Lecture notes for *Principles of Artificial Intelligence* (COMS 4720/5720)
Yan-Bin Jia, Iowa State University

Bayesian Networks (Bayes Nets)

Outline

I. Semantics

II. Network construction

* Figures are either from the textbook site or by the instructor.

I. Knowledge in an Uncertain Domain

- The full joint probability distribution can answer any question, but it also has several drawbacks:
  - exponential in the number n of variables and intractable as n grows very large
  - unnatural and tedious to specify probabilities of outcomes one by one
  - inadequate for representing human reasoning (good at conditional probabilities but poor at joint probabilities)

I. Knowledge in an Uncertain Domain

- The full joint probability distribution can answer any question, but it also has several drawbacks:
  - exponential in the number n of variables and intractable as n grows very large
  - unnatural and tedious to specify probabilities of outcomes one by one
  - inadequate for representing human reasoning (good at conditional probabilities but poor at joint probabilities)

- The number of probabilities can be greatly reduced by exploring the absolute and conditional independence relationships among the variables.

I. Knowledge in an Uncertain Domain

- The full joint probability distribution can answer any question, but it also has several drawbacks:
  - exponential in the number n of variables and intractable as n grows very large
  - unnatural and tedious to specify probabilities of outcomes one by one
  - inadequate for representing human reasoning (good at conditional probabilities but poor at joint probabilities)
- The number of probabilities can be greatly reduced by exploring the absolute and conditional independence relationships among the variables.
- These dependencies can be *concisely represented* by a Bayesian network, which can represent any full joint probability distribution.

Bayesian Network

A *Bayesian network* (aka a *Bayes net*) is a directed acyclic graph (DAG) such that

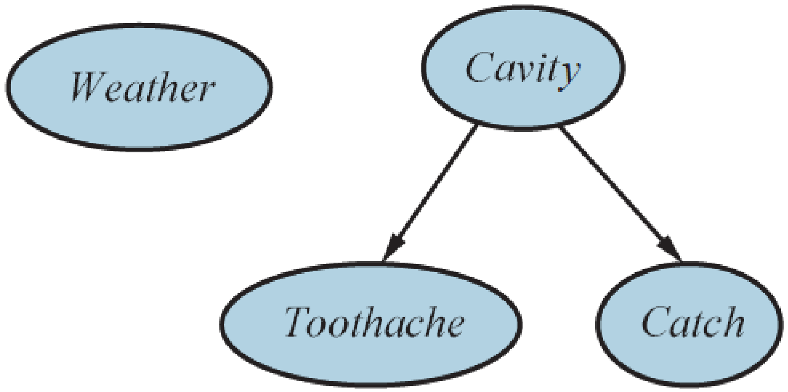
a) every node corresponds to a random variable, either discrete or continuous;

b) every edge (X, Y) specifies X (a cause) as a parent of Y (an effect);

c) every node X has associated probability information $\theta(X \mid \mathit{parent}(X))$ that quantifies the effect of the parents on X.

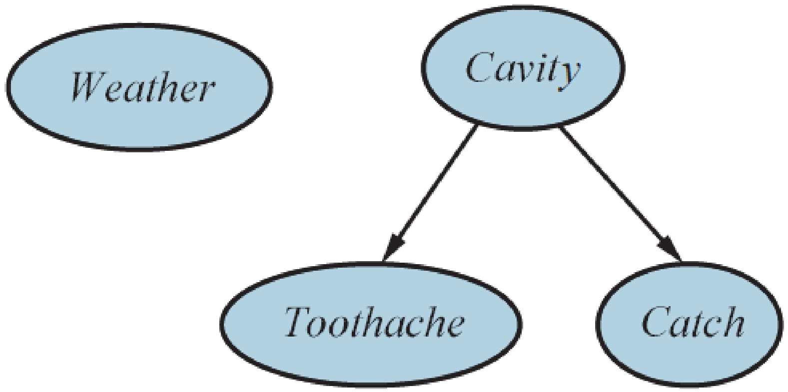
The network topology specifies the conditional independence relationships that hold in the domain.

BN as a Modeling Tool



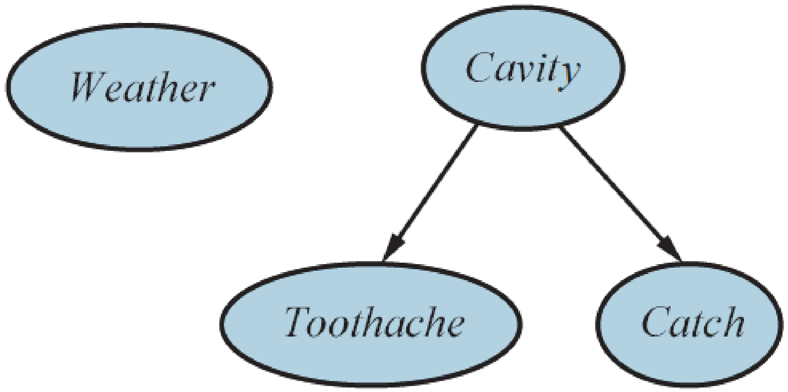
- The parents of a node X are those judged to be direct causes of X or have direct influence on X.

