Chapter 14: Bayesian Networks

These slides are adopted from Berkeley course materials and Russell and Norvig textbook

Independence

- Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form: $\forall x, y : P(x|y) = P(x)$
  - We write:

$$X \perp\!\!\!\perp Y$$

- Independence is a simplifying *modeling assumption.*

Example: Independence

Assume we have two random variables Temperature (hot/cold) and weather (Sun/Rain). Are T and W independent?

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

P(T)P(W)

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

Example: Independence

N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

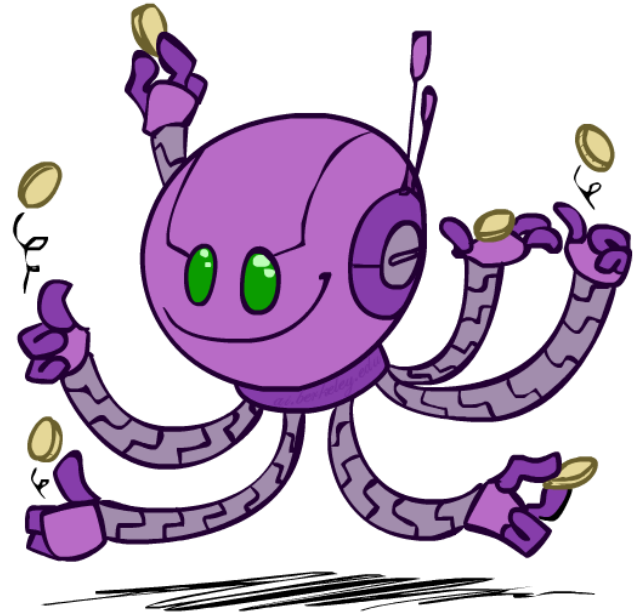
$P(X_2)$

H	0.5
T	0.5

$\ldots$

$P(X_n)$

H	0.5
T	0.5



$P(X_1, X_2, \ldots X_n)$

2^n

Conditional Independence

- Unconditional (absolute) independence very rare (why?)

- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.

$$X \perp\!\!\!\perp Y | Z$$

- X is conditionally independent of Y given Z

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

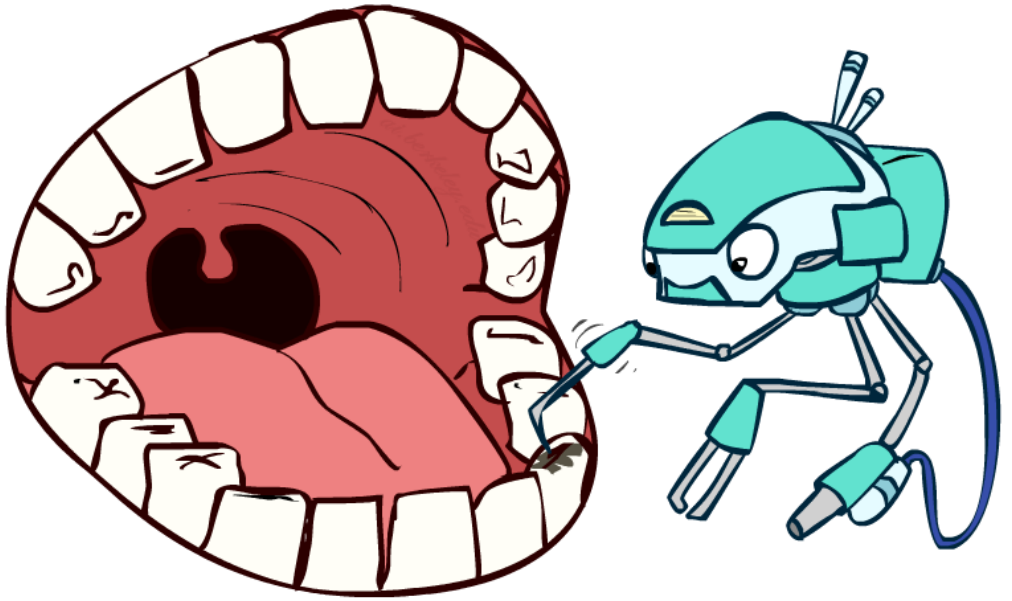
if and only if:

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

- or, equivalently, if and only if

Conditional Independence

P(Toothache, Cavity, Catch)



If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

- P(+catch | +toothache, +cavity) = P(+catch | +cavity)

The same independence holds if I don't have a cavity:

- P(+catch | +toothache, -cavity) = P(+catch| -cavity)

Catch is *conditionally independent* of Toothache given Cavity:

- P(Catch | Toothache, Cavity) = P(Catch | Cavity)
  - Equivalent statements:
    - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
    - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
    - One can be derived from the other easily

Conditional Independence

What about this domain:

- Traffic (T)
- Umbrella (U)
- Raining (R)

- $T \perp\!\!\!\perp U \mid R$

Probabilistic Models

Models describe how (a portion of) the world works

Models are always simplifications

- May not account for every variable
- May not account for all interactions between variables

What do we do with probabilistic models?

- We (or our agents) need to reason about unknown variables, given evidence
- Example: explanation (diagnostic reasoning)
- Example: prediction (causal reasoning)

Bayesian Networks

- The full joint probability distribution can answer any question about the domain, but this can become intractably large as the number of variables grows
- **Bayesian networks** represent the dependencies among variables.
- Bayesian networks can represent essentially *any* full joint probability distribution.

Bayesian Networks

- A Bayesian network is a directed acyclic graph (DAG) in which each node is annotated with quantitative probability information.

- The full specification is as follows:

  1. Each node corresponds to a random variable, which may be discrete or continuous.

  2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node

  X to node Y , X is said to be a *parent* of Y.

