

Compute $P(e|c)$

Conditional probability of effect $E=e$ given cause $C=c$ as evidence
assume **arc** exists from C_1 and C_2 to E

1. **Rewrite conditional probability** of e in terms of e and all of its parents given evidence c
2. **Re-express each joint probability** back to the probability of e given all of its parents
3. **Simplify** using independence and **Look Up required values** in the Bayesian Network.



Probabilistic Reasoning: Causal (Top-Down) Inference

Compute $P(e|c_1)$

$$\begin{aligned} 1. \quad &= P(e, c_1) / P(c_1) && \text{product rule} \\ &= (P(e, c_1, c_2) + P(e, c_1, \neg c_2)) / P(c_1) && \text{marginalizing} \\ &= P(e, c_1, c_2) / P(c_1) + P(e, c_1, \neg c_2) / P(c_1) && \text{algebra} \\ &= P(e, c_2 | c_1) + P(e, \neg c_2 | c_1) && \text{product rule} \end{aligned}$$

$$2. \quad = P(e|c_2, c_1) P(c_2|c_1) + P(e|\neg c_2, c_1) P(\neg c_2|c_1) \quad \text{cond. chain rule}$$

3. Simplify given C_1 and C_2 are independent

$$P(c_2|c_1) = P(c_2)$$

$$P(\neg c_2|c_1) = P(\neg c_2)$$

$$= P(e|c_2, c_1) P(c_2) + P(e|\neg c_2, c_1) P(\neg c_2) \quad \text{algebra}$$

now look up values to finish computation

Probabilistic Reasoning: Diagnostic (Bottom-Up) Inference

Compute $P(c|e)$

Conditional probability of cause $C=c$ given effect $E=e$ as evidence
assume arc exists from C to E

Idea: convert to causal inference using Bayes' rule

1. **Use Bayes' rule** $P(c|e) = P(e|c) P(c) / P(e)$
2. **Compute $P(e|c)$ using causal inference method**
3. **Look up value of $P(c)$ in Bayesian Net**
4. **Use normalization to avoid computing $P(e)$**
 - Requires computing $P(\neg c|e)$
 - Using steps as in 1 – 3 above

Computing Joint Probabilities: Using a Bayesian Network

- Basic task of probabilistic system is to compute conditional probabilities.
- * Any conditional probability can be computed:

$$P(v_1, \dots, v_k | v_{k+1}, \dots, v_n) = \sum P(V_1=v_1, \dots, V_n=v_n) / \sum P(V_{k+1}=v_{k+1}, \dots, V_n=v_n)$$

- These computations generally rely on the simplifications resulting from the independence of the RVs.
- Every variable that isn't an ancestor/successor of a query variable or an evidence variable is irrelevant to the query.
- What ancestors are irrelevant?

Independence in a Bayesian Network

Given a Bayesian Network how is independence established?

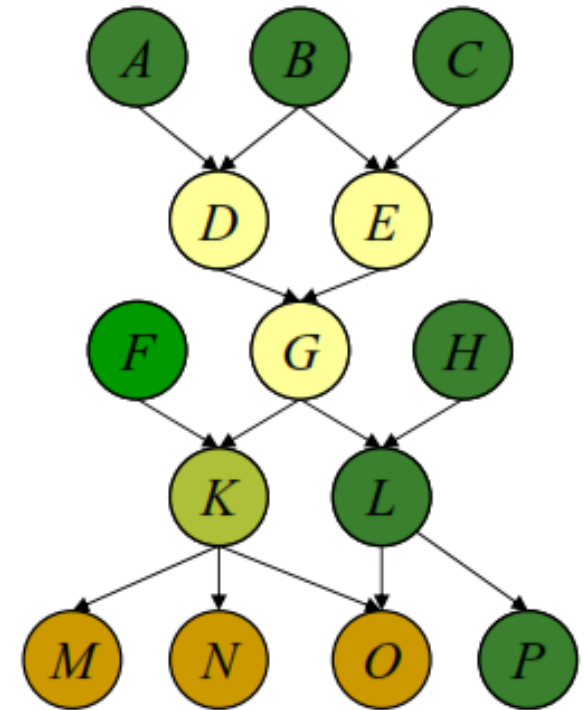
1. A node is conditionally independent (CI) of its non-descendants, given its parents.

e.g. Given **D** and **E**, **G** are CI of ?