BN as a Modeling Tool



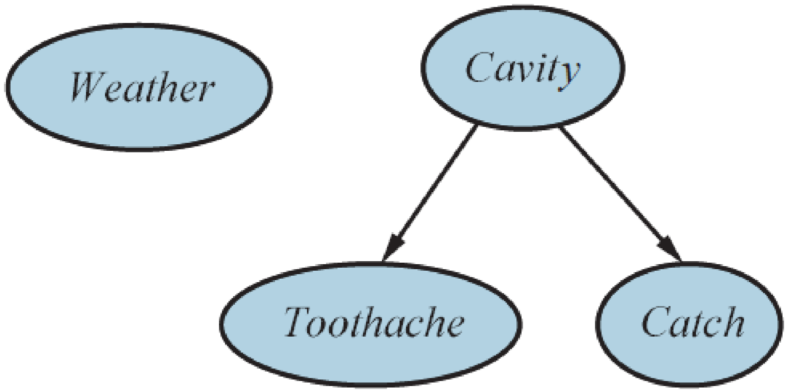
- The parents of a node X are those judged to be direct causes of X or have direct influence on X.
  - *Weather* is independent of the other three variables.

BN as a Modeling Tool



- The parents of a node X are those judged to be direct causes of X or have direct influence on X.
  - *Weather* is independent of the other three variables.
  - *Toothache* and *Catch* are conditionally dependent on *Cavity*, but conditionally independent of each other.

BN as a Modeling Tool

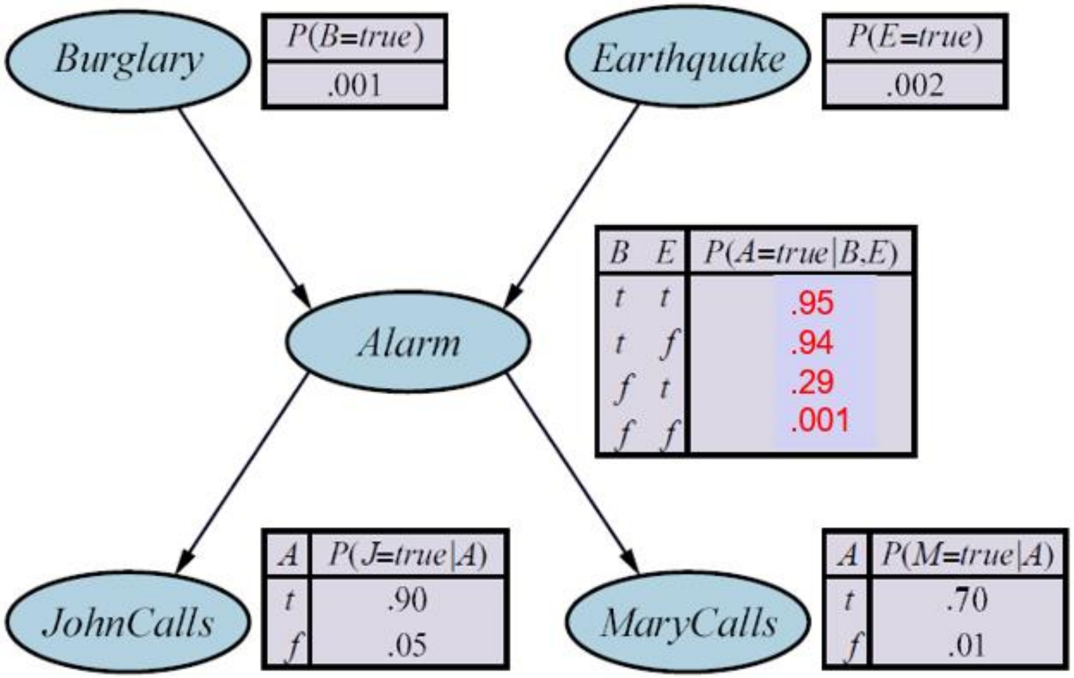


- The parents of a node X are those judged to be direct causes of X or have direct influence on X.
  - *Weather* is independent of the other three variables.
  - *Toothache* and *Catch* are conditionally dependent on *Cavity*, but conditionally independent of each other.

