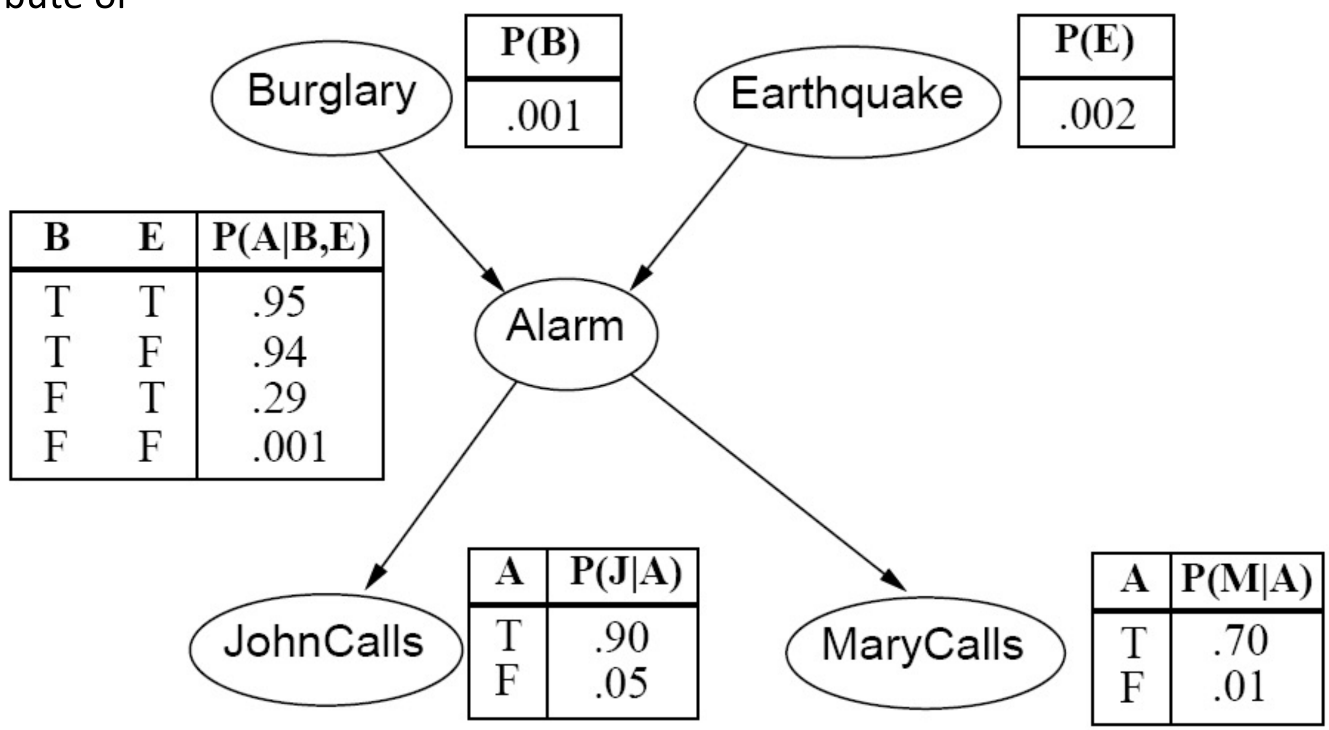
CS440/ECE448 Lecture 20: Bayesian Networks

Mark Hasegawa-Johnson, 3/2022



Outline

- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence

Review: Bayesian Classifier

- Class label $Y = y$, drawn from some set of labels
- Observation $X = x$, drawn from some set of features
- Bayesian classifier: choose the class label, y, that minimizes your probability of making a mistake:

$$\hat{y} = \underset{y}{\operatorname{argmin}}\, P(Y \neq y | X = x)$$

Minimum Probability of Error = Maximum A Posteriori

- The minimum probability of error (MPE) classifier is the one that minimizes your probability of making a mistake:

$$\hat{y} = \underset{y}{\operatorname{argmin}} P(Y \neq y | X = x)$$

- The maximum a posteriori (MAP) classifier is the one that maximizes your probability of being correct:

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(Y = y | X = x)$$

- Notice: they're the same! This is called the MPE=MAP rule.

Today: What if P(X,Y) is complicated?

Very, very common problem: P(X,Y) is complicated because both X and Y depend on some hidden variable H

$$P(Y=y|X=x) = \frac{\sum_h P(X=x, H=h, Y=y)}{\sum_{h,y'} P(X=x, H=h, Y=y')}$$

Why is this a problem?

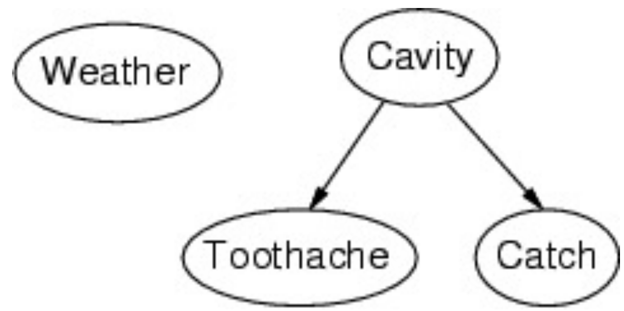
1. **SPACE COMPLEXITY**: $P(X=x, H=h, Y=y)$ requires $|X| \cdot |H| \cdot |Y|$ entries
   - Example: X has cardinality 1000, H has cardinality 1000, Y has cardinality 1000, then $P(X=x, H=h, Y=y)$ is a probability table with 1 billion entries.

2. **TIME COMPLEXITY**: The summation requires a lot of time.

Outline

- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence

Bayesian networks: Structure



- **Nodes:** random variables
- **Arcs:** interactions
  - An arrow from one variable to another indicates direct ***causal*** influence of variable #1 on variable #2
  - Must form a directed, acyclic graph

Conditional independence and the joint distribution

- Key property: each node is conditionally independent of its *non-descendants* given its *parents*
- Suppose the nodes $X_1, \ldots, X_n$ are sorted in topological order
- To get the joint distribution $P(X_1, \ldots, X_n)$, use chain rule:

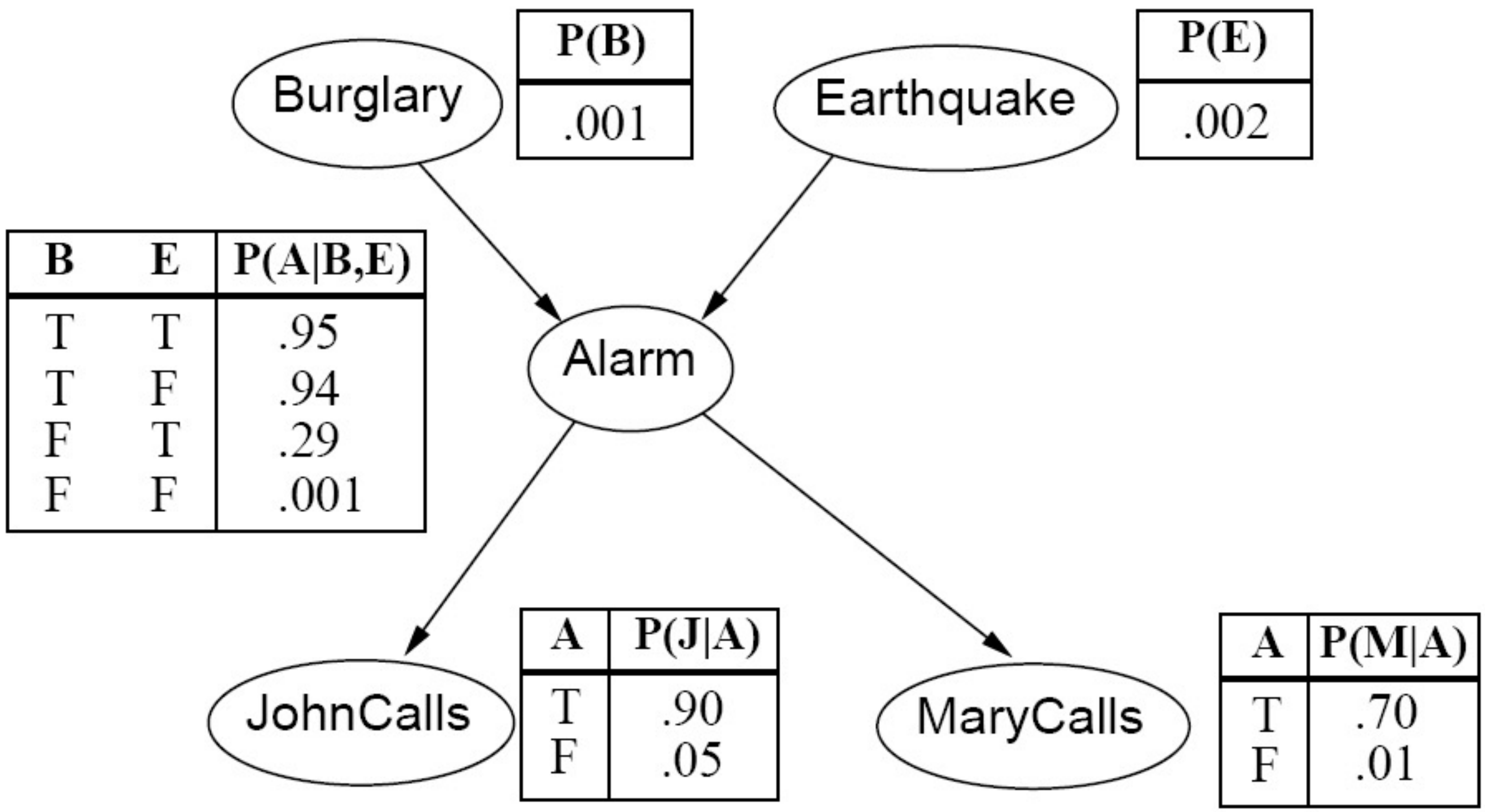
$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid X_1, \ldots, X_{i-1}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i \mid Parents(X_i)\right)$$

Example: Los Angeles Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary, Earthquake, Alarm, John, Mary*
- What are the direct influence relationships?
  - 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

Example: Burglar Alarm



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Space complexity: LA Burglar Alarm

- How much space do we need to store the model without dependencies?
  - 5 variables
  - Each is binary
  - $P(B, E, A, J, M)$ is a table with $2^5 = 32$ entries
  - Since they add up to 1, we could store just $2^5 - 1 = 31$ entries
- How much space do we need to store the Bayes net parameters?
  - $P(B), P(E)$: two numbers
  - $P(A|B = b, E = e)$: one entry for each setting of $b \in \{F, T\}$, $e \in \{F, T\}$
  - $P(J|A = a), P(M|A = a)$: two numbers for each setting of $a \in \{F, T\}$
  - Total: $1 + 1 + 4 + 2 + 2 = 10$ entries

