Quantifying uncertainty & Bayesian networks

CE417: Introduction to Artificial Intelligence
Sharif University of Technology
Spring 2016

Soleymani

“Artificial Intelligence: A Modern Approach”, 3rd Edition, Chapter 13 & 14.1-14.2

Outline

- Uncertainty
- Probability & basic notations
- Inference in probability theory
  - Bayes rule
- Baysian networks
  - representing probabilistic knowledge as a Bayesian network

Automated taxi example

- A_t :"leave for airport t minutes before flight"

- "Will A_t get us to the airport on time?"
  - "A_{90} will get us there on time if the car does not break down or run out of gas and there's no accident on the bridge and I don't get into an accident and …"

- Qualification problem

Probability: summarizing uncertainty

- Problems with logic
  - laziness: failure to enumerate exceptions, qualifications, etc.
  - Theoretical or practical ignorance: lack of complete theory or lack of all relevant facts and information

- Probability summarizes the uncertainty

- Probabilities are made w.r.t the current knowledge state (not w.r.t the real world)
  - Probabilities of propositions can change with new evidence
    - e.g., $P(A_{90}$ gets us there on time$) = 0.7$

      $P(A_{90}$ gets us there on time $|$ time$= 5$ p.m.$) = 0.6$

Automated taxi example: rational decision

- A_{90} can be a rational decision depending on performance measure
  - Performance measure: getting to the airport in time, avoiding a long wait in the airport, avoiding speeding tickets along the way, …

- Maximizing expected utility
  - Decision theory = probability theory + utility theory
    - Utility theory: assigning a degree of usefulness for each state and preferring states with higher utility

Basic probability notations (sets)

- Sample space (Ω): set of all possible worlds

- Probability model
  - $0 \leq P(\omega) \leq 1$ $\quad(\forall \omega \in \Omega)$ ω is an atomic event
  - $P(\Omega) = 1$
  - $P(A) = \sum_{\omega \in A} P(\omega)$ $\quad(\forall A \subseteq \Omega)$ A is a set of possible worlds

Basic probability notations (propositions)

- S: set of all propositional sentences

- Probability model
  - $0 \leq P(a) \leq 1 \qquad (\forall a \in S)$
  - If T is a tautology $P(T) = 1$
  - $P(\phi) = \sum_{\omega \in \Omega:\ \omega \vDash \phi} P(\omega) \quad (\forall \phi \in S)$

Basic probability notations (propositions)

- Each proposition corresponds to a set of possible worlds
  - A possible world is defined as an assignment of values to all random variables under consideration.

- Elementary proposition
  - e.g., $Dice1 = 4$

- Complex propositions
  - e.g., $Dice1 = 4 \lor Dice2 = 6$

Random variables

- Random variables: Variables in probability theory
- Domain of random variables: **Boolean**, **discrete** or **continuous**
  - Boolean
    - e.g., The domain of $Cavity$: $\{true, false\}$
      - $Cavity = true$ is abbreviated as $cavity$
  - Discrete
    - e.g., The domain of $Weather$ is $\{sunny, rainy, cloudy, snow\}$
  - Continuous
- Probability distribution: the function describing probabilities of possible values of a random variable
  - $P(Weather = sunny) = 0.6,\ P(Weather = rain) = 0.1, \dots$

Probabilistic inference

- Joint probability distribution
  - specifies probability of every atomic event
- Prior and posterior probabilities
  - belief in the absence or presence of evidences
- Bayes' rule
  - used when we don't have $P(a|b)$ but we have $P(b|a)$
- Independence

Probabilistic inference

- If we consider **full joint distribution** as KB
  - Every query can be answered by it

- **Probabilistic inference**: Computing **posterior** distribution of variables given evidence
  - An agent needs to make decisions based on the obtained evidence

Joint probability distribution

- Joint probability distribution
  - Probability of all combinations of the values for a set of random vars.

$\mathbf{P}(Weather, Cavity)$: joint probability as a 4 × 2 matrix of values

	Weather= sunny	Weather= rainy	Weather= cloudy	Weather= snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

- Queries can be answered by summing over atomic events

Prior and posterior probabilities

- Prior or unconditional probabilities of propositions: belief in the absence of any other evidence
  - e.g., $P(cavity) = 0.2$
    $P(Weather = sunny) = 0.72$

- Posterior or conditional probabilities: belief in the presence of evidences
  - e.g., $P(cavity \mid toothache)$

$$P(a|b) = P(a \wedge b)/P(b) \quad \text{if } P(b) > 0$$

Sum rule and product rule

- Product rule (obtained from the formulation of the conditional probability):

$$P(a, b) = P(a|b)\,P(b) = P(b|a)\,P(a)$$

- Sume rule:

$$P(a) = \sum_{b} P(a, b)$$

Inference: example

- Joint probability distribution:

	$tootache$		$\neg tootache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- Conditional probabilities:

$$\mathbf{P}(\boldsymbol{y}|\mathbf{e}) = \frac{\sum_{z} \mathbf{P}(\mathbf{z}, \mathbf{e}, \mathbf{y})}{\sum_{z} \sum_{y} \mathbf{P}(\mathbf{z}, \mathbf{e}, \mathbf{y})}$$

$P(\neg cavity \mid toothache)$

$$= \frac{P(\neg cavity \land toothache)}{P(toothache)}$$

$$= \frac{\sum_{catch=T,F} P(\neg cavity \land toothache, catch)}{\sum_{catch=T,F} \sum_{cavity=T,F} P(cavity \land toothache, catch)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Bayes' rule

- Bayes' rule: $P(a|b) = P(b|a)\,P(a) \,/\, P(b)$
  - $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$
- In many problems, it may be difficult to compute $P(a|b)$ directly, yet we might have information about $P(b|a)$.
- Computing **diagnostic** probability from **causal** probability:

$$P(Cause|Effect) = P(Effect|Cause)\,P(Cause) \,/\, P(Effect)$$

Bayes' rule: example

- Meningitis(M) & Stiff neck (S)
  - $P(m) = \frac{1}{5000}$
  - $P(s) = 0.01$
  - $P(s|m) = 0.7$