- The *parameters* required for model construction are *conditional probabilities* that quantify cause-effect relations, which are
  - psychologically meaningful
  - often measurable

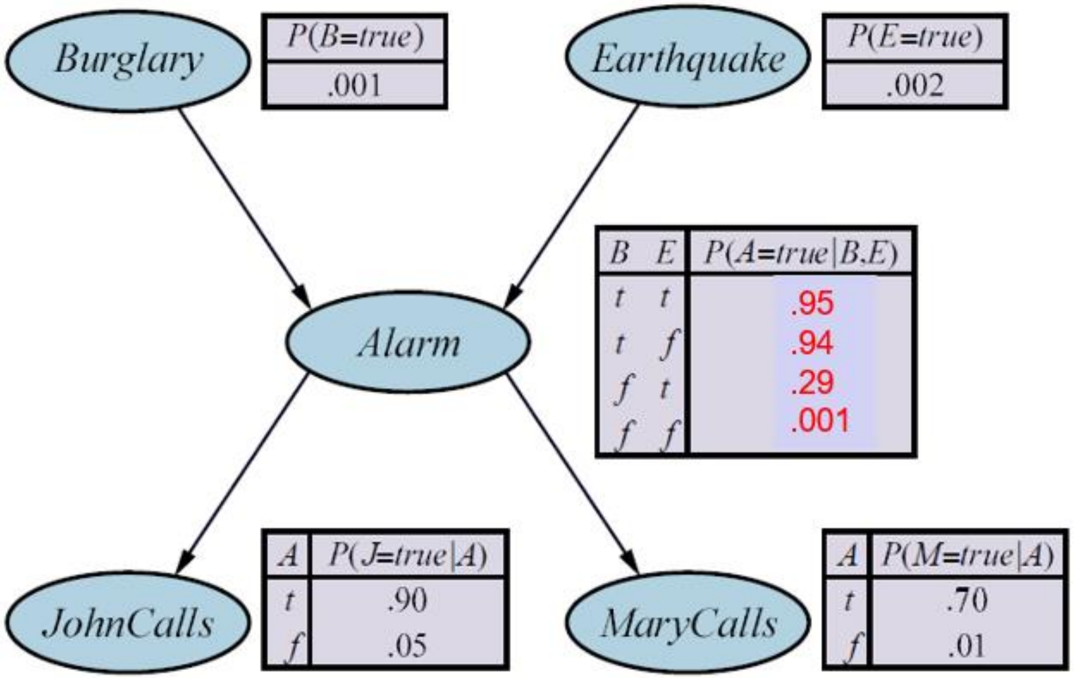
Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.



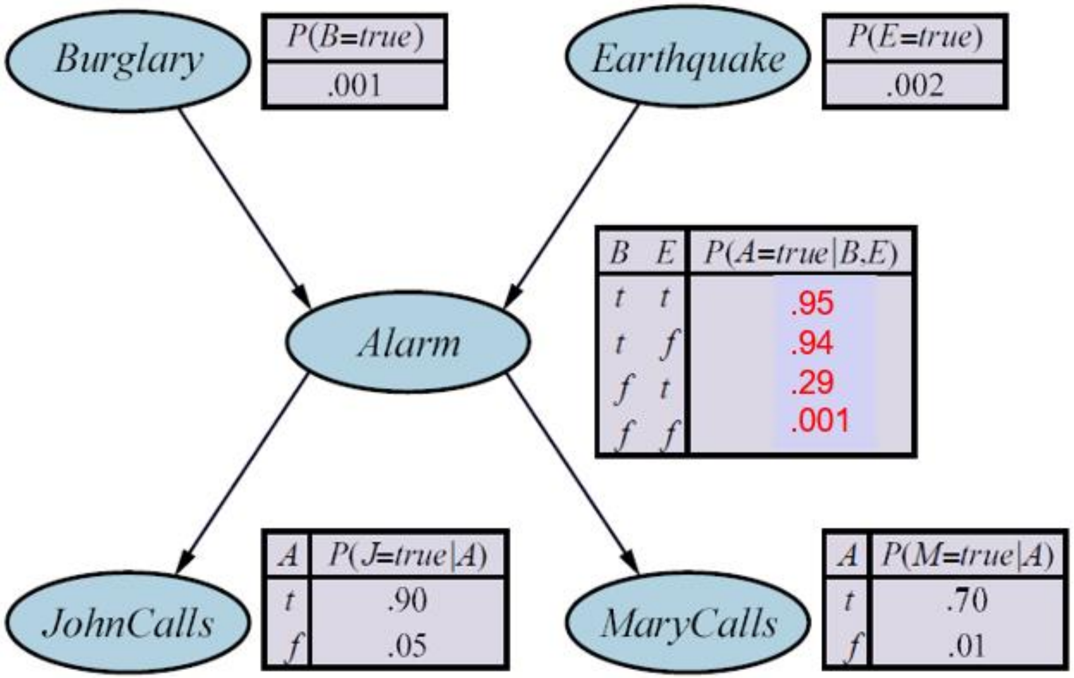
Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.