  3. Each node Xi has a conditional probability distribution **P**(Xi |Parents(Xi)) that quantifies

  the effect of the parents on the node.

Bayesian Networks

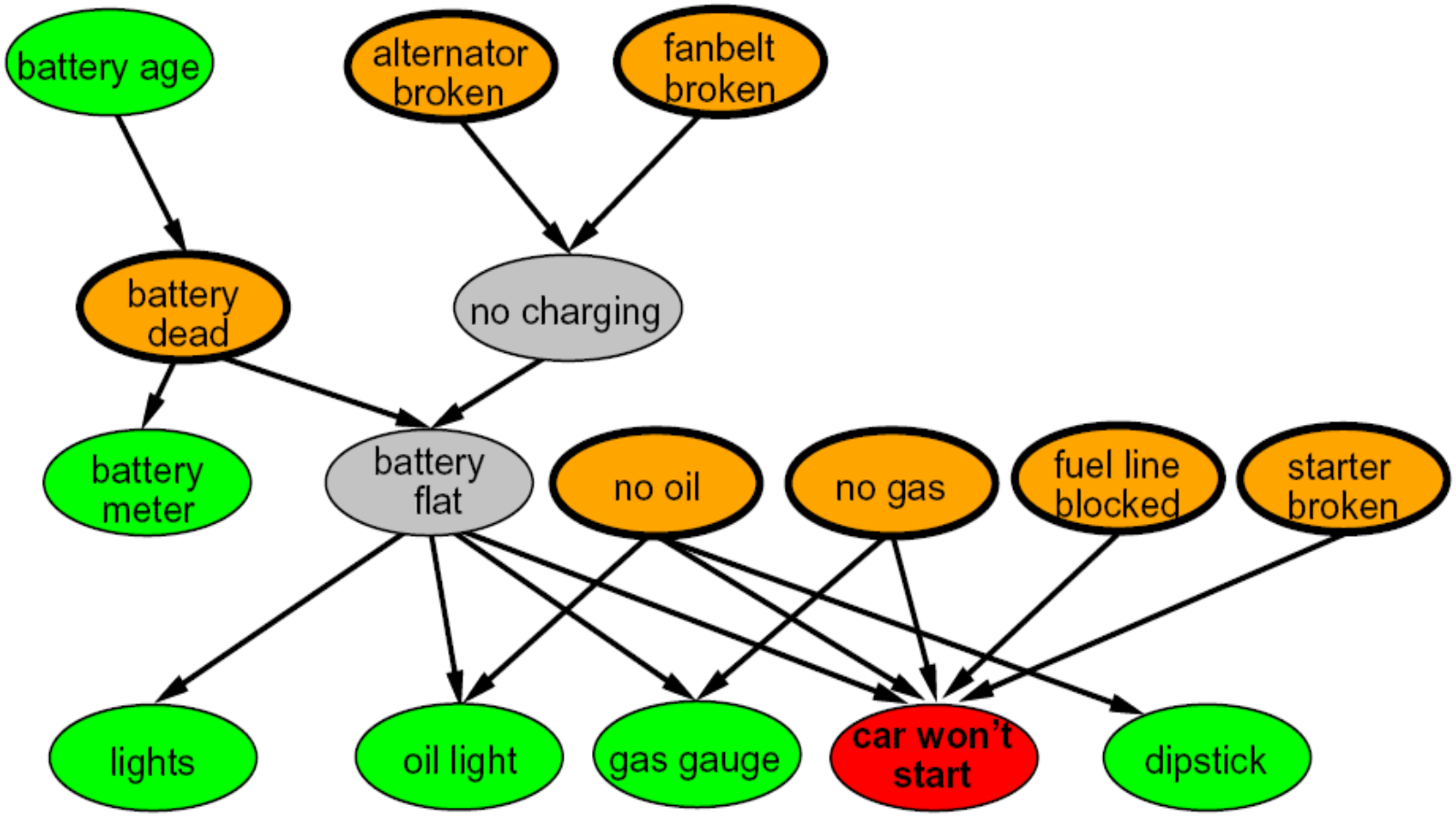
- The topology of the network—the set of nodes and links—specifies the conditional independence relationships that hold in the domain, in a way that will be made precise shortly.

- The *intuitive* meaning of an arrow is typically that X has a *direct influence* on Y, which suggests

  that causes should be parents of effects.

Given the Bayesian network topology and the conditional probability tables, the full joint distribution for all the variables can be specified.

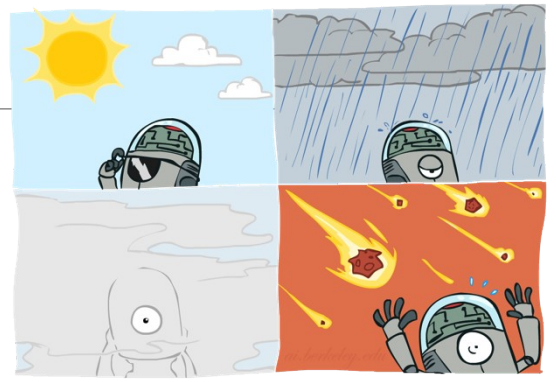
Example Bayesian Network: Car



Graphical Model Notation

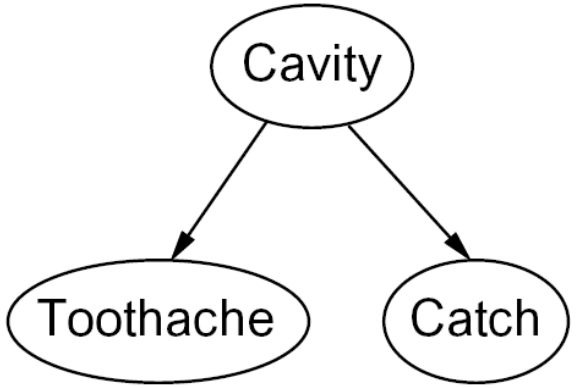
Nodes: variables (with domains)