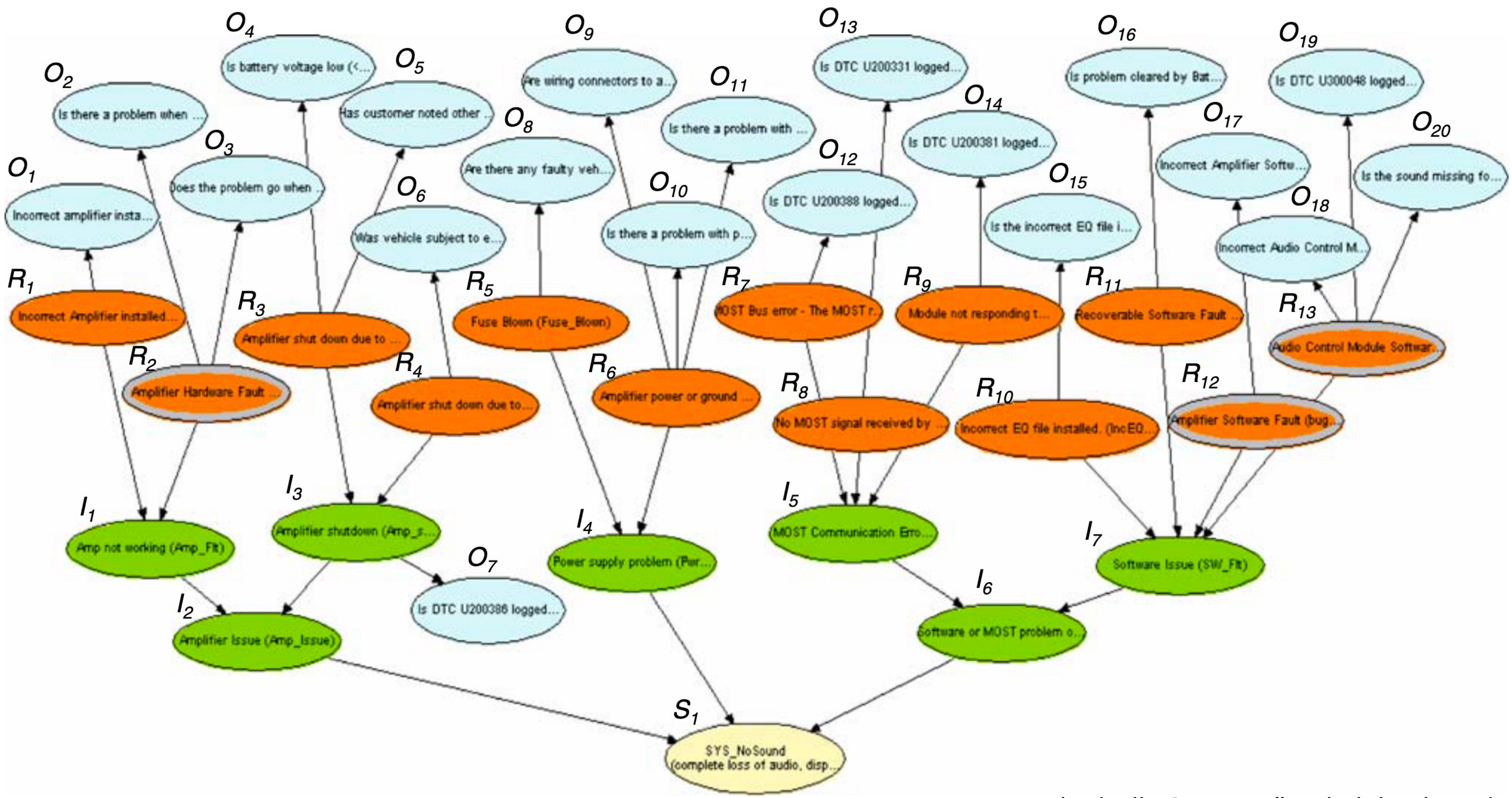


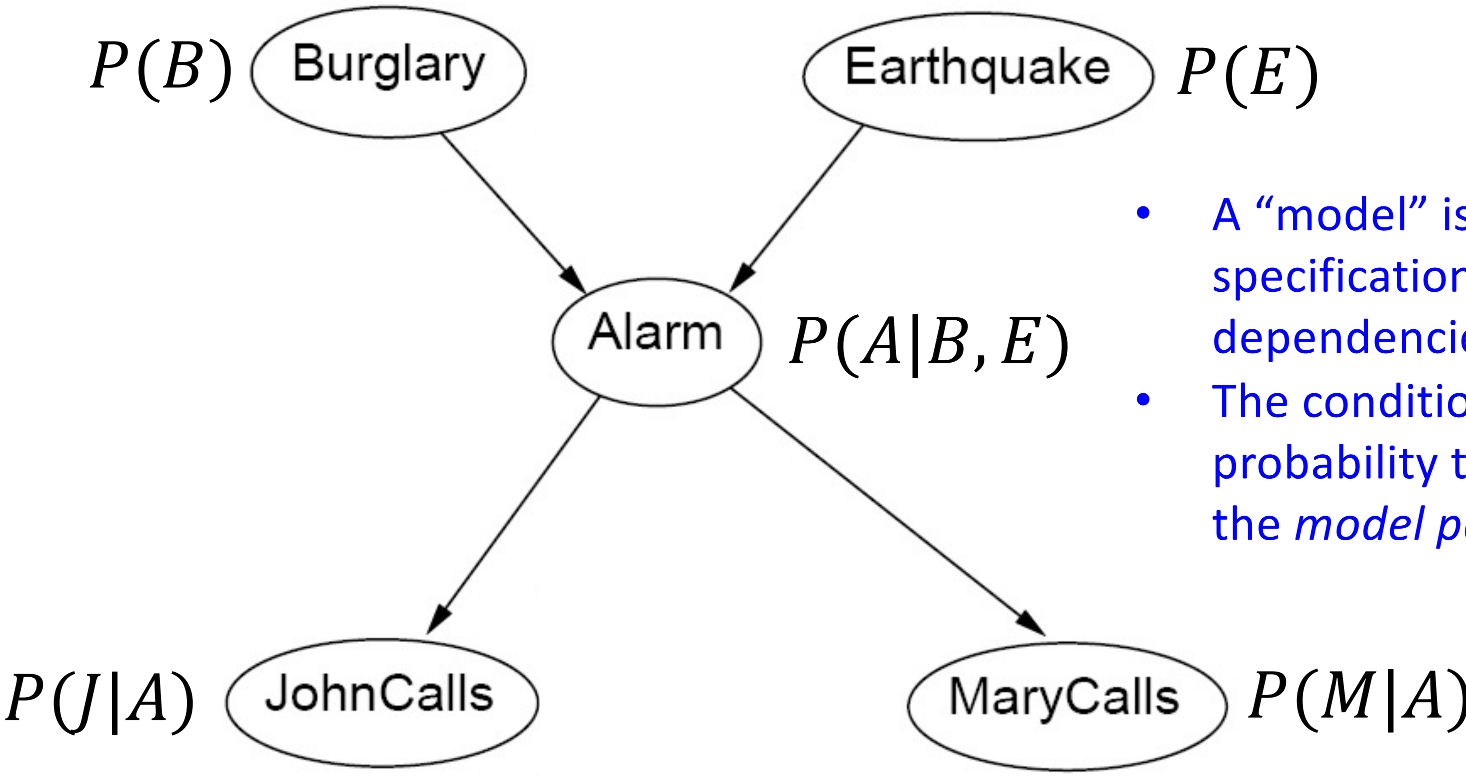
Fig. 6 Bayesian diagnostic model for the symptom "no sound"

Huang, McMurran, Dhadyalla & Jones, "Probability-based vehicle fault diagnosis: Bayesian network method," 2008

Space complexity, Huang et al. “no sound” diagnosis model

- How much space do we need to store the model without dependencies?
  - 41 binary variables: table would require $2^{41} - 1 = 2{,}199{,}023{,}255{,}551$ entries
- How much space do we need to store the Bayes net parameters?
  - One binary variable with four binary parents, requires one entry for each of the $2^4 = 16$ values of its parent variables
  - Two binary variable with three binary parents, each require 8 entries
  - Five binary variables with two binary parents, each require 4 entries
  - Twenty binary variables with one binary parent, each require 2 entries
  - Thirteen binary variables with no parents, each require 1 entry
  - Total: $16 + 2\times 8 + 5\times 4 + 20\times 2 + 13 = 105$ entries

Example: Burglar Alarm



- A “model” is a complete specification of the dependencies.
- The conditional probability tables are the *model parameters.*