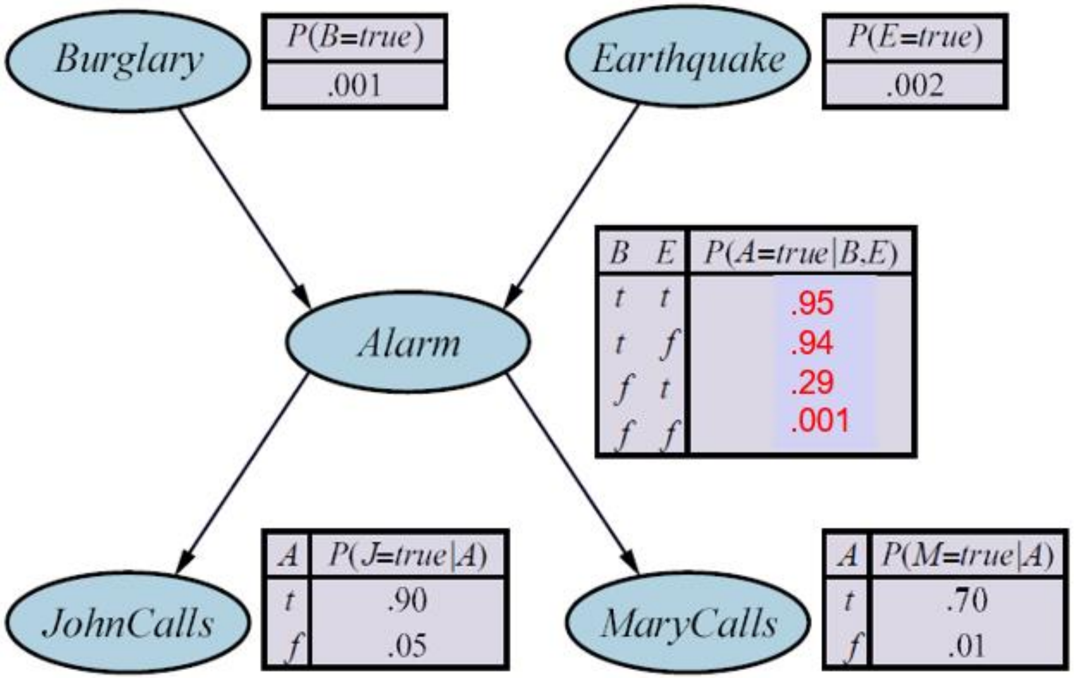
Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.



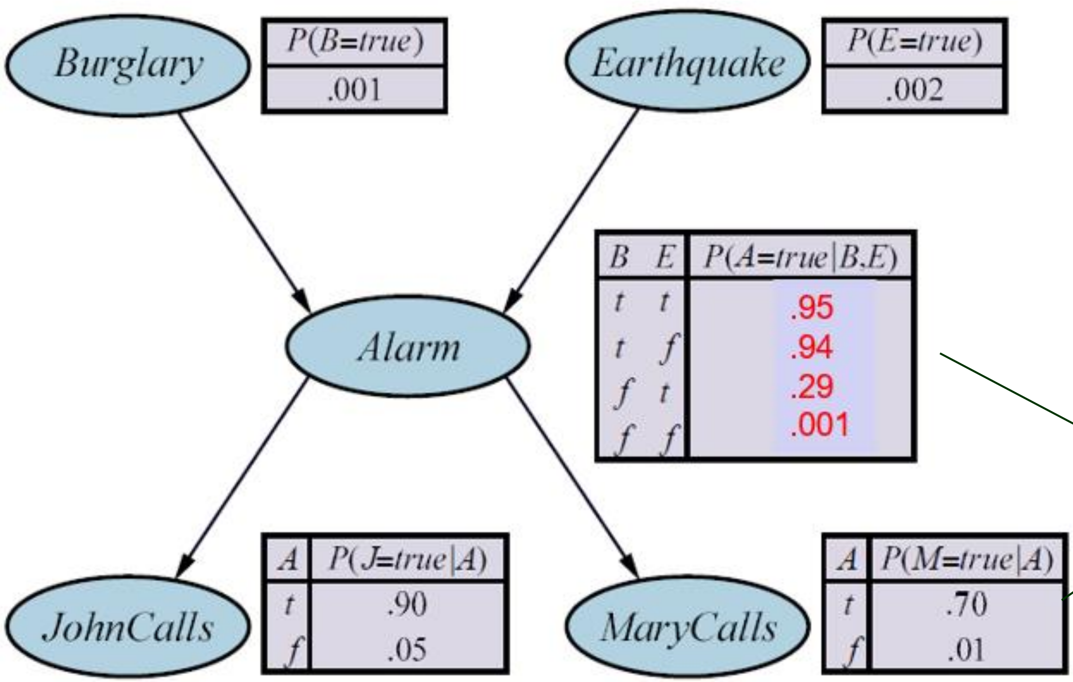
Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.
  - John nearly always calls but sometimes confuses the alarm with the telephone ringing.
  - Mary often misses the alarm because she likes playing loud music.