- $P(m|s) =?$

$$P(m|s) = P(s|m)P(m)/P(s) \ = \ 0.7 \times 0.0002/0.01 \ = \ 0.0014$$

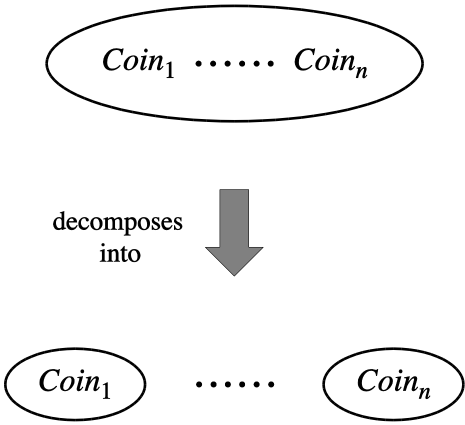
Independence

- Propositions a and b are independent iff

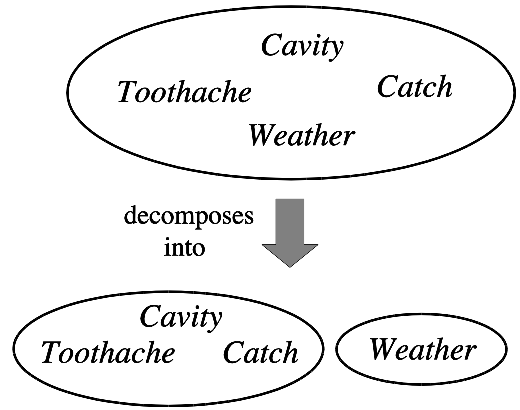
$$P(a|b) = P(a)$$
$$P(b|a) = P(b)$$

$$P(a, b) = P(a)\,P(b)$$

- n independent biased coins
  - The number of required independent probabilities is reduced from $2^n - 1$ to n



Independence



$$P(toothache, catch, cavity, cloudy)$$
$$= P(toothache, catch, cavity)\, P(cloudy)$$

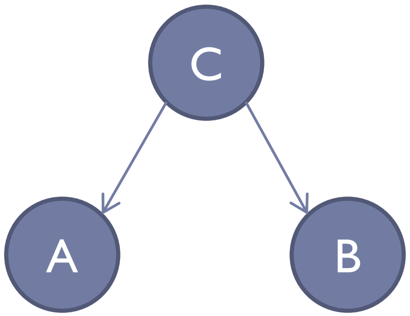
The number of required independent probabilities is reduced from $21 = (2^3 - 1) \times (4 - 1)$ to $10 = (2^3 - 1) + (4 - 1)$

Independent and conditional independent

- Independent random variables:

$$P(a, b) = P(a)P(b)$$

- Conditionally independent random variables: A and B are only <u>conditionally independent given C</u>

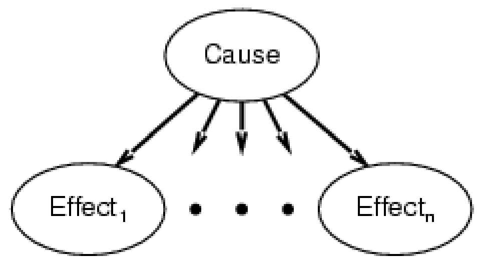


$$P(a, b|c) = P(a|c)P(b|c)$$

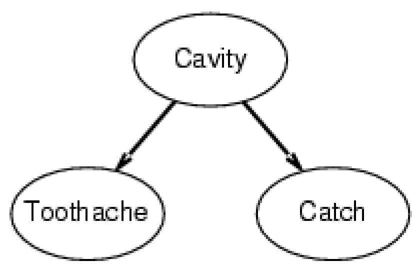
Conditional independence

Naïve Bayes model:

$$\mathbf{P}(Cause, Effect_1, \dots, Effect_n) = \mathbf{P}(Cause) \prod_{i=1}^{n} \mathbf{P}(Effect_i | Cause)$$

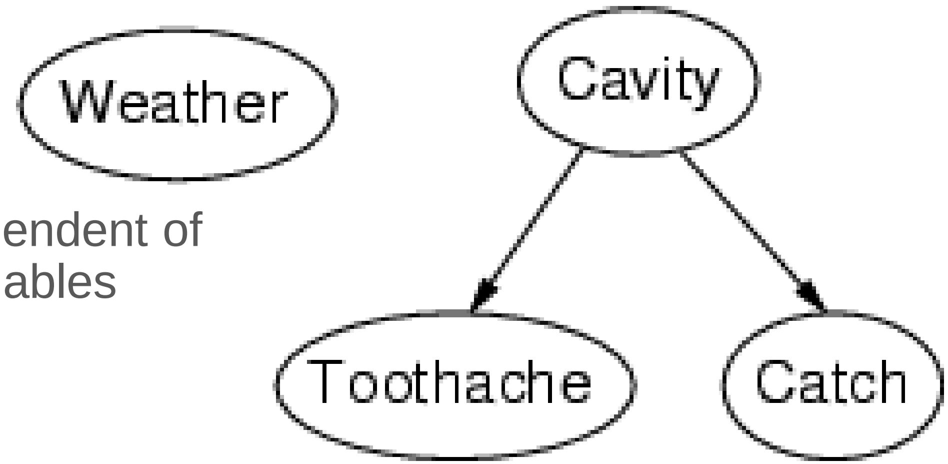


- Example



Cavity example

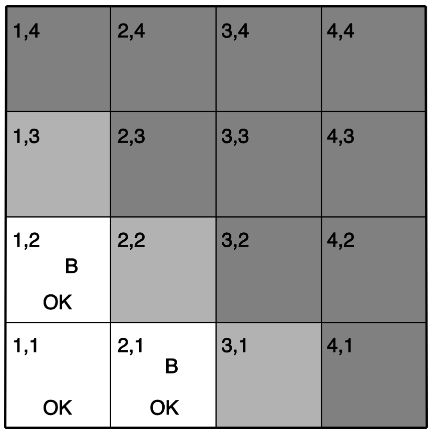
- Topology of network encodes conditional independencies