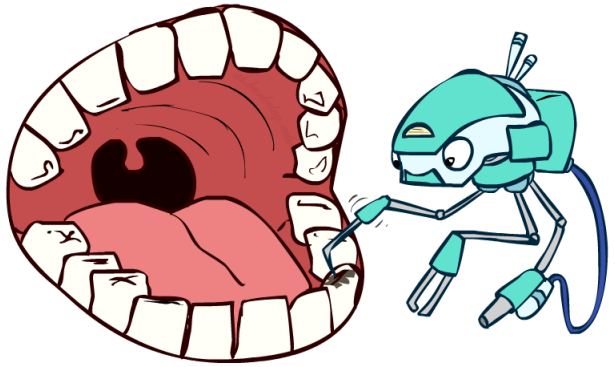
- Can be assigned (observed) or unassigned (unobserved).



Arcs: interactions

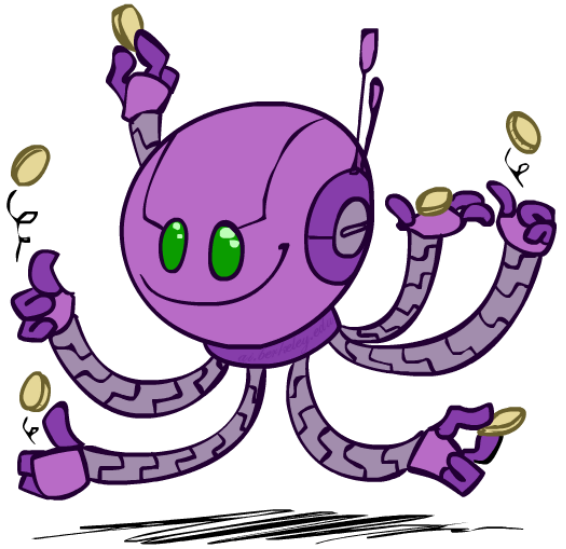
- Similar to CSP constraints
- Indicate “direct influence” between variables
- Formally: encode conditional independence





Example: Coin Flips

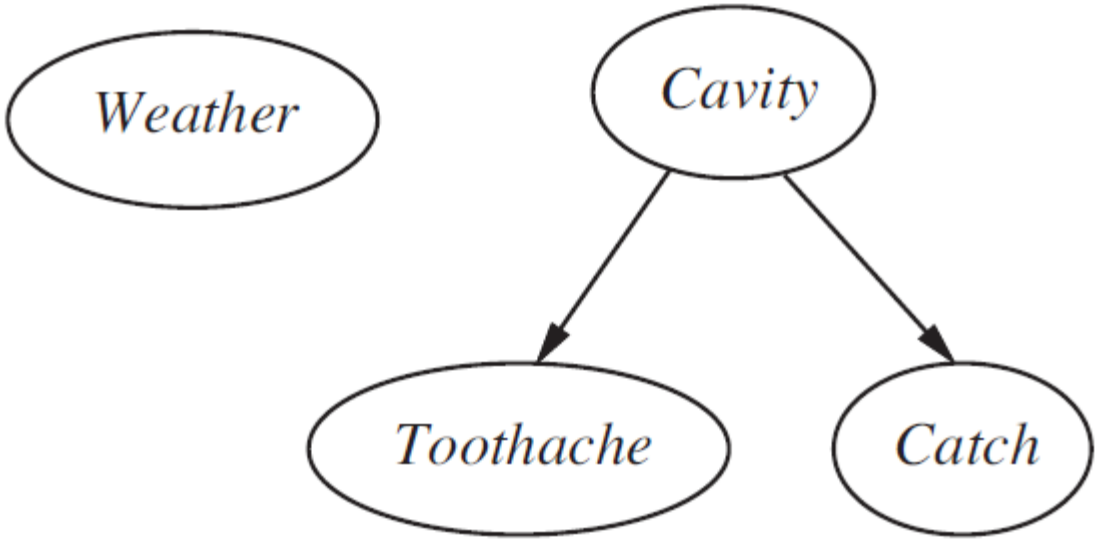
N independent coin flips



No interactions between variables: absolute independence

Bayesian Network Example

- Toothache and Catch are conditionally independent given Cavity.
- Weather is independent of the other variables.



Example: Burglary Network

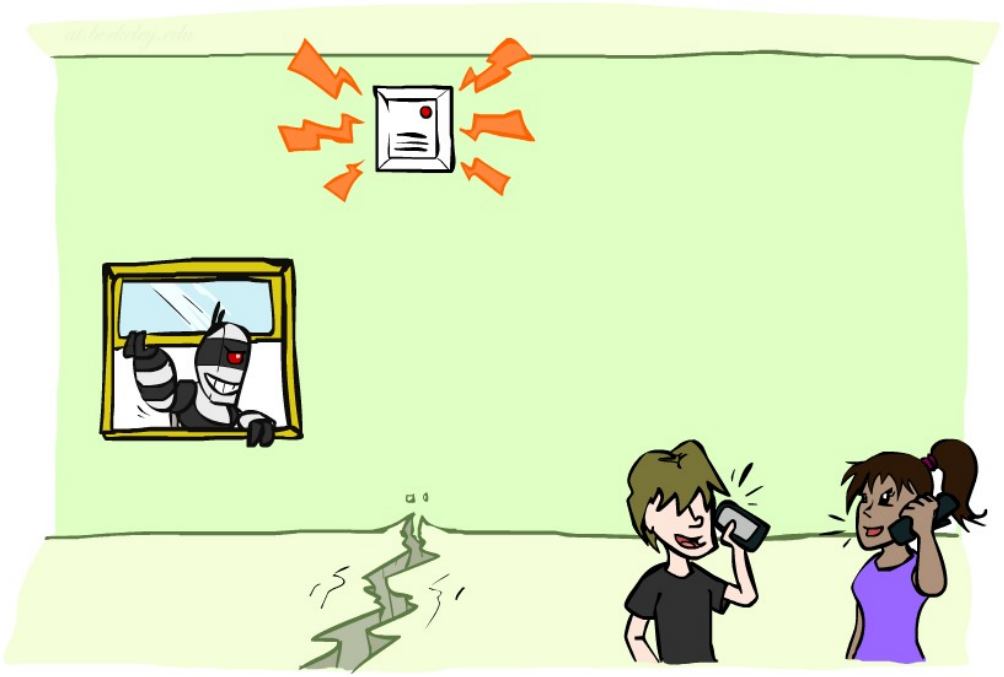
- You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.
- You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm.
- John nearly always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
- Mary, on the other hand, likes rather loud music and often misses the alarm altogether. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Draw the corresponding Bayesian Network.