Burglar Alarm Problem

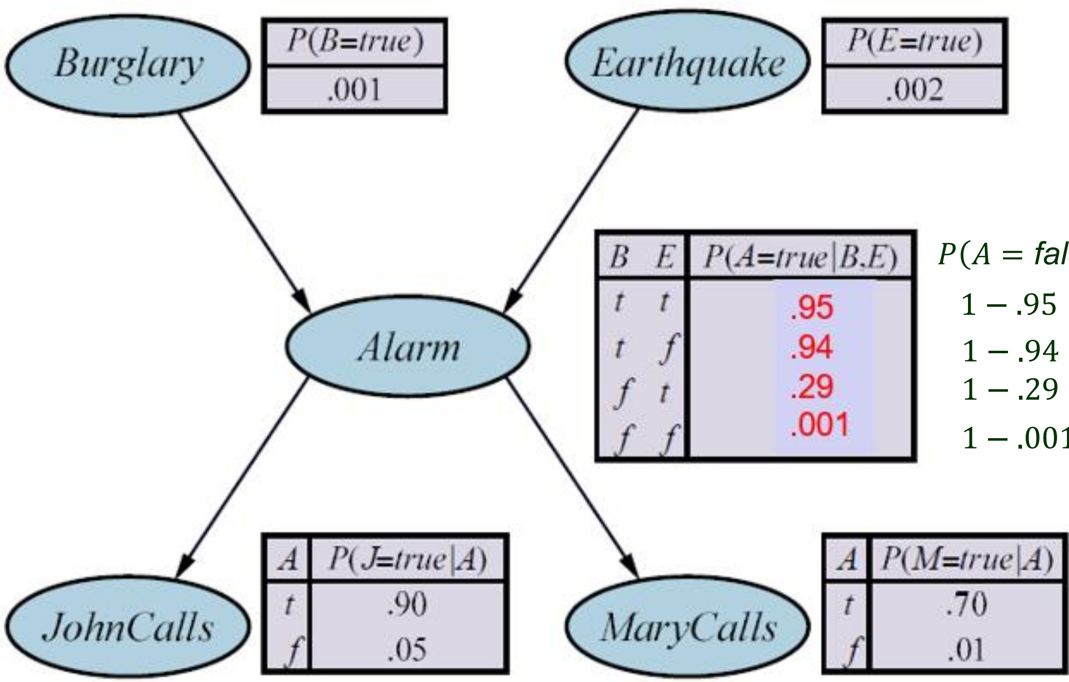
- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.
  - John nearly always calls but sometimes confuses the alarm with the telephone ringing.
  - Mary often misses the alarm because she likes playing loud music.



conditional probability tables (CPTs)

Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.
  - ♣ John nearly always calls but sometimes confuses the alarm with the telephone ringing.
  - ♣ Mary often misses the alarm because she likes playing loud music.