A, B, C, F, H

e.g. Given **F** and **G**, **K** are CI of ?

A, B, C, D, E, H, L, P



Independence in a Bayesian Network

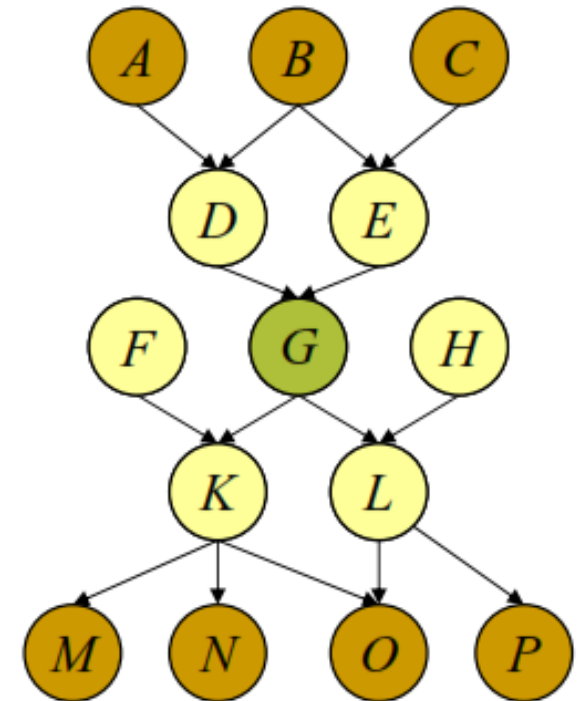
Given a Bayesian Network how is independence established?

2. A node is conditionally independent of all other nodes in the network given its parents, children, and children's parents, which is called a **Markov blanket**
e.g. **What is the Markov blanket for G?**

D, E, F, H, K, L

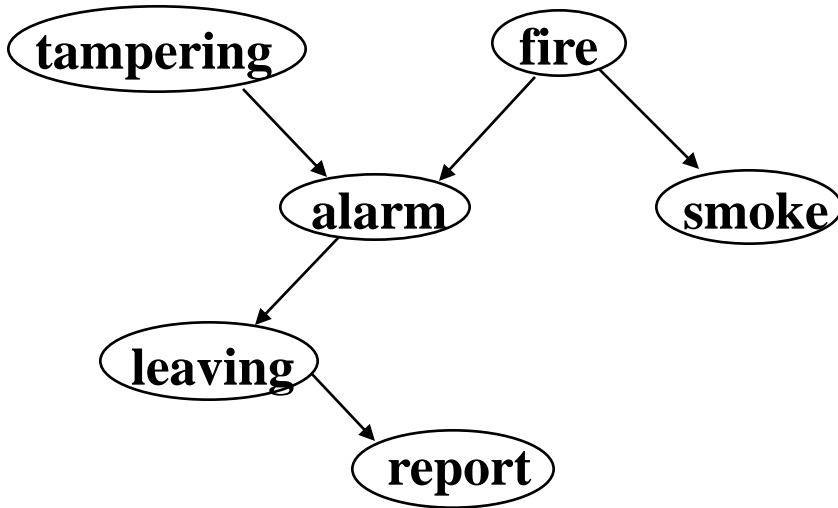
Given this blanket G is CI of ?

A, B, C, M, N, O, P



What about absolute independence?