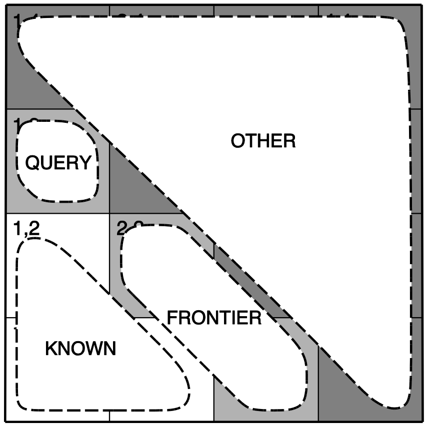


$Weather$ is independent of the other variables

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

Wumpus example



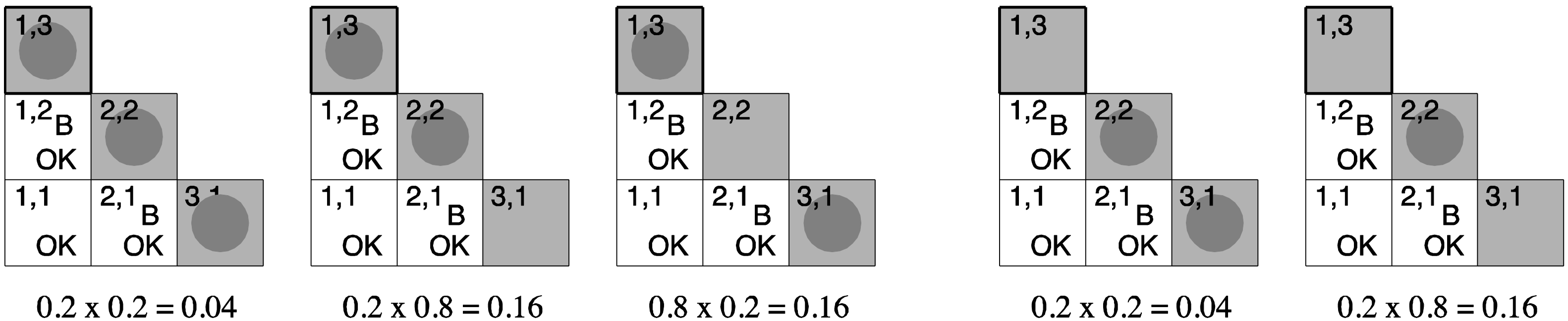


$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

$known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

$P\left(P_{1,3} \middle| known, b\right) = ?$

Wumpus example



(a) (b)

Possible worlds with $P_{1,3} = true$ Possible worlds with $P_{1,3} = false$

$$P(P_{1,3} = True \mid known, b) \propto 0.2 \times [0.2 \times 0.2 + 0.2 \times 0.8 + 0.8 \times 0.2]$$
$$P(P_{1,3} = False \mid known, b) \propto 0.8 \times [0.2 \times 0.2 + 0.2 \times 0.8]$$
$$\Rightarrow P(P_{1,3} = True \mid known, b) = 0.31$$

Bayesian networks

- Importance of independence and conditional independence relationships (to simplify representation)

- Bayesian network: a graphical model to represent dependencies among variables
  - compact specification of full joint distributions
  - easier for human to understand

- **Bayesian network** is a directed acyclic graph
  - Each node shows a random variable
  - Each link from X to Y shows a "direct influence“ of X on Y (X is a parent of Y)
  - For each node, a conditional probability distribution $\mathbf{P}(X_i|Parents(X_i))$ shows the effects of parents on the node

Burglary example

"A **burglar alarm**, respond occasionally to minor **earthquakes**.

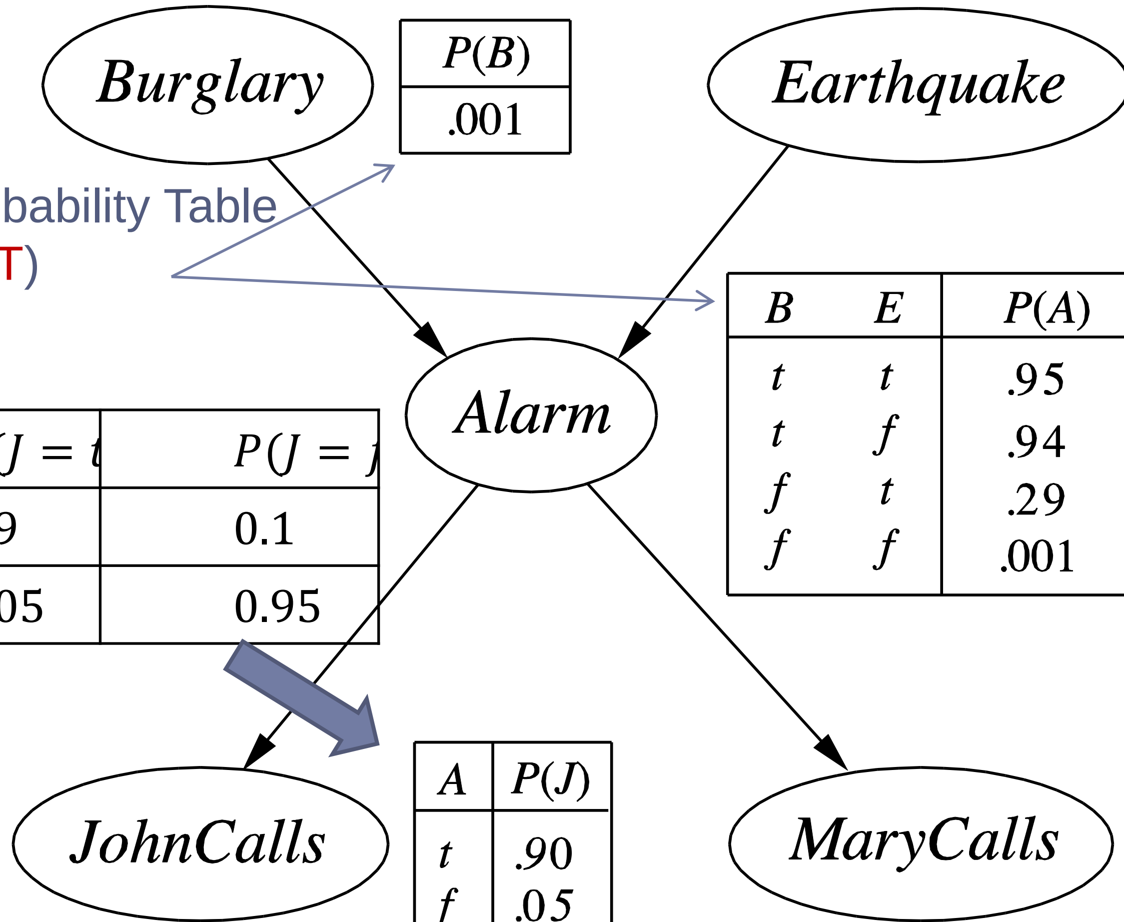
Neighbors **John** and **Mary** call you when hearing the alarm.