Outline

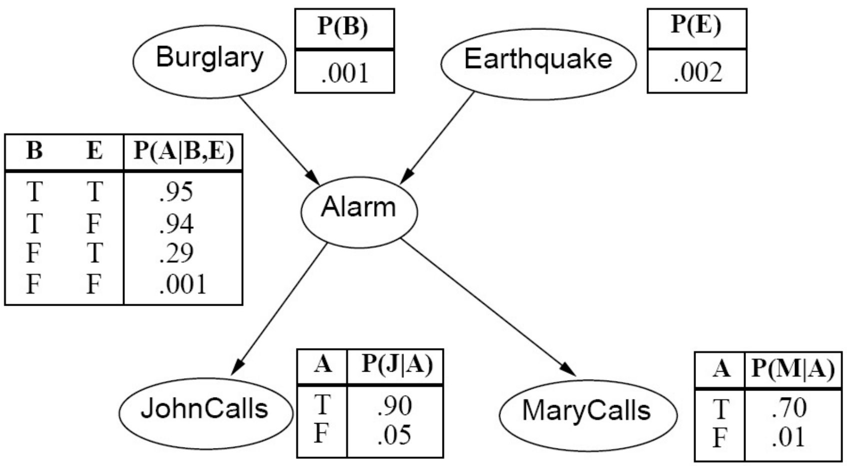
- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence

Classification using probabilities

- Suppose Mary has called to tell you that you had a burglar alarm. Should you call the police?
  - Make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision. You decide that you have a burglar in your house if and only if

$$P(Burglary = T|Mary = T) > P(Burglary = F|Mary = T)$$

Using a Bayes network to estimate a posteriori probabilities

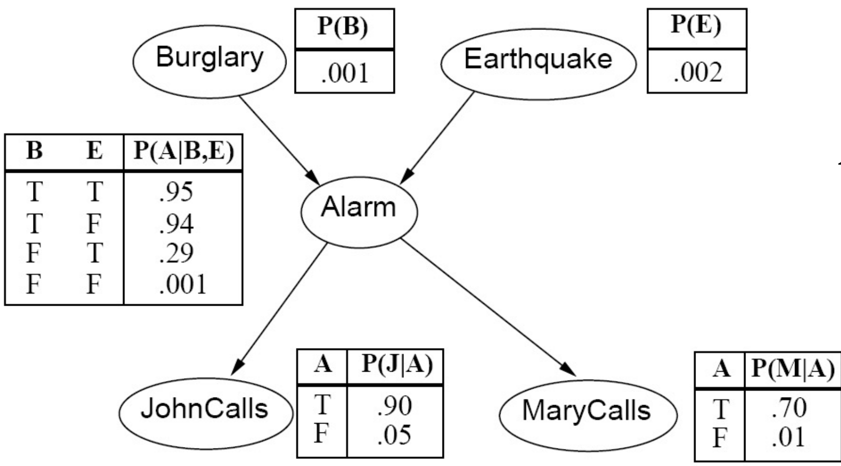


- Notice: we don't know $P(B|M)$! We have to figure out what it is.
- This is called "inference".
- First step: find the joint probability of B, M, and any other variables that are necessary in order to link these two together.

$$P(B,E,A,M) = P(B)P(E)P(A|B,E)P(M|A)$$

$P(BEAM)$	$M=F, A=F$	$M=F, A=T$	$M=T, A=F$	$M=T, A=T$
$B=F, E=F$	0.986045	2.99×10^{-4}	9.96×10^{-3}	6.98×10^{-4}
$B=F, E=T$	1.4×10^{-3}	1.7×10^{-4}	1.4×10^{-5}	4.06×10^{-4}
$B=T, E=F$	5.93×10^{-5}	2.81×10^{-4}	5.99×10^{-7}	6.57×10^{-4}
$B=T, E=T$	9.9×10^{-8}	5.7×10^{-7}	10^{-9}	1.33×10^{-6}

Using a Bayes network to estimate a posteriori probabilities



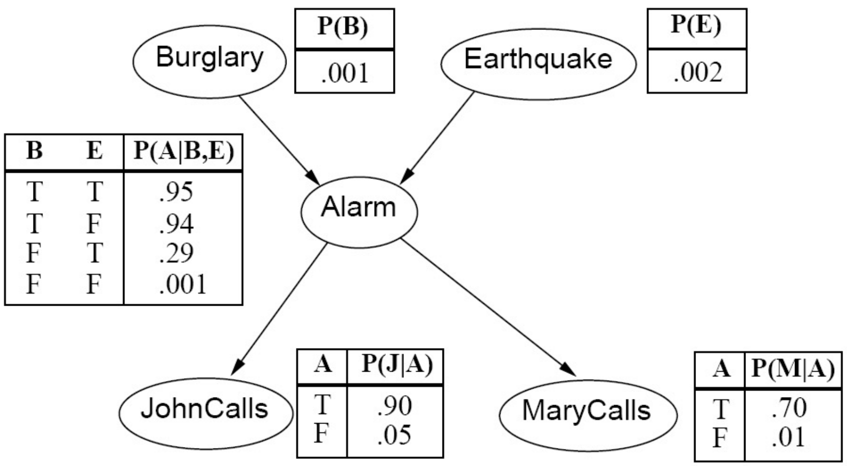
Second step: marginalize (add) to get rid of the variables you don't care about.

$$P(B,M) = \sum_{e\in\{F,T\}} \sum_{a\in\{F,T\}} P(B,E=e,A=a,M)$$

$P(B,M)$	$M=F$	$\boldsymbol{M=T}$
$B=F$	0.987922	0.011078
$B=T$	0.000341	0.000659

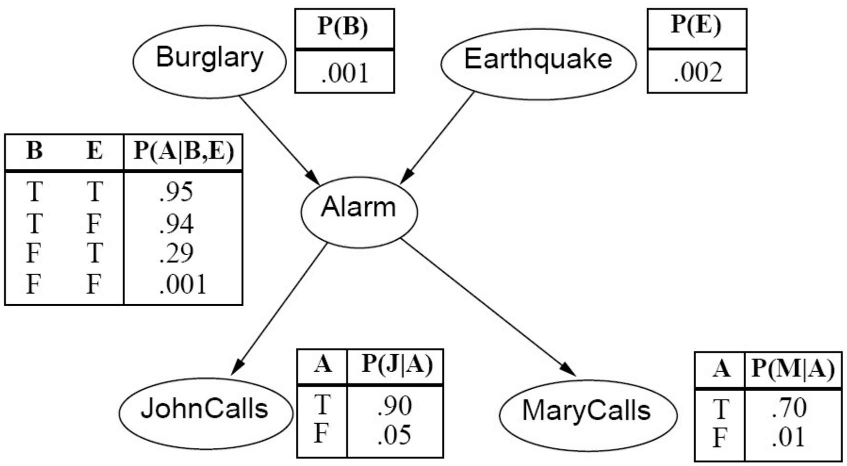
Using a Bayes network to estimate a posteriori probabilities

Third step: ignore (delete) the column that didn't happen.