Example: Burglary Network

Variables

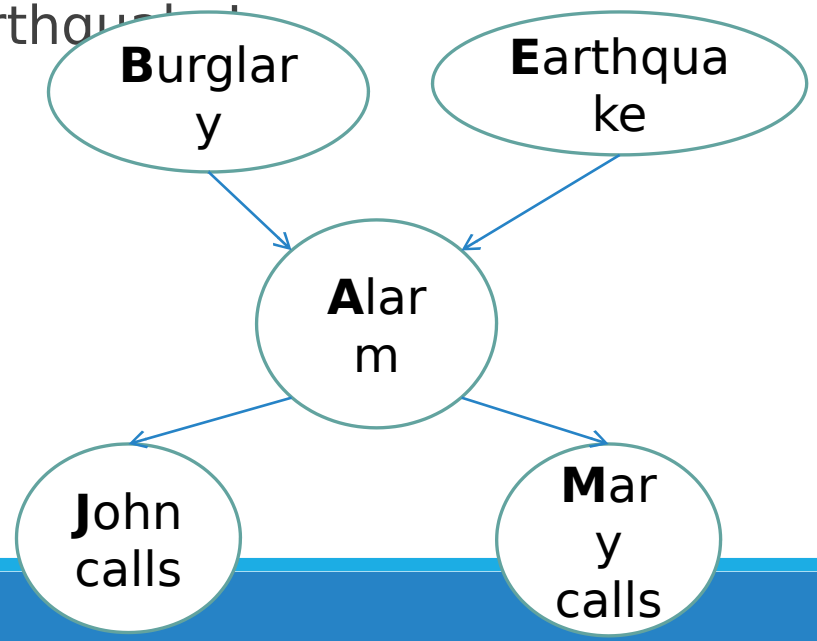
- B: Burglary
- A: Alarm goes off
- M: Mary calls
- J: John calls
- E: Earthquake!