John nearly always calls when hearing the alarm.

Mary often misses the alarm."

- Variables:
  - $Burglary$
  - $Earthquake$
  - $Alarm$
  - $JohnCalls$
  - $MaryCalls$

Burglary example



$P(B)$
.001

$P(E)$
.002

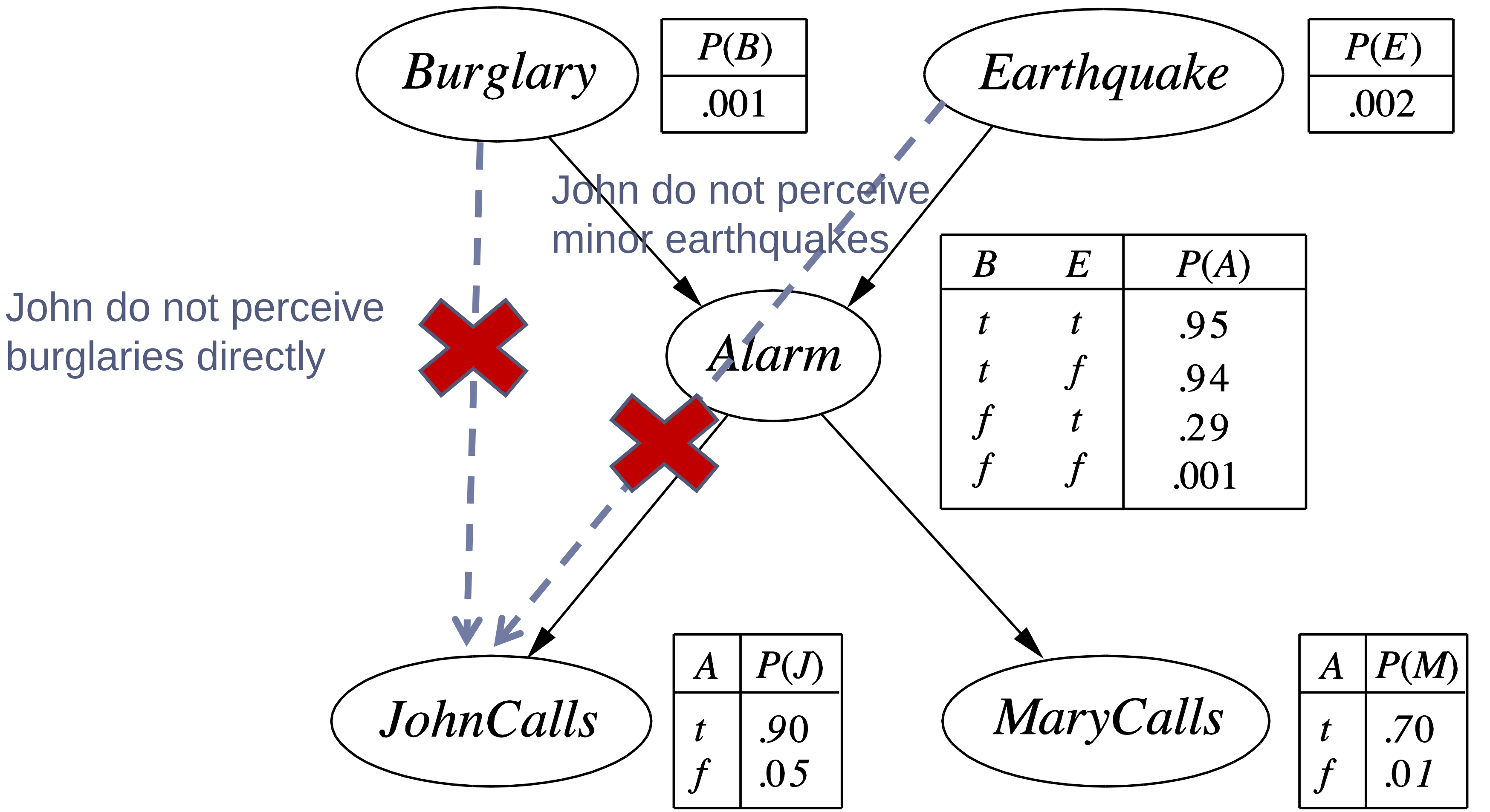
B	E	$P(A)$
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	$P(J = t$	$P(J = f$
t	0.9	0.1
f	0.05	0.95

A	$P(J)$
t	.90
f	.05

A	$P(M)$
t	.70
f	.01

Burglary example



Semantics of Bayesian networks

- The full joint distribution can be defined as the product of the local conditional distributions (using chain rule):

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$$

- Chain rule is derived by successive application of product rule:

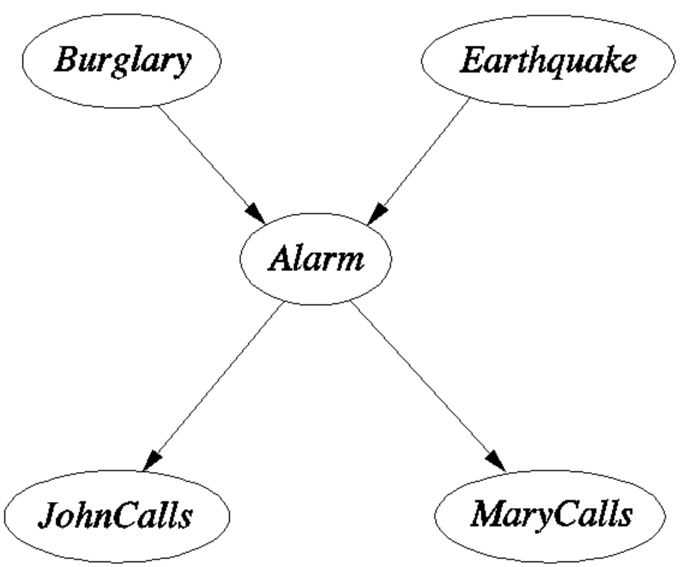
$$\begin{aligned}
&\boldsymbol{P}(X_1, \ldots, X_n) \\
&= \boldsymbol{P}(X_1, \ldots, X_{n-1})\, \boldsymbol{P}(X_n | X_1, \ldots, X_{n-1}) \\
&= \boldsymbol{P}(X_1, \ldots, X_{n-2})\, \boldsymbol{P}(X_{n-1} | X_1, \ldots, X_{n-2})\, \boldsymbol{P}(X_n | X_1, \ldots, X_{n-1}) \\
&= \ldots
\end{aligned}$$

$$= P(X_1) \prod_{i=2}^{n} \boldsymbol{P}(X_i | X_1, \ldots, X_{i-1})$$

Burglary example (joint probability)

We can compute joint probabilities from CPTs:

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$
$$= \; P\,(j \mid a)\; P\,(m \mid a)\; P\,(a \mid \neg b \wedge \; \neg e)\; P\,(\neg b)\; P\,(\neg e)$$
$$= \; 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.000628$$



Burglary example

- "John calls to say my alarm is ringing, but Mary doesn't call. Is there a burglar?"
  - $P(b|j \wedge \neg m)$?