P(B=true)
.001

P(E=true)
.002

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

$P(A = \textit{false} \mid B, E)$

$1 - .95 = .05$

$1 - .94 = .06$

$1 - .29 = .71$

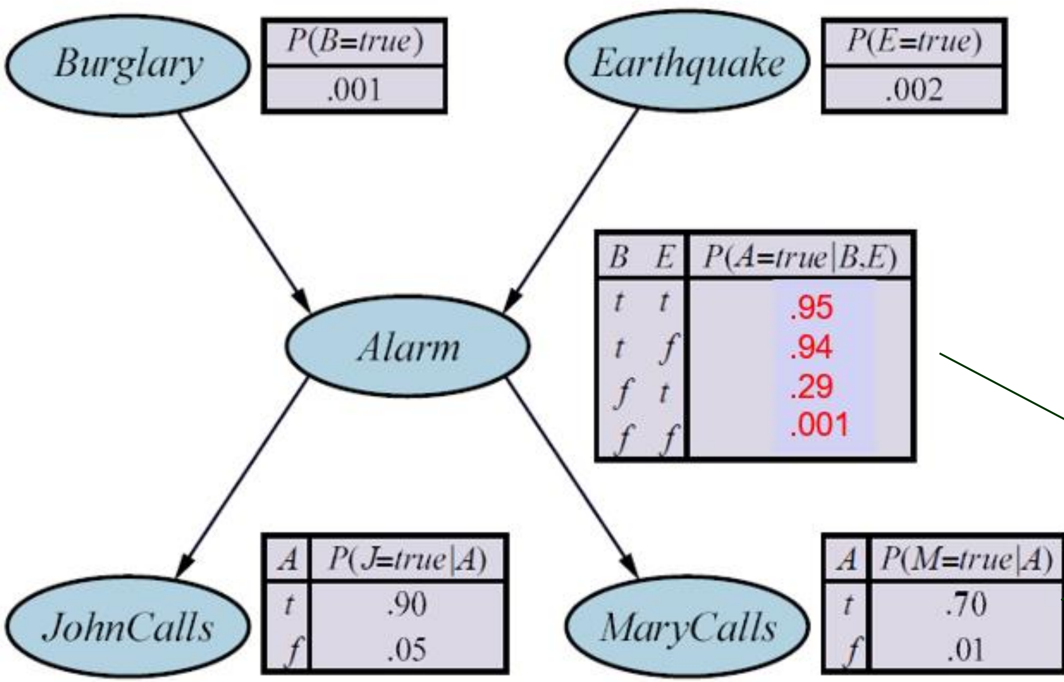
$1 - .001 = .999$

A	P(J=true|A)
t	.90
f	.05

A	P(M=true|A)
t	.70
f	.01

Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.
  - John nearly always calls but sometimes confuses the alarm with the telephone ringing.
  - Mary often misses the alarm because she likes playing loud music.



P(B=true)
.001

P(E=true)
.002

B	E	P(A=true|B,E)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

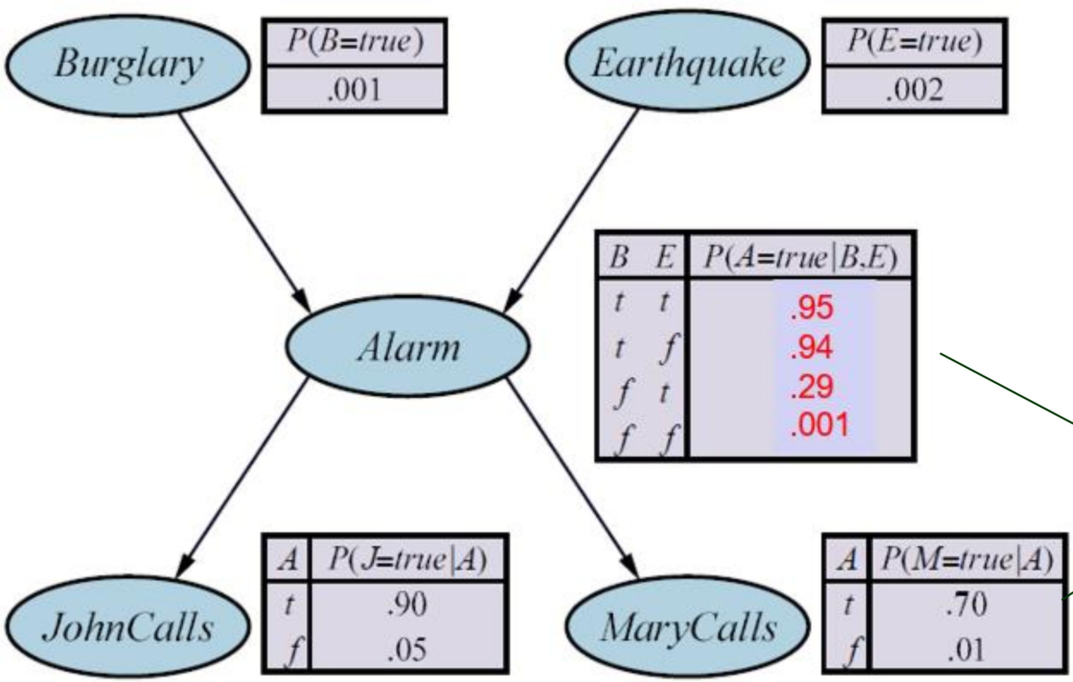
A	P(J=true|A)
t	.90
f	.05

A	P(M=true|A)
t	.70
f	.01

conditional probability tables (CPTs)

Burglar Alarm Problem

- A newly installed burglar alarm is fairly reliable at detecting a burglary.
- But it can also be occasionally set off by earthquakes.
- Neighbors John and Mary have promised a call when they hear the alarm.
  - John nearly always calls but sometimes confuses the alarm with the telephone ringing.
  - Mary often misses the alarm because she likes playing loud music.