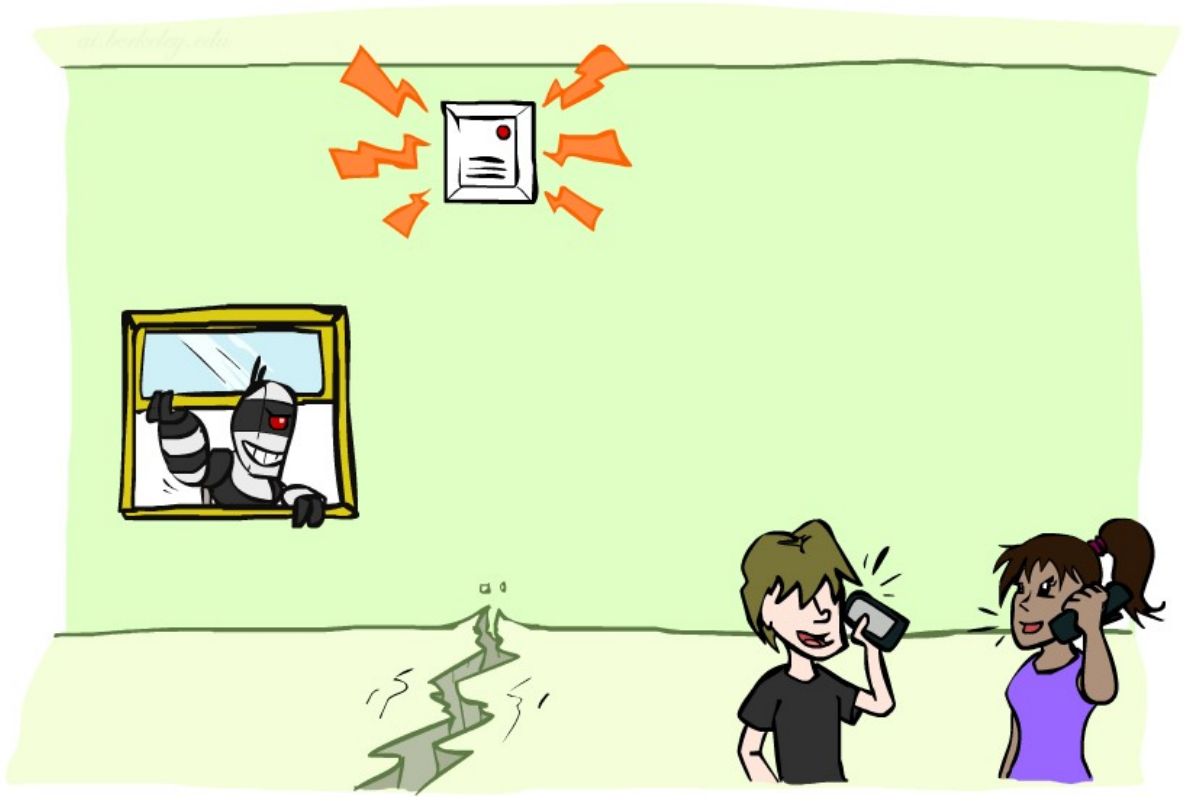


Example: Burglary Network

Variables

- B: Burglary
- A: Alarm goes off
- M: Mary calls
- J: John calls
- E: Earthqu



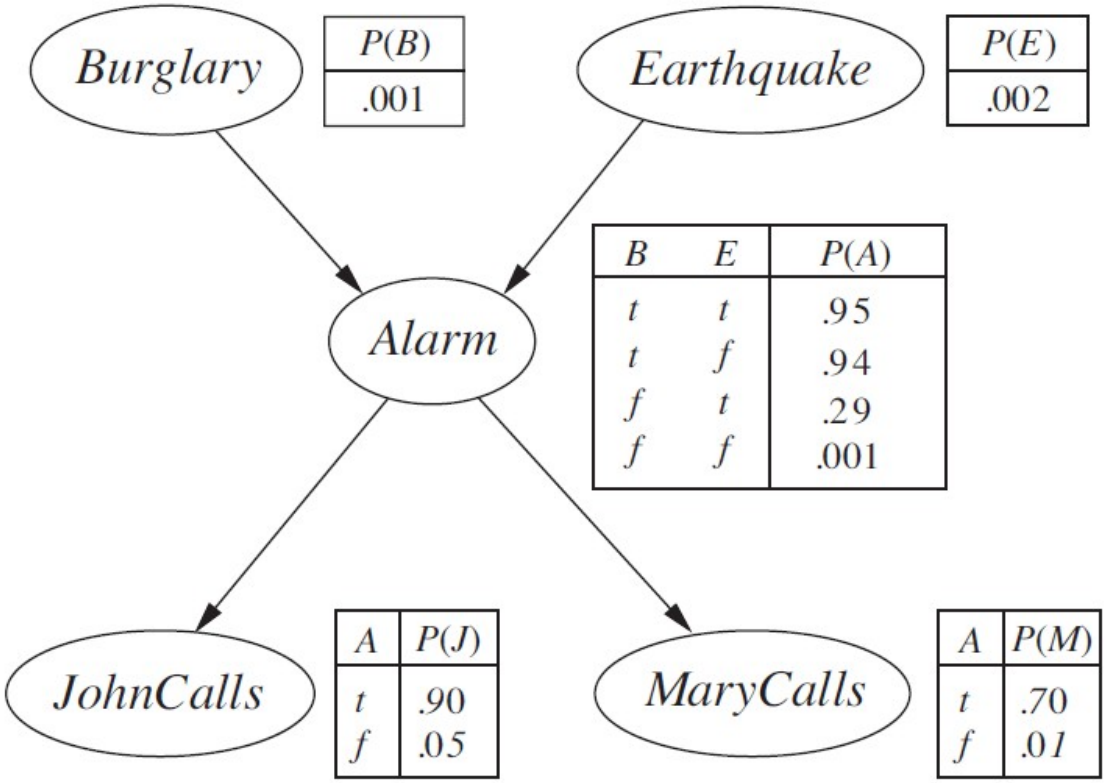


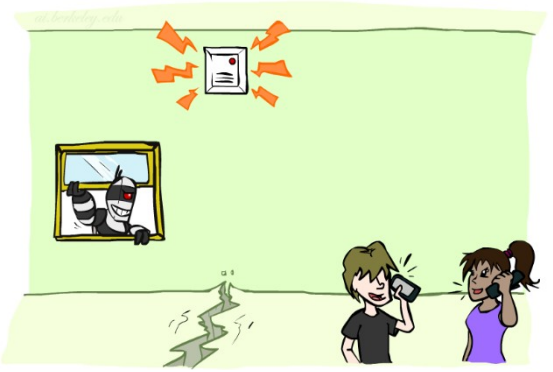
Example: Burglary Network

- The network structure shows that burglary and earthquakes directly affect the probability of the alarm's going off, but whether John and Mary call depends only on the alarm.

- The network thus represents our assumptions that they do not perceive burglaries directly, they do not notice minor earthquakes, and they do not confer before calling.

Example: Burglary Network



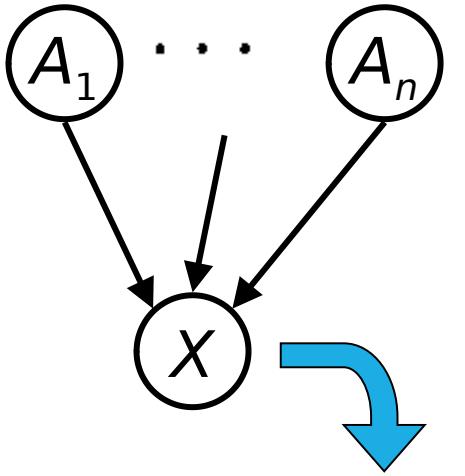


Conditional Probability Table (CPT)

- Each row in a CPT contains the conditional probability of each node value for a **conditioning case**.
- A conditioning case is just a possible combination of values for the parent nodes.
- Each row must sum to 1, because the entries represent an exhaustive set of cases for the variable.
- For Boolean variables, once you know that the probability of a true value is p, the probability of false must be 1 – p, so we the second number is often omitted.
- A table for a Boolean variable with k Boolean parents contains 2^k independently specifiable probabilities.
- A node with no parents has only one row, representing the prior probabilities of each possible value of the variable.