- This conditional probability can be computed from the joint probabilities as discussed earlier:

$$\mathbf{P}(\boldsymbol{y}|\mathbf{e}) = \frac{\sum_{\boldsymbol{z}} \mathbf{P}(\boldsymbol{z}, \mathbf{e}, \mathbf{y})}{\sum_{\boldsymbol{z}} \sum_{\boldsymbol{y}} \mathbf{P}(\boldsymbol{z}, \mathbf{e}, \mathbf{y})}$$

$Y \cup Z \cup E = X$

X shows the set of random variables in the Bayesian network

Compactness

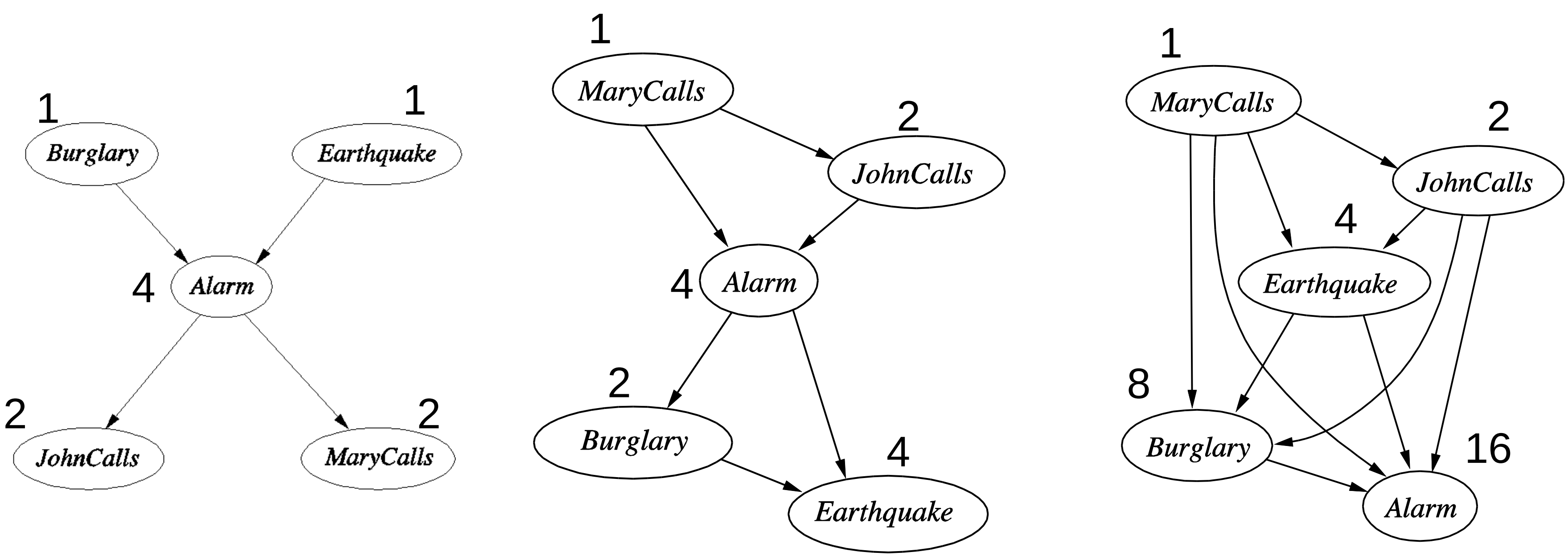
- Locally structured

- A CPT for a Boolean variable with k Boolean parents requires:
  - 2^k rows for the combinations of parent values
  - $k = 0$ (one row showing prior probabilities of each possible value of that variable)

- If each variable has no more than k parents
  - Baysian network requires $O(n \times 2^k)$ numbers (linear with n)
  - Full joint distribution requires $O(2^n)$ numbers

Node ordering: burglary example

- The structure of the network and so the number of required probabilities for different node orderings can be different



$1 + 1 + 4 + 2 + 2 = 10$ $\quad$ $1 + 2 + 4 + 2 + 4 = 13$ $\quad$ $1 + 2 + 4 + 8 + 16 = 31$

Constructing Bayesian networks

I. **Nodes**:

determine the set of variables and order them as $X_1, \ldots, X_n$

(More compact network if causes precede effects)

II. **Links**:

for $i = 1$ to n

1) select a minimal set of parents for X_i from $X_1, \ldots, X_{i-1}$ such that

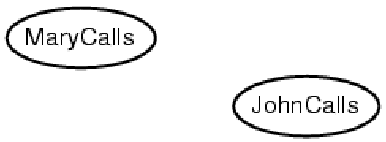
$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

2) For each parent insert a link from the parent to X_i

3) CPT creation based on $\mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$

Node ordering: Burglary example

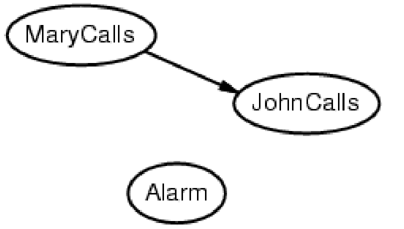
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

Node ordering: Burglary example

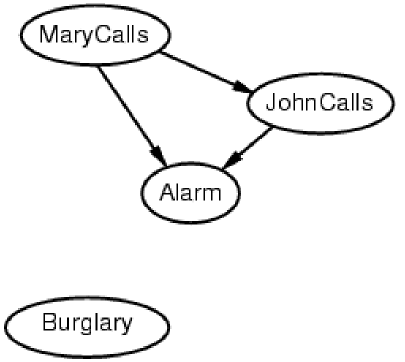
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$?
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$?