Problem Estimate the probability of a burglary given the evidence of who has or has not called.

conditional probability tables (CPTs)

Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \ldots, X_n$ in the network

Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \dots, X_n$ in the network
- an entry in the joint distribution is defined as

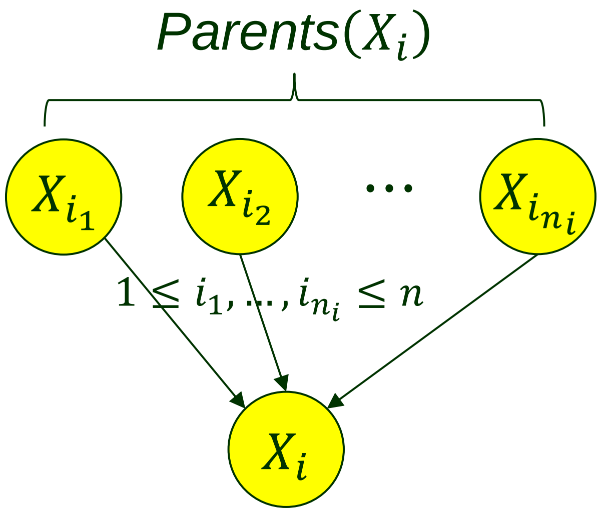
$$P(x_1, \dots, x_n) \quad = P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$$

Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \dots, X_n$ in the network
- an entry in the joint distribution is defined as

$$P(x_1, \dots, x_n) \quad = P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$$



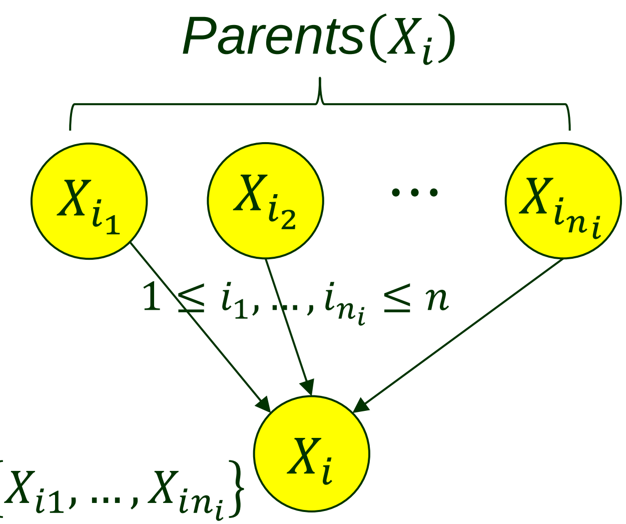
Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \dots, X_n$ in the network
- an entry in the joint distribution is defined as

$$P(x_1, \dots, x_n) \quad = P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$$

$$\equiv \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$



where

$$parents(X_i) = \{x_j \mid X_j \in Parents(X_i)\} = \{X_{i1}, \dots, X_{in_i}\}$$

// the values of *Parents*(X_i) that appear in $x_1, \dots, x_n$

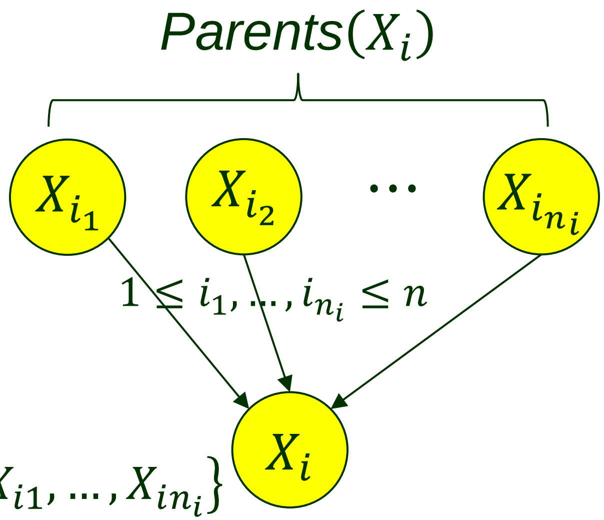
Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \ldots, X_n$ in the network
- an entry in the joint distribution is defined as

$$P(x_1, \ldots, x_n) = P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$$

$$\equiv \prod_{i=1}^{n} \theta(x_i \mid \mathit{parents}(X_i))$$



where

$\mathit{parents}(X_i) = \{x_j \mid X_j \in \mathit{Parents}(X_i)\} = \{X_{i1}, \ldots, X_{in_i}\}$

// the values of $\mathit{Parents}(X_i)$ that appear in $x_1, \ldots, x_n$

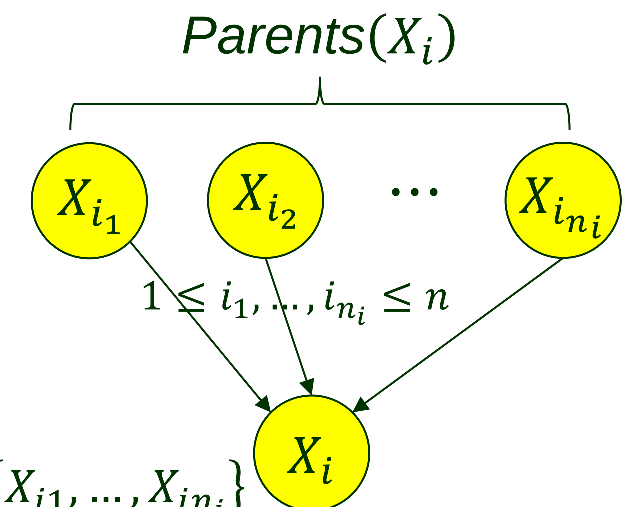
$\theta(x_i \mid \mathit{parents}(X_i))$

// probability of $X_i = x_i$ given the values of the parents of X_i

Semantics of a Bayes Net

How does the syntax correspond to a joint distribution over the variables?