Bayesian Network Semantics

- Bayesian network is a directed acyclic graph.
- It consists of a set of nodes, one per variable X
- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents' values $P(X|a_1 \dots a_n)$
  - CPT: conditional probability table
  - Description of a noisy "causal" process



A Bayes network = Topology (graph) + Local Conditional Probabilities

Joint Probability Distribution

- The joint probability distribution for any N variables $X_1, X_2, \ldots X_N$:

$$P(x_1, \ldots, x_n) = P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1}, \ldots, x_1)$$

- This can be written as:

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1) \,. \end{aligned}$$

- The rule above is named the chain rule.
- The chain rule holds for any set of random variables.

Bayesian Network Semantics

- What do Bayesian networks mean?
- Bayesian networks is a representation of joint distributions.
- The chain rule which hold for any set of variables state that:

$$\begin{aligned} P(x_1, \ldots, x_n) &= P(x_n \,|\, x_{n-1}, \ldots, x_1) P(x_{n-1} \,|\, x_{n-2}, \ldots, x_1) \cdots P(x_2 \,|\, x_1) P(x_1) \\ &= \prod_{i=1}^{n} P(x_i \,|\, x_{i-1}, \ldots, x_1) \,. \end{aligned}$$

- However, for Bayesian networks, using the conditional independence:

$$\mathbf{P}(X_i \,|\, X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \,|\, Parents(X_i))$$

Bayesian Network Semantics

- Bayesian networks implicitly encode joint distributions.
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

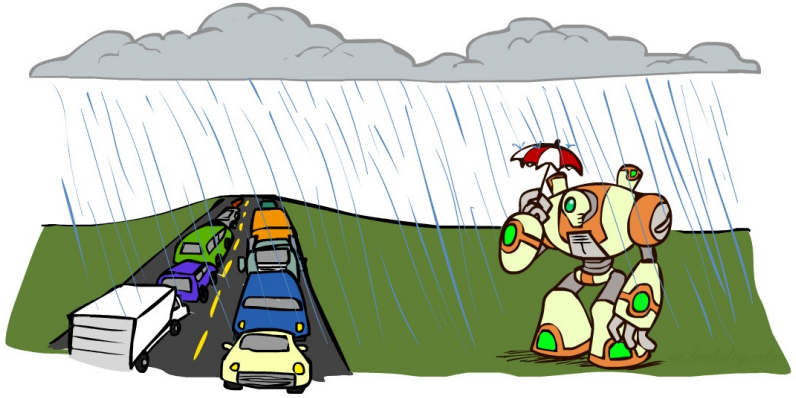
Conditional Independence and the Chain Rule

Chain rule:

$$P(X_1, X_2, \dots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\dots$$

Example:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

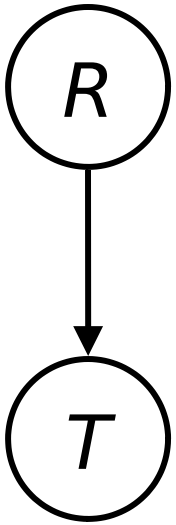


With assumption of conditional independence:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

Bayesian networks / graphical models help us express conditional independence assumptions.

Example: Traffic



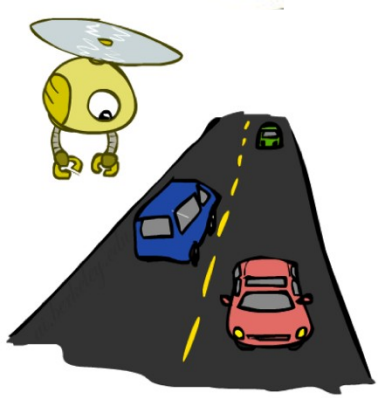
$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

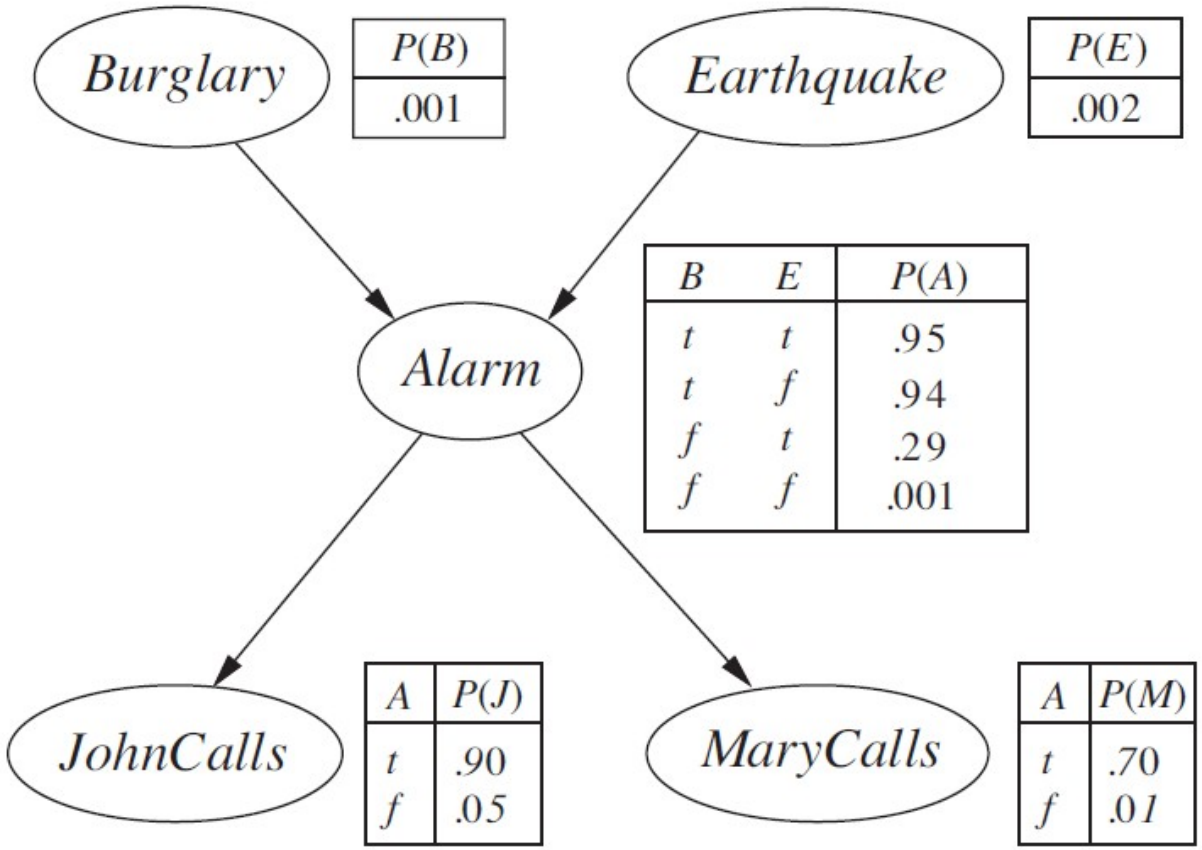
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(+r, -t) =$ P(+r)P(-t|+r) = ¼*1/4