Node ordering: Burglary example

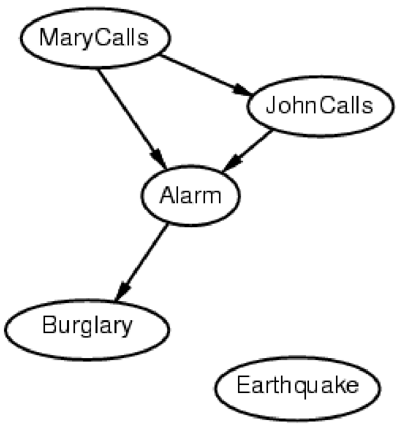
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$?
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$?

Node ordering: Burglary example

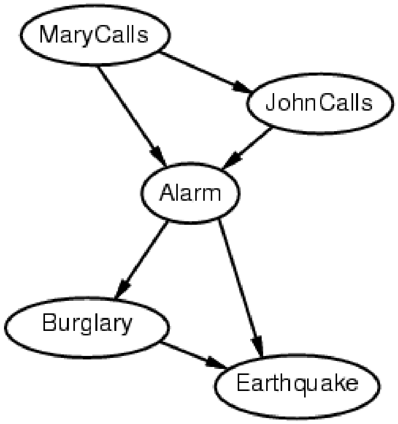
- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A)$?
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A, B)$?

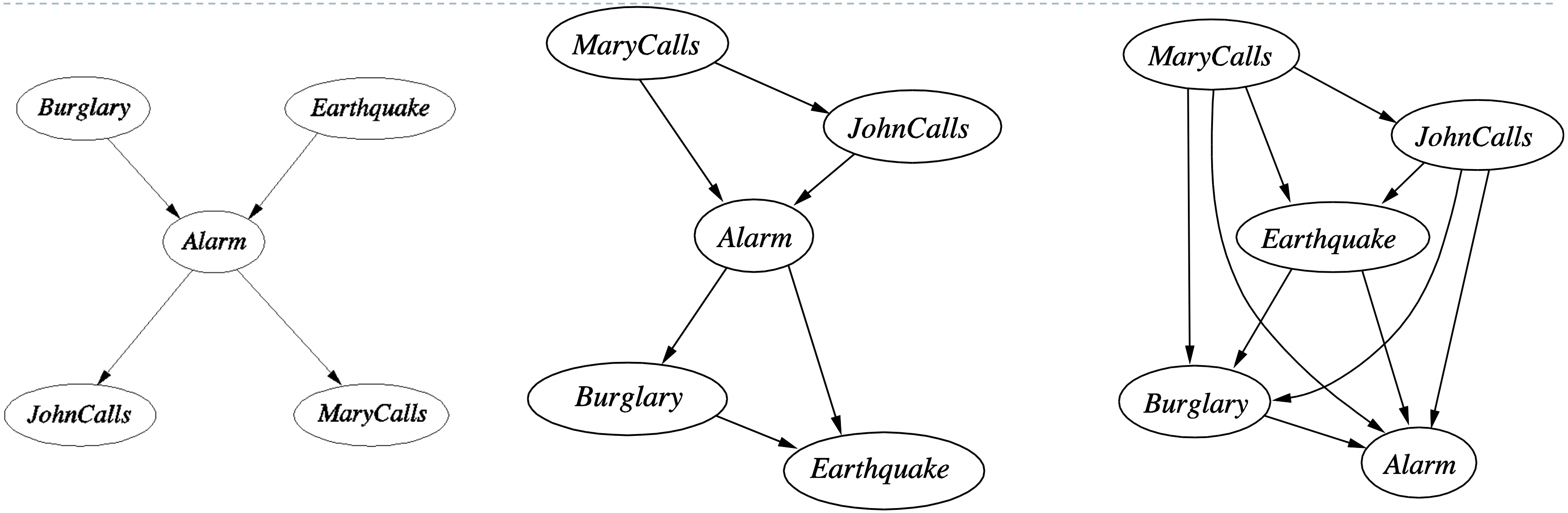
Node ordering: Burglary example

- Suppose we choose the ordering M, J, A, B, E



- $\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? **No**
- $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Yes**
- $\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A)$? **No**
- $\mathbf{P}(E \mid B, A, J, M) = \boldsymbol{P}(E \mid A, B)$? **Yes**

Causal models



- Some new links represent relationships that require difficult and unnatural probability judgments
  - Deciding conditional independence is hard in non-causal directions

Why using Bayesian networks?

- Compact representation of probability distributions: smaller number of parameters
  - Instead of storing full joint distribution requiring large number of parameters

- Incorporation of domain knowledge and causal structures

- Algorithm for systematic and efficient inference/learning
  - Exploiting the graph structure and probabilistic semantics
  - Take advantage of **conditional and marginal independences** among random variables

Inference query

- Nodes: $\boldsymbol{X} = \{X_1, \dots, X_n\}$
- Evidence: an assignment of values to a set $\boldsymbol{X_V}$ of nodes in the network

- Likelihood: $p(\boldsymbol{x_v}) = \sum_{\boldsymbol{X_H}} p(\boldsymbol{X_H}, \boldsymbol{x_v})$ $(\boldsymbol{X} = \boldsymbol{X_H} \cup \boldsymbol{X_V})$
- A posteriori belief: $p(\boldsymbol{X_H} | \boldsymbol{x_v}) = \frac{p(\boldsymbol{X_H}, \boldsymbol{x_v})}{\sum_{\boldsymbol{X_H}} p(\boldsymbol{X_H}, \boldsymbol{x_v})}$
- $p(\boldsymbol{Y} | \boldsymbol{x_v}) = \frac{\sum_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{x_v})}{\sum_{\boldsymbol{Y}} \sum_{\boldsymbol{Z}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{x_v})}$ $(\boldsymbol{X_H} = \boldsymbol{Y} \cup \boldsymbol{Z})$

Partial joint and conditional examples

- Partial joint probability distribution (joint probability distribution for a subset variables) can be computed from the full joint distribution through marginalization
  - $P(j, \neg m, b)$?