$P(B,M)$	$\boldsymbol{M=T}$
$B=F$	0.011078
$B=T$	0.000659

Using a Bayes network to estimate a posteriori probabilities



Fourth step: use the definition of conditional probability.

$$P(B = T|M = T)$$

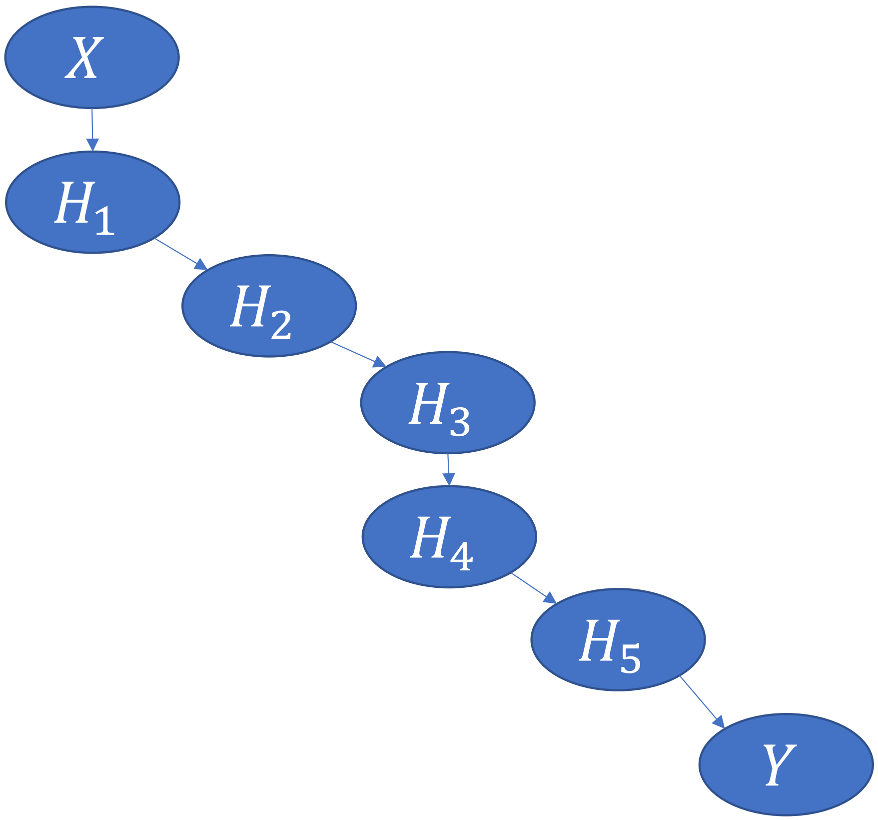
$$= \frac{P(B = T, M = T)}{P(B = T, M = T) + P(B = F, M = T)}$$

$P(B\|M)$	$\boldsymbol{M = T}$
$B = F$	0.943883
$B = T$	0.056117

Some unexpected conclusions

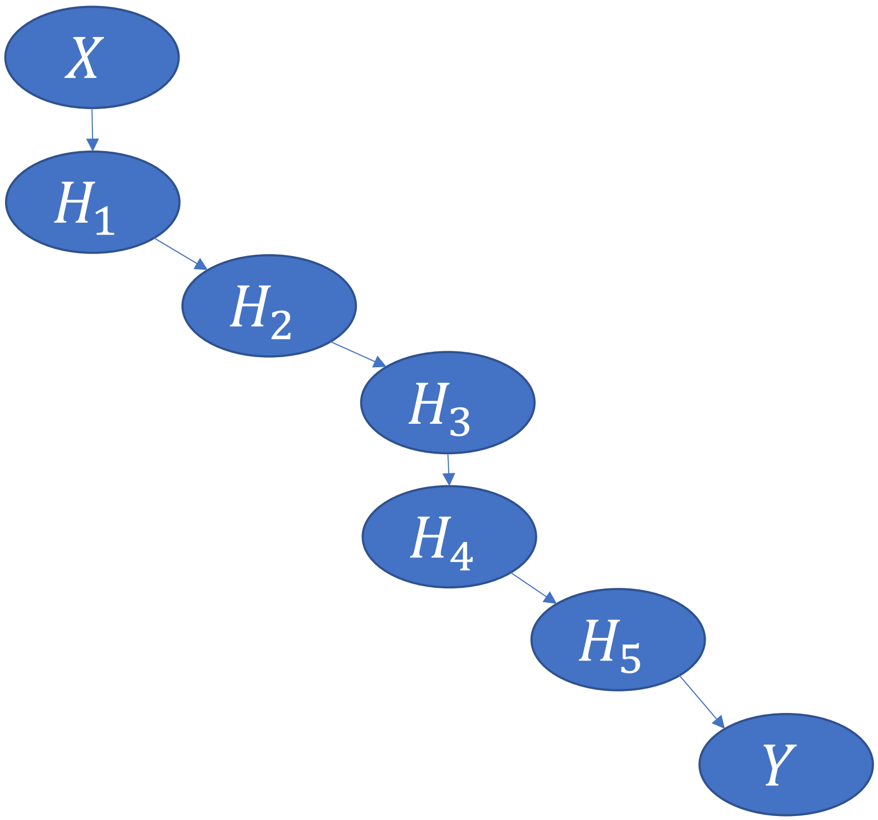
- Burglary is so unlikely that, if only Mary calls or only John calls, the probability of a burglary is still only about 5%.
- If both Mary and John call, the probability is ~50%.

Belief propagation: The general algorithm



Given an arbitrary Bayes net, you want to find the joint probability of two variables, X and Y, that are connected by a chain of intermediate variables, H_1 through H_N.