Example



$$P(\text{tampering}) = 0.02$$

$$P(\text{fire}) = 0.01$$

$$P(\text{smoke}|\text{fire}) = 0.9$$

$$P(\text{smoke}|\sim\text{fire}) = 0.01$$

$$P(\text{alarm}|\text{fire}, \text{tamper}) = 0.5$$

$$P(\text{alarm}|\text{fire}, \sim\text{tamper}) = 0.99$$

$$P(\text{alarm}|\sim\text{fire}, \text{tamper}) = 0.85$$

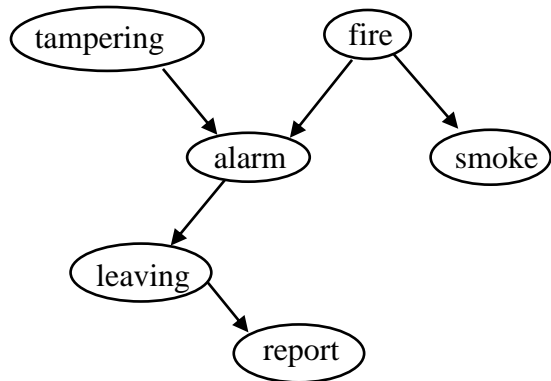
$$P(\text{alarm}|\sim\text{fire}, \sim\text{tamper}) = 0.0001$$

$$P(\text{leaving}|\text{alarm}) = 0.88$$

$$P(\text{leaving}|\sim\text{alarm}) = 0.001$$

$$P(\text{report}|\text{leaving}) = 0.75$$

$$P(\text{report}|\sim\text{leaving}) = 0.01$$



Example

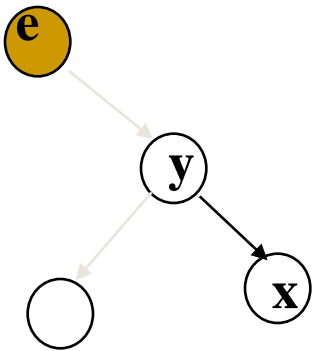
$$\begin{aligned}
 P(\text{leaving}|\text{smoke}) &= P(\text{leaving}|\text{alarm}, \text{smoke}) * P(\text{alarm}|\text{smoke}) \\
 &\quad + P(\text{leaving}|\sim\text{alarm}, \text{smoke}) * (1 - P(\text{alarm}|\text{smoke})) \\
 &= P(\text{leaving}|\text{alarm}) * P(\text{alarm}|\text{smoke}) \\
 &\quad + P(\text{leaving}|\sim\text{alarm}) * (1 - P(\text{alarm}|\text{smoke})) \\
 &= 0.88 * P(\text{alarm}|\text{smoke}) + 0.001 * (1 - P(\text{alarm}|\text{smoke}))
 \end{aligned}$$

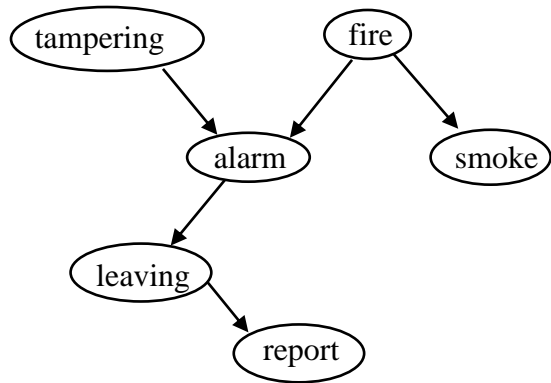
$$\begin{aligned}
 P(\text{alarm}|\text{smoke}) &= \\
 &P(\text{alarm}|\text{fire}, \text{tamper}, \text{smoke}) * P(\text{fire}, \text{tamper}|\text{smoke}) \\
 &+ P(\text{alarm}|\text{fire}, \sim\text{tamper}, \text{smoke}) * P(\text{fire}, \sim\text{tamper}|\text{smoke}) \\
 &+ P(\text{alarm}|\sim\text{fire}, \text{tamper}, \text{smoke}) * P(\sim\text{fire}, \text{tamper}|\text{smoke}) \\
 &+ P(\text{alarm}|\sim\text{fire}, \sim\text{tamper}, \text{smoke}) * P(\sim\text{fire}, \sim\text{tamper}|\text{smoke})
 \end{aligned}$$