$$P(j, \neg m, b) = \sum_A \sum_E P(j, \neg m, b, A, E)$$

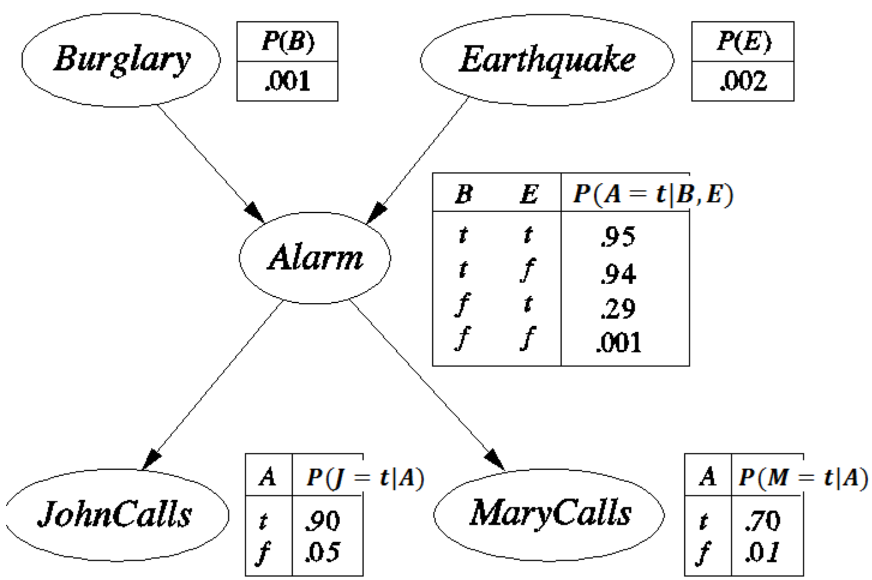
- Conditional probability distribution:
  - Can be computed from the full joint distribution through marginalization and definition of conditionals
  - Burglary example: $P(b|j, \neg m)$?

$$P(b|j, \neg m) = \frac{P(j, \neg m, b)}{P(j, \neg m)} = \frac{\sum_A \sum_E P(j, \neg m, b, A, E)}{\sum_B \sum_A \sum_E P(j, \neg m, b, A, E)}$$

Burglary example: full joint probability

$$P(b|j,\neg m) = \frac{P(j,\neg m,b)}{P(j,\neg m)} = \frac{\sum_A \sum_E P(j,\neg m,b,A,E)}{\sum_B \sum_A \sum_E P(j,\neg m,b,A,E)}$$

$$= \frac{\sum_A \sum_E P(j|A)P(\neg m|A)P(A|b,E)P(b)P(E)}{\sum_B \sum_A \sum_E P(j|A)P(\neg m|A)P(A|B,E)P(B)P(E)}$$



Short-hands

j: $JohnCalls = True$

$\neg b$: $Burglary = False$

…

Inference in Bayesian networks

- Computing $p(\boldsymbol{X_H}|\boldsymbol{x_v})$ in an arbitrary GM is NP-hard

- Exact inference: enumeration intractable (NP-Hard)
  - Special cases are tractable

Enumeration

- $P(\boldsymbol{Y}|\boldsymbol{x}_{\boldsymbol{v}}) \propto P(\boldsymbol{Y}, \boldsymbol{x}_{\boldsymbol{v}})$

- $P(\boldsymbol{Y}, \boldsymbol{x}_{\boldsymbol{v}}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Y}, \boldsymbol{x}_{\boldsymbol{v}}, \boldsymbol{Z})$
  - exponential in general (with respect to the number of nodes in $\boldsymbol{Z}$)
  - we cannot find a general procedure that works **efficiently** for arbitrary networks

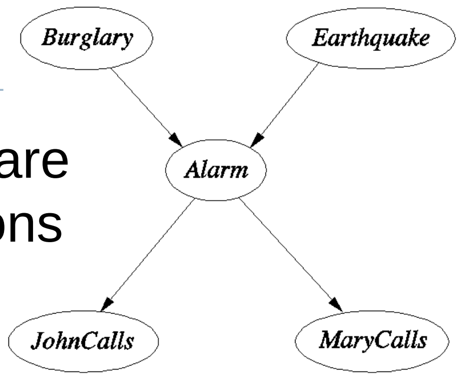
- Sometimes the structure of the network allows us to infer more efficiently
  - avoiding exponential cost

- Can be improved by re-using calculations
  - similar to dynamic programming

Distribution of products on sums

- Exploiting the factorization properties to allow sums and products to be interchanged
  - $a \times b + a \times c$ needs three operations while $a \times (b + c)$ requires two

Variable elimination: example



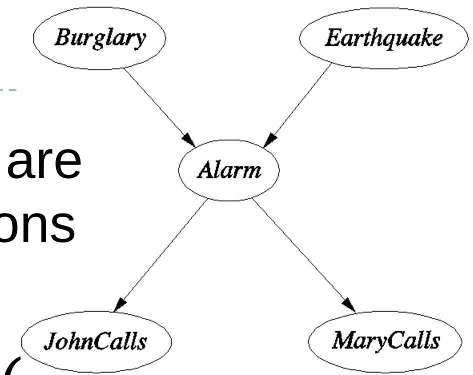
$$P(b|j) \propto P(b,j)$$

Intermediate results are probability distributions

$$P(b,j) = \sum_A \sum_E \sum_M P(b)P(E)P(A|b,E)P(j|A)P(M|A)$$

$$= P(b) \sum_E P(E) \sum_A P(A|b,E)P(j|A) \sum\nolimits_M P(M|A)$$

Variable elimination: example



$$P(B|j) \propto P(B,j)$$

Intermediate results are probability distributions

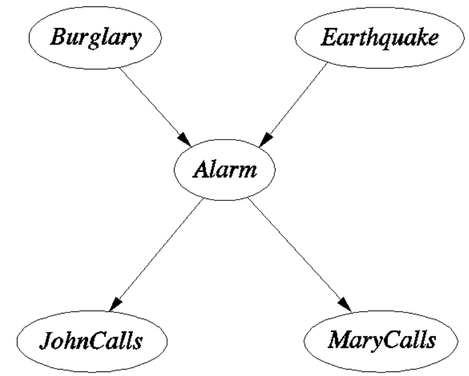
$$P(B,j) = \sum_A \sum_E \sum_M \underbrace{P(B)}_{\boldsymbol{f}_1(B)} \underbrace{P(E)}_{\boldsymbol{f}_2(E)} \underbrace{P(A|B,E)}_{\boldsymbol{f}_3(A,B,E)} \underbrace{P(j|A)}_{\boldsymbol{f}_4(A)} \underbrace{P(M|A)}_{\boldsymbol{f}_5(A,M)}$$

$$= P(B) \sum_E P(E) \sum_A P(A|B,E)P(j|A) \underbrace{\sum\nolimits_M P(M|A)}_{\boldsymbol{f}_6(A) \Rightarrow \begin{bmatrix}1\\1\end{bmatrix}}$$

$$\boldsymbol{f}_7(B,E) = \sum\nolimits_A \boldsymbol{f}_3(A,B,E) \times \boldsymbol{f}_4(A) \times \boldsymbol{f}_6(A)$$

$$\boldsymbol{f}_8(B) = \sum\nolimits_E \boldsymbol{f}_2(E) \times \boldsymbol{f}_7(B,E)$$