Belief propagation: The general algorithm



Initialize:

Start with P(X)

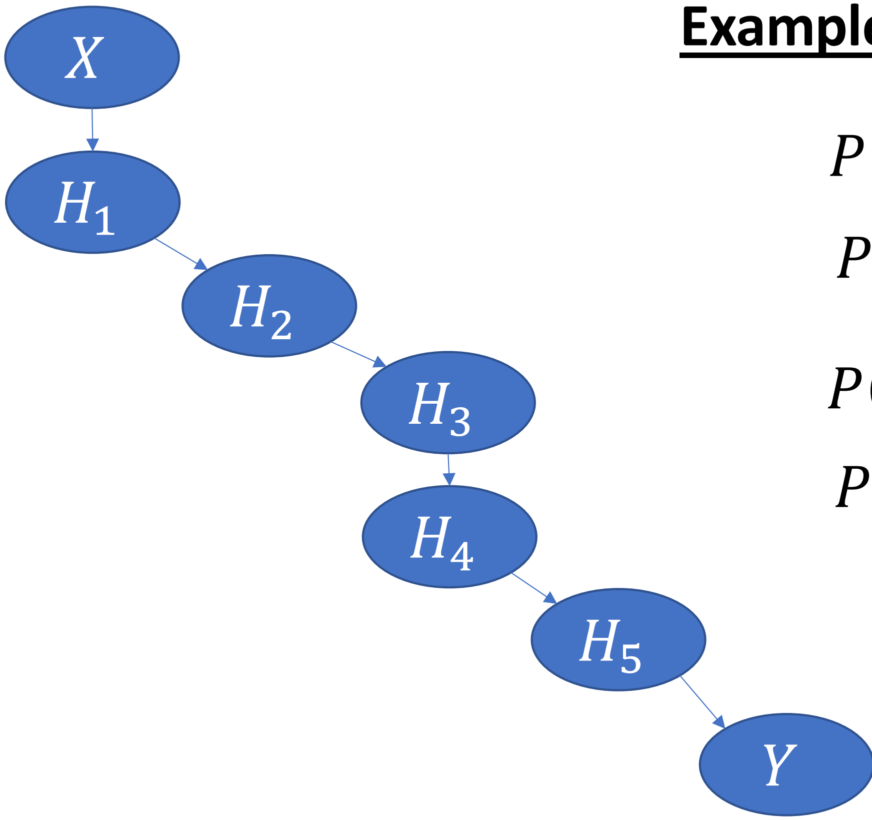
Iterate:

1. PRODUCT: Multiply in the next variable
2. SUM: Marginalize out any variables you no longer need

Terminate:

When you have P(X,Y)

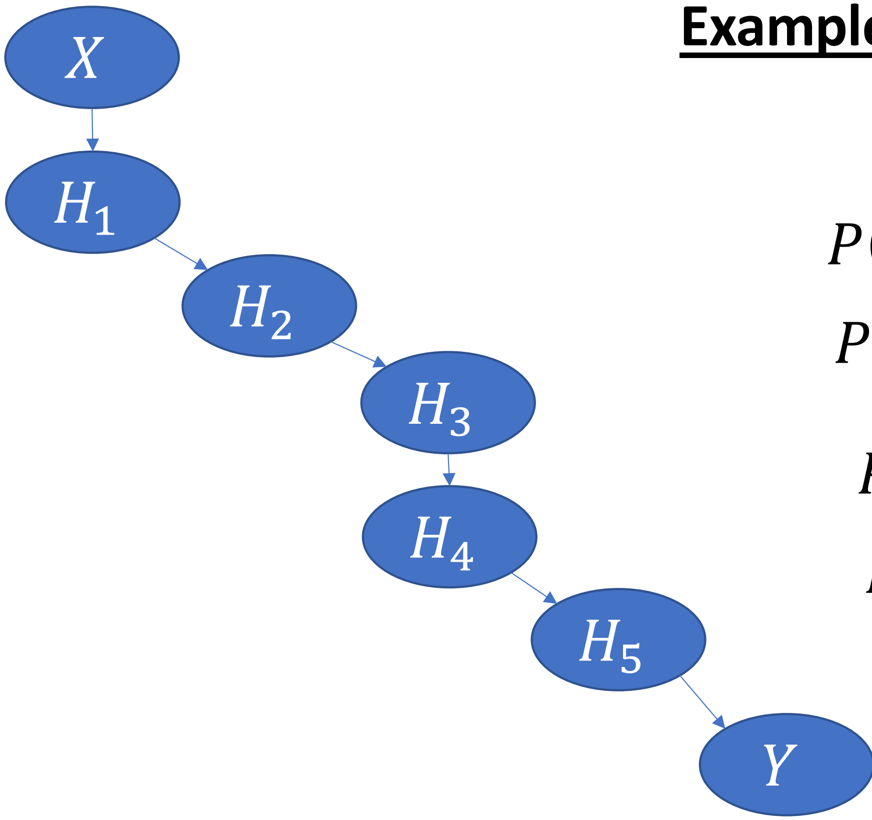
Belief propagation: The general algorithm



Example:

$$P(X, H_1) = P(X)P(H_1|X)$$
$$P(X, H_1, H_2) = P(X, H_1)P(H_2|H_1)$$
$$P(X, H_2) = \sum_{h_1} P(X, H_1 = h_1, H_2)$$
$$P(X, H_2, H_3) = P(X, H_2)P(H_3|H_2)$$
$$P(X, H_3) = \sum_{h_2} P(X, H_2 = h_2, H_3)$$
$$\vdots$$

Belief propagation: The general algorithm



Example:

⋮

$$P(X, H_4, H_5) = P(X, H_4)P(H_5|H_4)$$

$$P(X, H_5) = \sum_{h_4} P(X, H_4 = h_4, H_5)$$

$$P(X, H_5, Y) = P(X, H_5)P(Y|H_5)$$

$$P(X, Y) = \sum_{h_5} P(X, H_5 = h_5, Y)$$

Belief propagation: Space and time complexity

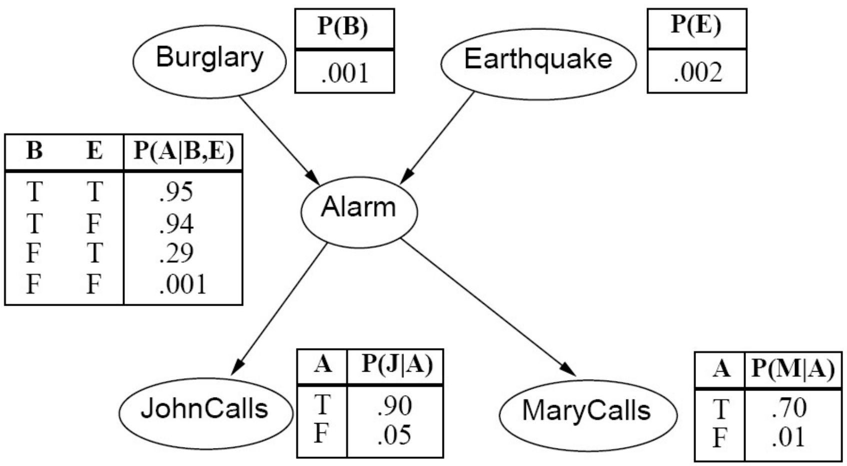
- If there is just one path from X to Y (as shown in the example), then space and time complexity of belief propagation are each K^3, where K is the maximum cardinality of any of the random variables.
  - Each product operation results in a table of 3 variables, with $K^3 - 1$ entries
  - Each summation is over K entries, for each of K^2 combinations