Reasoning in a Belief Net

Updating belief in x based on evidence e from **non-descendants**:

$$\begin{aligned} P(x | e) &= \sum_v P(x, y=v | e) && \text{marginalize} \\ &= \sum_v P(x | y=v, e) * P(y=v | e) && \text{chain rule} \\ &= \sum_v P(x | y=v) * P(y=v | e) && \text{cond. independence} \end{aligned}$$





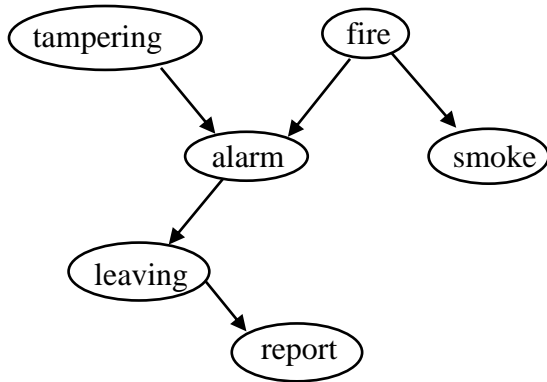
Example

$P(\text{alarm}|\text{smoke}) =$

$$\begin{aligned}
 & P(\text{alarm}|\text{fire, tamper}) * P(\text{fire, tamper}|\text{smoke}) \\
 & + P(\text{alarm}|\text{fire, } \sim\text{tamper}) * P(\text{fire, } \sim\text{tamper}|\text{smoke}) \\
 & + P(\text{alarm}|\sim\text{fire, tamper}) * P(\sim\text{fire, tamper}|\text{smoke}) \\
 & + P(\text{alarm}|\sim\text{fire, } \sim\text{tamper}) * P(\sim\text{fire, } \sim\text{tamper}|\text{smoke})
 \end{aligned}$$

$$\begin{aligned}
 P(\text{alarm}|\text{smoke}) = & 0.5 * P(\text{fire, tamper}|\text{smoke}) \\
 & + 0.99 * P(\text{fire, } \sim\text{tamper}|\text{smoke}) \\
 & + 0.85 * P(\sim\text{fire, tamper}|\text{smoke}) \\
 & + 0.0001 * P(\sim\text{fire, } \sim\text{tamper}|\text{smoke})
 \end{aligned}$$

Example



$P(\text{fire, tamper}|\text{smoke})$

$$= P(\text{fire}|\text{tamper, smoke}) * P(\text{tamper}|\text{smoke})$$

$$= P(\text{fire}|\text{tamper, smoke}) * P(\text{tamper})$$

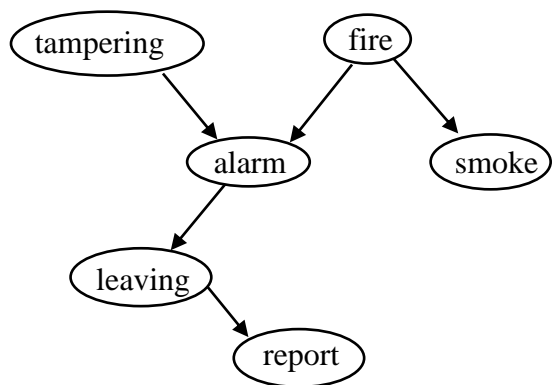
$$= 0.02 * P(\text{fire}|\text{tamper, smoke})$$

Conditioning on Descendents

If e includes descendents of x , separate into e_d and $e_{\sim d}$ where e_d involves only descendents of x and $e_{\sim d}$ contains only non-descendents. By Bayes' theorem,

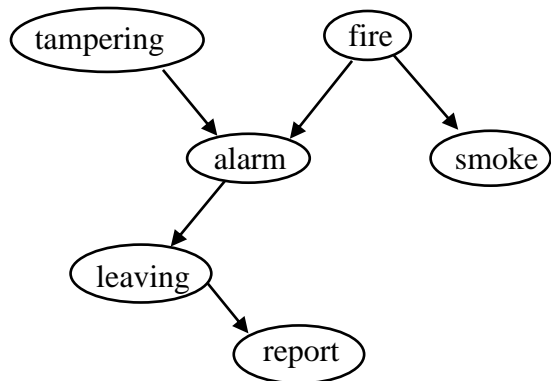
$$P(x | e_d, e_{\sim d}) = \frac{P(e_d | x, e_{\sim d}) * P(x | e_{\sim d})}{P(e_d | e_{\sim d})}$$

Probabilities in r.h.s. match specification of belief net.