Variable elimination: Order of summations



- An inefficient order:

$$P(B,j) = \sum_{M}\sum_{E}\sum_{A} P(B)P(E)P(A|B,E)P(j|A)P(M|A)$$

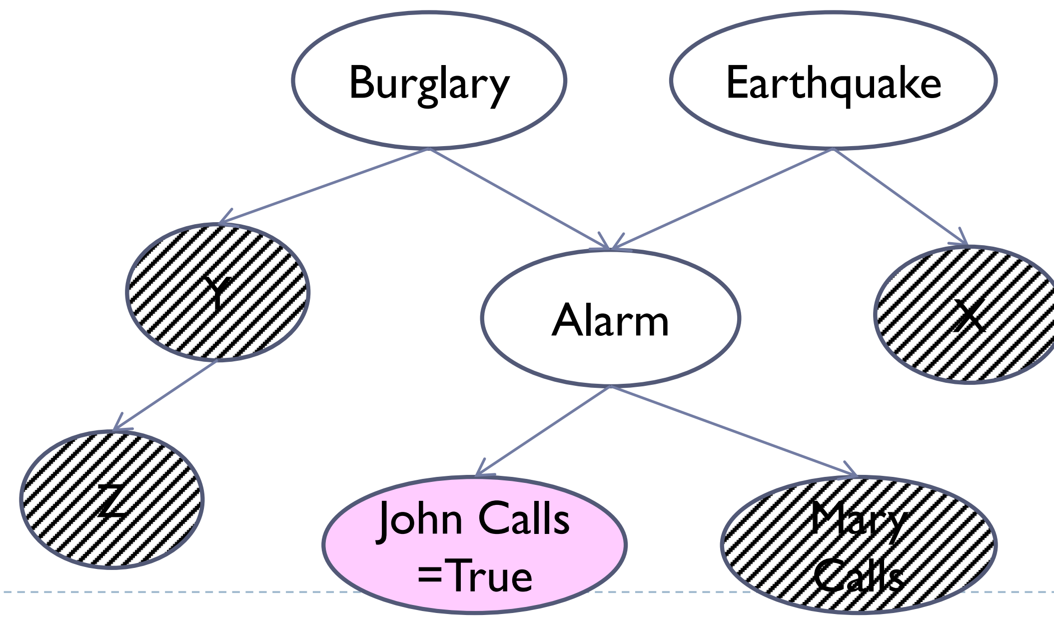
$$= P(B)\sum\nolimits_{M}\sum\nolimits_{E} P(E)\sum_{A} \underbrace{P(A|B,E)P(j|A)P(M|A)}_{\boldsymbol{f}(A,B,E,M)}$$

Variable elimination: Pruning irrelevant variables

- Any variable that is not an ancestor of a query variable or evidence variable is irrelevant to the query.

- Prune all non-ancestors of query or evidence variables:

  $P(b, j)$



Variable elimination algorithm

- Given: BN, evidence e, a query $P(\boldsymbol{Y}|\boldsymbol{x_v})$
- Prune non-ancestors of $\{\boldsymbol{Y}, \boldsymbol{X_V}\}$
- Choose an **ordering** on variables, e.g., $X_1, \dots, X_n$
- For i = 1 to n, If $X_i \notin \{\boldsymbol{Y}, \boldsymbol{X_V}\}$
  - Collect factors $\boldsymbol{f}_1, \dots, \boldsymbol{f}_k$ that include X_i
  - Generate a new factor by eliminating X_i from these factors:

$$\boldsymbol{g} = \sum_{X_i} \prod_{j=1}^{k} \boldsymbol{f}_j$$

- Normalize $P(\boldsymbol{Y}, \boldsymbol{x_v})$ to obtain $P(\boldsymbol{Y}|\boldsymbol{x_v})$

After this summation, X_i is eliminated

Variable elimination algorithm

- Given: BN, evidence e, a query $P(\boldsymbol{Y}|\boldsymbol{x}_v)$
- Prune non-ancestors of $\{\boldsymbol{Y}, \boldsymbol{X}_V\}$
- Choose an **ordering** on variables, e.g., $X_1, \ldots, X_n$
- For i = 1 to n, If $X_i \notin \{\boldsymbol{Y}, \boldsymbol{X}_V\}$
  - Collect factors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k$ that include X_i
  - Generate a new factor by eliminating X_i from these factors:

$$\boldsymbol{g} = \sum_{X_i} \prod_{j=1}^{k} \boldsymbol{f}_j$$

- Normalize $P(\boldsymbol{Y}, \boldsymbol{x}_v)$ to obtain $P(\boldsymbol{Y}|\boldsymbol{x}_v)$

- Evaluating expressions in a proper order
- Storing intermediate results
- Summation only for those portions of the expression that depend on that variable

Variable elimination

- Eliminates by summation non-observed non-query variables one by one by distributing the sum over the product
- Complexity determined by the size of the largest factor
- Variable elimination can lead to significant costs saving but its efficiency depends on the network structure .
  - there are still cases in which this algorithm we lead to exponential time.

Summary

- Probability is the most common way to represent uncertain knowledge
- Whole knowledge: Joint probability distribution in probability theory (instead of truth table for KB in two-valued logic)
- Independence and conditional independence can be used to provide a compact representation of joint probabilities
- Bayesian networks provides a compact representation of joint distribution including network topology and CPTs
  - Using independencies and conditional independencies
  - They can also lead to more efficient inference using these independencies