- If there are multiple paths from X to Y, or if there are multiple X variables (many different relevant observations), then belief propagation becomes NP-complete
  - It's necessary to create a probability table containing all the variables in all the paths between X and Y
  - That table has $K^{2N+1} - 1$ entries, where N is the number of different paths that connect X to Y

Outline

- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence

Using a Bayes network to estimate a posteriori probabilities



Fourth step: use the definition of conditional probability.

$$P(B=T|M=T)$$

$$=\frac{P(B=T,M=T)}{P(B=T,M=T)+P(B=F,M=T)}$$

$P(B\|M)$	$\boldsymbol{M=T}$
$B=F$	0.943883
$B=T$	0.056117

Some unexpected conclusions

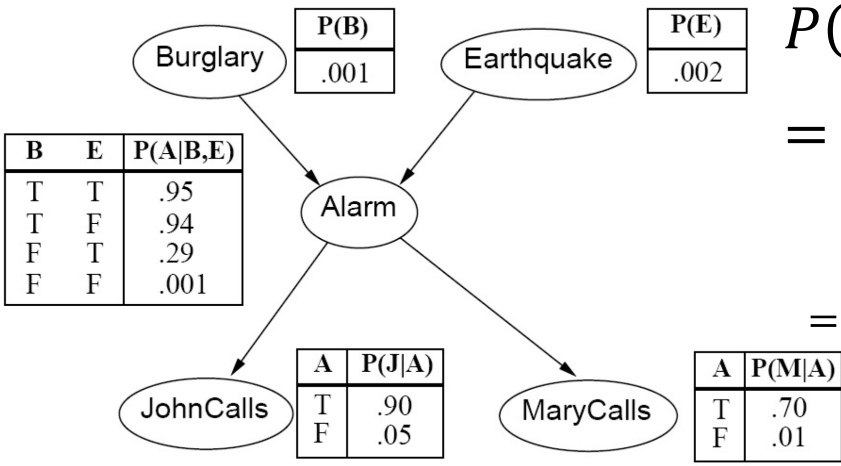
- If only Mary calls or only John calls, the probability of a burglary is about 5% or 6%.

unless …

- If you know that there was an earthquake, then it's very likely that the alarm was caused by the earthquake. In that case, the probability you had a burglary is vanishingly small, even if twenty of your neighbors call you.
- This is called the "explaining away" effect. The earthquake "explains away" the burglar alarm.

The “Explaining Away” Effect

Probability of a Burglary, given that Mary called, and given a known earthquake:



$$P(B=T|M=T,E=T)$$

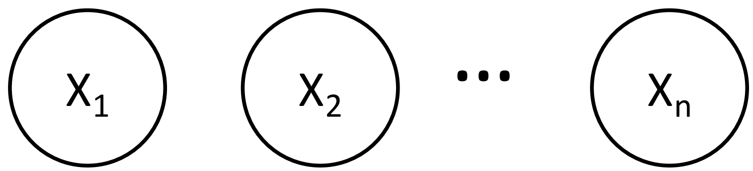
$$= \frac{\sum_{a\in\{F,T\}} P(M=T,A=a,E=T,B=T)}{\sum_{a\in\{F,T\},b\in\{F,T\}} P(M=T,A=a,E=T,B=b)}$$

$$= \frac{(0.001)(0.002)(0.95)(0.7) + (0.001)(0.002)(0.05)(0.01)}{\begin{pmatrix} (0.001)(0.002)(0.95)(0.7) + (0.001)(0.002)(0.05)(0.01) \\ +(0.999)(0.002)(0.29)(0.7) + (0.999)(0.002)(0.71)(0.01) \end{pmatrix}}$$

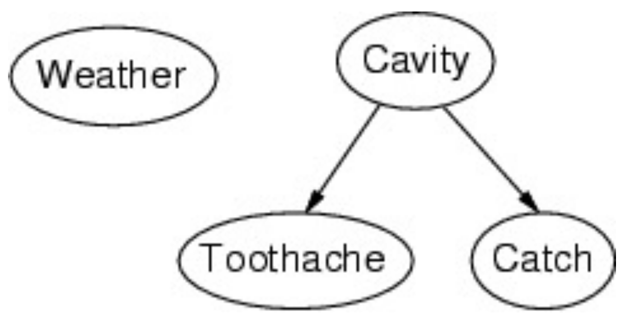
$$= 0.003$$

Independence

- By saying that X_i and X_j are independent, we mean that
$$\mathrm{P}(X_j, X_i) = \mathrm{P}(X_i)\mathrm{P}(X_j)$$
- X_i and X_j are independent if and only if they have no common ancestors
- Example: *independent coin flips*



- Another example: Weather is independent of all other variables in this model.

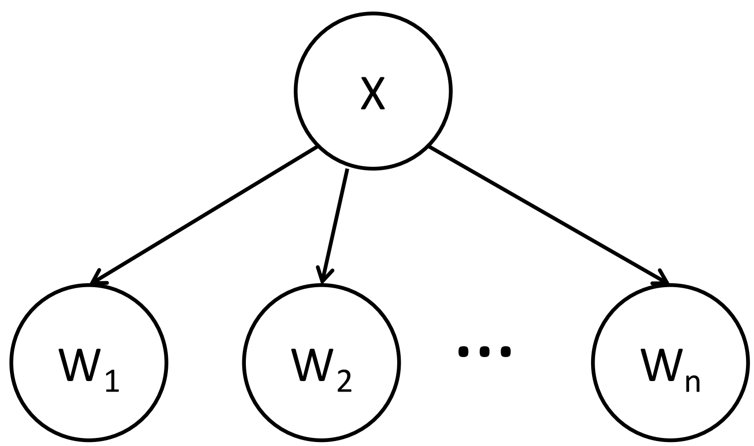


Conditional independence

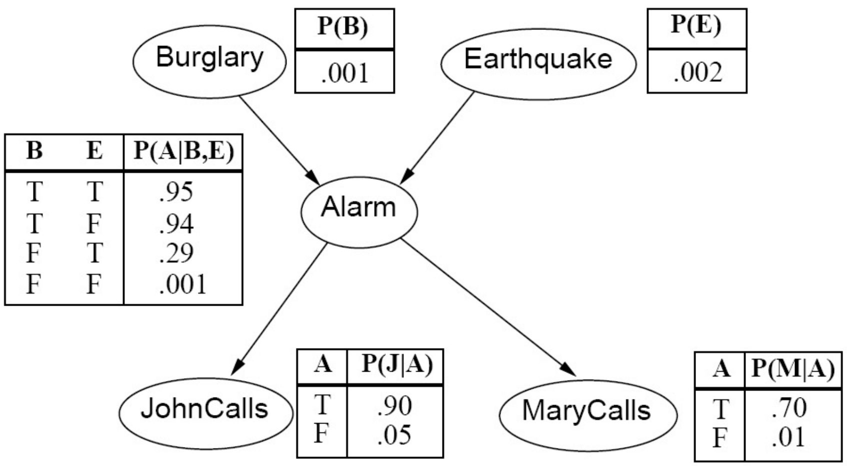
- By saying that W_i and W_j are conditionally independent given X, we mean that

$$\mathrm{P}\left(W_i, W_j \middle| X\right) = \mathrm{P}(W_i|X)\mathrm{P}(W_j|X)$$