Example: Burglary Network

Given this Bayesian network, calculate P(j,m,a, b, e)



Example: Burglary Network

$$
\begin{aligned}
P(j, m, a, \neg b, \neg e) &= P(j \mid a) P(m \mid a) P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e) \\
&= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628
\end{aligned}
$$

Bayesian Networks Construction

1. *Nodes:*

Determine the set of variables that are required to model the domain, and order them, {X1, . .,Xn}.

Any order will work, but the resulting network will be more compact if the variables are ordered such that causes precede effects.

2. *Links:*

For i = 1 to n do:

• Choose, from X1, . . . ,Xi−1, a minimal set of parents for Xi, such that the following equation is satisfied:

$$\mathbf{P}(X_i \mid X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i \mid Parents(X_i))$$

For each parent insert a link from the parent to Xi.

• CPTs: Write down the conditional probability table, **P**(Xi|Parents(Xi)).

Intuitively, the parents of node Xi should contain all those nodes in X1, . . . , Xi−1 that *directly influence* Xi.

Bayesian Networks Construction

- Because each node is connected only to earlier nodes, this construction method guarantees

that the network is acyclic.

- Bayesian networks contain no redundant probability values, therefore there is no chance for inconsistency between probability values.

Bayesian Networks Compactness

- **Bayesian Networks is locally structured sparse**) system where each subcomponent interacts directly with only a bounded number of other components, regardless of the total number of components.
- Local structure is usually associated with linear rather than exponential growth in complexity.
- For Bayesian networks, assume each random variable is directly influenced by at most k others (has k parents), for some constant k.
- If we assume n Boolean variables, then the amount of information needed to specify each conditional probability table will be at most 2^k numbers, and the complete network can be specified by $n2^k$ numbers.
- In contrast, the joint distribution contains 2^n numbers.
- For example, suppose we have n=30 nodes, each with five parents (k =5).
- Then, the Bayesian network requires 960 numbers, but the full joint distribution requires over a billion

Exact Inference in Bayesian Networks

- Any probabilistic inference system computes the posterior probability distribution for a set of **query variables**, given some observed **event**(some assignment of values to a set of **evidence variables).**

- X denotes the query variable.
- **E** denotes the set of evidence variables E1, . . . ,Em, and **e** is a particular observed event.
- **Y** denotes the hidden variables which are nonevidence, nonquery variables Y1, . . . , Yl.
- Thus, the complete set of variables is **X** ∪**E** ∪**Y**.
- A typical query asks for the posterior probability distribution **P**(X | **e**)

Exact Inference in Bayesian Networks- Example

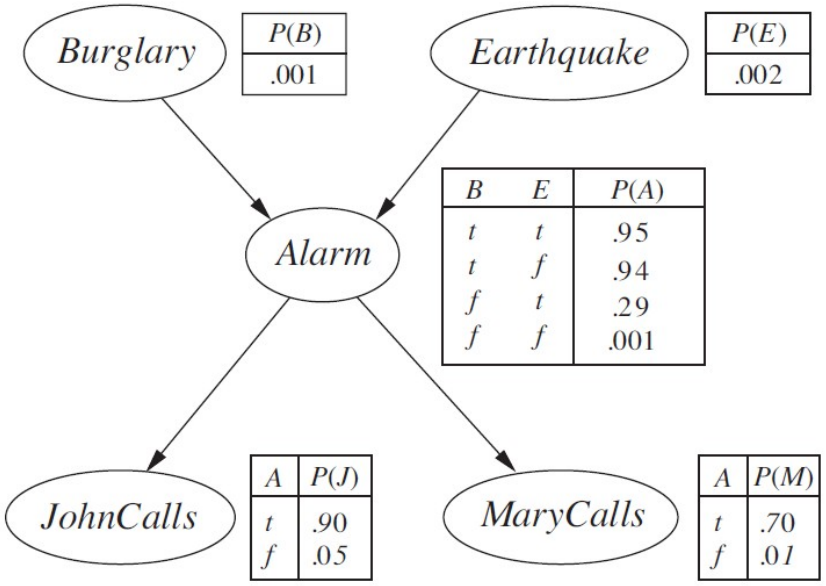
- In the burglary network, we might observe the event in which JohnCalls =true and

MaryCalls =true.

- Calculate the probability that a burglary has

given than Mary calls and John calls.

P(b|j,m)=?



Inference by enumeration

- Any conditional probability can be computed by summing terms from the full joint distribution as follows:

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})$$

(Using Bayes rule and Marginalization)

P(Burglary | JohnCalls =true, MaryCalls =true).