Example

$$\begin{aligned}
 & P(\text{fire} | \text{tamper}, \text{smoke}) \\
 &= \frac{P(\text{smoke} | \text{fire}, \text{tamper}) * P(\text{fire} | \text{tamper})}{P(\text{smoke} | \text{tamper})} \\
 &= \frac{P(\text{smoke} | \text{fire}) * P(\text{fire})}{P(\text{smoke} | \text{tamper})} \\
 &= \frac{0.9 * 0.01}{P(\text{smoke} | \text{tamper})}
 \end{aligned}$$



Example

$$P(\text{smoke}|\text{tamper}) =$$

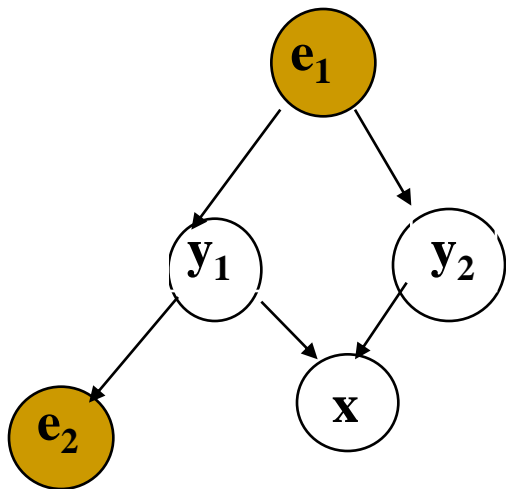
$$\begin{aligned}
 & P(\text{smoke}|\text{tamper, fire}) * P(\text{fire}|\text{tamper}) + \\
 & P(\text{smoke}|\text{tamper, } \sim\text{fire}) * P(\sim\text{fire}|\text{tamper}) \\
 = & P(\text{smoke}|\text{fire}) * P(\text{fire}) + P(\text{smoke}|\sim\text{fire}) * (1 - P(\text{fire})) \\
 = & 0.9 * 0.01 + 0.01 * 0.99 \\
 = & 0.0189
 \end{aligned}$$

$$P(\text{fire}|\text{tamper, smoke}) = 0.9 * 0.01 / 0.0189 = 0.476$$

$$P(\text{alarm} | \text{smoke}) = \dots$$

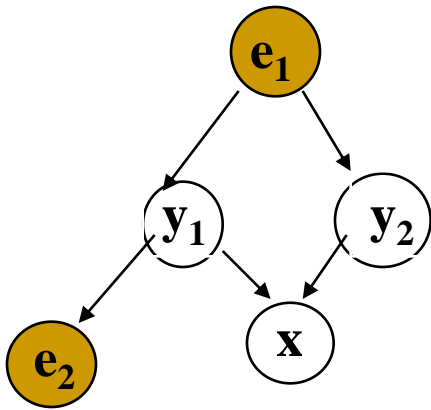
$$P(\text{leaving} | \text{smoke}) = \dots$$