- W_i and W_j are conditionally independent given X if and only if they have no common ancestors other than the ancestors of X.
- Example: *naïve Bayes model:*



Conditional Independence ≠ Independence



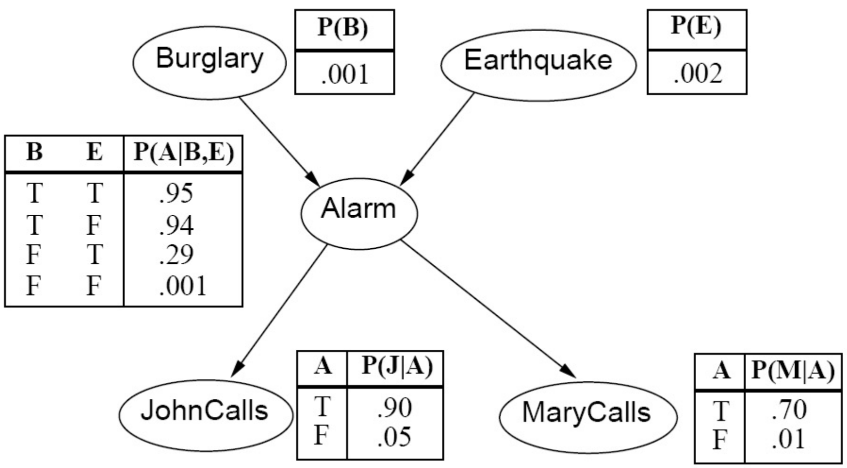
B and E are **independent**:

$$P(B|E) = P(B)$$

B and E are **not conditionally independent given A**:

$$P(B|E,A) \neq P(B|E)$$

Conditional Independence ≠ Independence



J and M are **<u>conditionally independent given A:</u>**

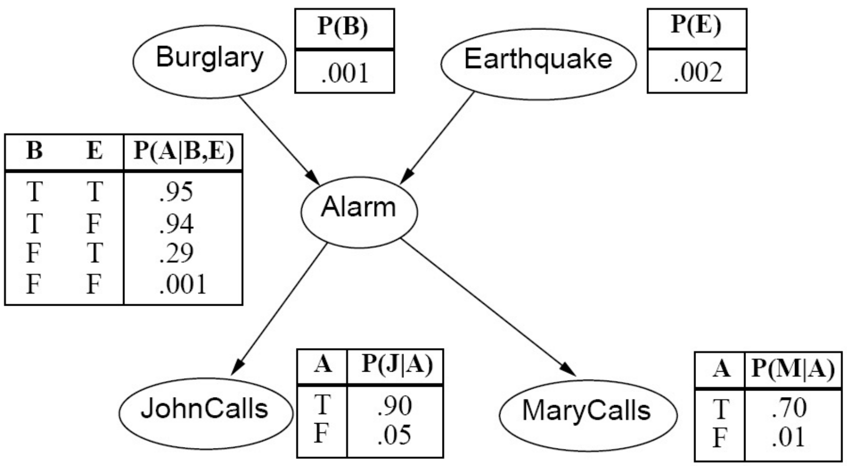
$$P(J|A, M) = P(J|A)$$

$$P(M|A, J) = P(M|A)$$

J and M are **<u>not independent</u>**!

$$P(J|M) \neq P(J)$$

Conditional Independence ≠ Independence



B and M are **conditionally independent given A:**

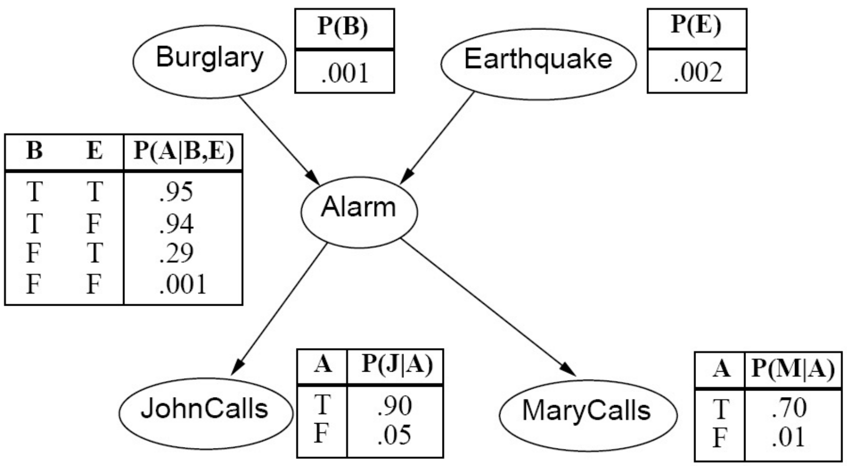
$$P(B|A, M) = P(B|A)$$

$$P(M|A, B) = P(M|A)$$

B and M are **not independent**!

$$P(B|M) \neq P(B)$$

Conditional Independence ≠ Independence



- B and E (no common ancestor, common descendant A):
  - Independent
  - Not conditionally independent given A
- J and M (common ancestor A, no common descendant):
  - Not independent
  - Conditionally independent given A
- B and M (B is the ancestor of M):
  - Not independent
  - Conditionally independent given A

Conditional Independence ≠ Independence

- Variables in a Bayes net are **independent** if they have no common ancestors
  - If they have a common ancestor (e.g., J and M), they are not independent
  - If one is the ancestor of the other (e.g., B and M), they are not independent
- Variables in a Bayes net are **conditionally independent** given knowledge of:
  - Their common ancestors, and
  - A variable that is a descendant of one, and an ancestor of the other

Outline

- Why Bayes nets? The complexity of a true Bayes classifier
- Space complexity
- Time complexity
- Independence and Conditional independence