- The hidden variables for this query are Earthquake and Alarm:

$$\mathbf{P}(B \mid j, m) = \alpha\, \mathbf{P}(B, j, m) = \alpha \sum_{e} \sum_{a} \mathbf{P}(B, j, m, e, a,)$$

Inference by enumeration

$$\mathbf{P}(B \mid j, m) = \alpha \mathbf{P}(B, j, m) = \alpha \sum_e \sum_a \mathbf{P}(B, j, m, e, a,)$$

For Burglary=True:

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

- To compute this expression, we have to add four terms, each computed by multiplying five numbers.
- In the worst case, where we have to sum out almost all the variables, the complexity of the algorithm for a network with n Boolean variables is $O(n2^n)$.

Inference by enumeration

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

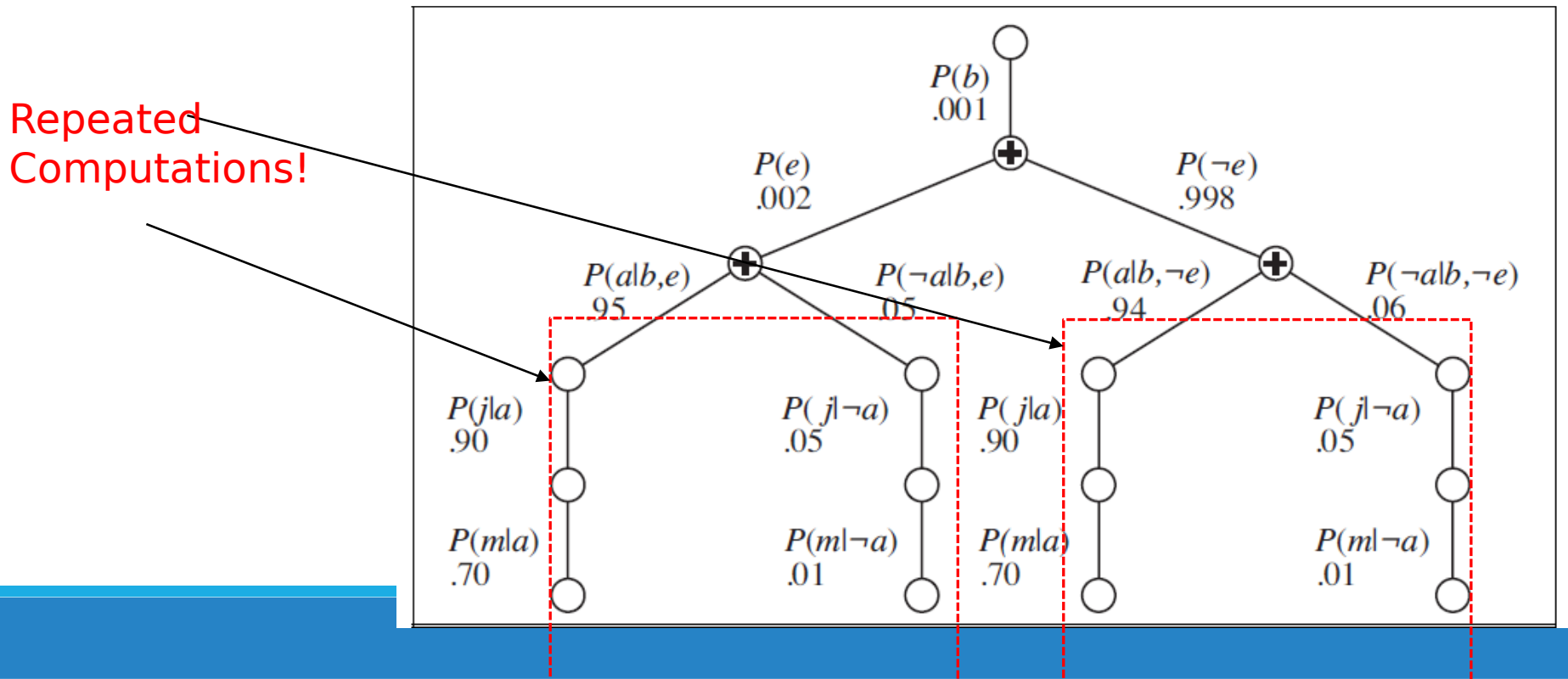
- A simple improvement can de done by taking out P(b) and P(e) out of the summations:

$$P(b \mid j, m) = \alpha\, P(b) \sum_e P(e) \sum_a P(a \mid b, e)P(j \mid a)P(m \mid a)\ .$$

Inference by enumeration

- The structure of computations done by inference by enumeration:

$$P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(j \mid a) P(m \mid a) .$$



Inference by enumeration Example

$$P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(j \mid a) P(m \mid a) .$$

= x0.00059224259

=x0.001491857649

After normalization:

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 $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ /* **Y** = *hidden variables* */

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(bn.VARS, $\mathbf{e}_{x_i}$)
      where $\mathbf{e}_{x_i}$ is **e** extended with $X = x_i$
  return NORMALIZE($\mathbf{Q}(X)$)

---

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($vars$), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is **e** extended with $Y = y$

Figure 14.9 The enumeration algorithm for answering queries on Bayesian networks.

Inference by enumeration

- The ENUMERATION-ASK algorithm evaluates the computation trees using depth-first recursion.
- The algorithm is very similar to the backtracking algorithm for solving CSPs.
- The space complexity of ENUMERATION-ASK is linear in the number of variables.

- However, its time complexity for a network with n Boolean variables is always $O(2^n)$—better than the $O(n\ 2^n)$ for the simple approach described earlier.

- Inference by enumeration repeats evaluating some expressions such as the products $P(j \mid a)P(m \mid a)$ and $P(j \mid \neg a)P(m \mid \neg a)$ are computed twice, once for each value of e.

Sections 14.1., 14.2, and 14.4.