Another Example



$$P(x \mid e_1, e_2) = ?$$

Recursive Estimation



$$\begin{aligned}
 P(\mathbf{x} | \mathbf{e}_1, \mathbf{e}_2) &= \sum_{y_1, y_2} P(\mathbf{x} | y_1, y_2) P(y_1, y_2 | \mathbf{e}_1, \mathbf{e}_2) \\
 &= \sum_{y_1, y_2} P(\mathbf{x} | y_1, y_2) P(y_1 | y_2, \mathbf{e}_1, \mathbf{e}_2) P(y_2 | \mathbf{e}_1, \mathbf{e}_2) \\
 &= \sum_{y_1, y_2} P(\mathbf{x} | y_1, y_2) P(y_1 | \mathbf{e}_1, \mathbf{e}_2) P(y_2 | \mathbf{e}_1) \\
 &= \sum_{y_1, y_2} P(\mathbf{x} | y_1, y_2) \frac{P(\mathbf{e}_2 | y_1, \mathbf{e}_1) P(y_1 | \mathbf{e}_1)}{P(\mathbf{e}_2 | \mathbf{e}_1)} P(y_2 | \mathbf{e}_1) \\
 &= \sum_{y_1, y_2} P(\mathbf{x} | y_1, y_2) \frac{P(\mathbf{e}_2 | y_1) P(y_1 | \mathbf{e}_1)}{\sum_{y_1} P(\mathbf{e}_2 | y_1) P(y_1 | \mathbf{e}_1)} P(y_2 | \mathbf{e}_1)
 \end{aligned}$$

Tradeoff of FJPDT vs. BB Network

1. Recall that the FJPDT can be used to answer any query about the domain. Since any probability from the joint can be calculated from a belief network, we can conclude that any query about the domain can be answered using a belief network.
2. The tradeoff is that by keeping the joint we have to estimate and save a very large number of probabilities. A belief network can be much more concise, but you need to calculate, rather than look up in a table, values for the joint.
3. E.g. Assuming Boolean variables, with k parents of each node, and n nodes (which implies n CPTs), the complete network can be specified with $n \cdot 2^k$ probabilities. The joint requires 2^n probabilities. If $n = 20$ and $k = 5$, then the belief network requires 640 numbers, whereas the full joint requires over a million.

Review: Bayesian Nets

- **Bayesian Nets are the bread and butter of AI-uncertainty community** (like resolution to AI-logic)
- **Bayesian Nets are a compact representation**
 - Don't require exponential storage to hold all of the probs.
 - In the full joint probability distribution (FJPD) table are a decomposed representation of the FJPD table
 - Conditional prob. dist. tables in non-root nodes are only exponential in the max number of parents of any node
- **Bayesian Nets are fast at computing joint probs:** $P(V_1, \dots, V_k)$ i.e. prior probability of V_1, \dots, V_k
 - Computing the probability of an atomic event can be done in linear time with the number of nodes in the net
- **Conditional probabilities can be computed:**
 $P(Q|E_1, \dots, E_k)$
cond. prob. of query Q given evidence E_1, \dots, E_k
 - Requires enumerating all of the relevant joint probabilities, which takes *exponential* time in the number of variables

Review: Conditional Probability

- $P(A|B)$ = the *conditional (or posterior) probability* of A given that all we know is B.

$$P(A | B) = \frac{P(A \wedge B)}{P(B)}, P(B) > 0$$

- ◆ Once we receive some evidence concerning a proposition, prior probabilities are no longer applicable.
- ◆ We need to assess the conditional probability of that proposition given that what we know is the available evidence.

Review: Chain Rule

- *Chain rule* is derived by successive applications of the product rule:

$$\begin{aligned} \mathbf{P}(X_1, \dots, X_n) &= \mathbf{P}(X_1, \dots, X_{n-1}) \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \mathbf{P}(X_1, \dots, X_{n-2}) \mathbf{P}(X_{n-1} | X_1, \dots, X_{n-2}) \mathbf{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \dots \\ &= \mathbf{P}(X_1) \prod_{i=2}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

Review: Conditional independence

- **X and Y are independent random variables if:**
 - $P(X,Y) = P(X)P(Y)$ or equivalently
 - $P(X|Y) = P(X)$
- X is conditionally independent of Y given Z if
 - $P(X|Y,Z) = P(X|Z)$ or equivalently
 - $P(X,Y|Z) = P(X|Z) P(Y|Z)$
 - $P(X,Y|Z) = P(X|Y,Z)P(Y|Z)$
- The **product rule** is an alternative formulation of conditional probability:

$$P(A\&B) = P(A|B) P(B) = P(B|A) P(A)$$

- **Bayes' Rule**

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