- n variables $X_1, \ldots, X_n$ in the network
- an entry in the joint distribution is defined as



$$P(x_1, \ldots, x_n) = P(X_1 = x_1 \wedge \cdots \wedge X_n = x_n)$$

$$\equiv \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

where

$parents(X_i) = \{x_j \mid X_j \in Parents(X_i)\} = \{X_{i1}, \ldots, X_{in_i}\}$

// the values of $Parents(X_i)$ that appear in $x_1, \ldots, x_n$

$\theta(x_i \mid parents(X_i))$

// probability of $X_i = x_i$ given the values of the parents of X_i

Every entry in the joint distribution is the product of the appropriate elements of the local conditional distribution.

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

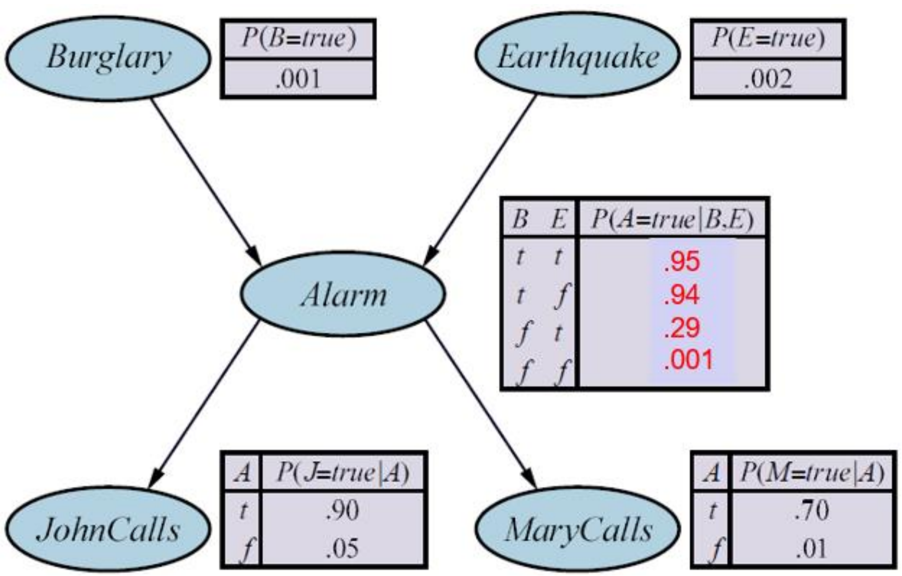
Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.

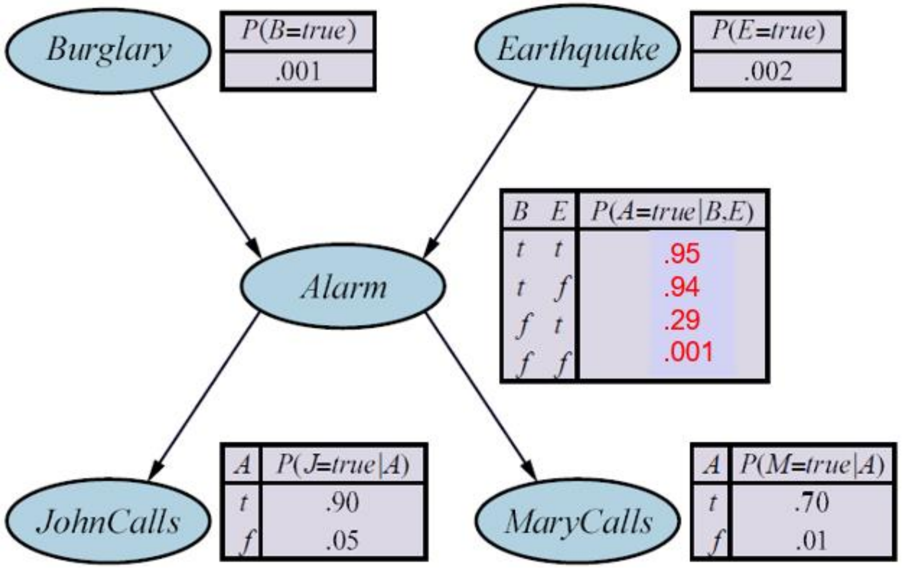
$$P(j, m, a, \neg b, \neg e)$$



BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.



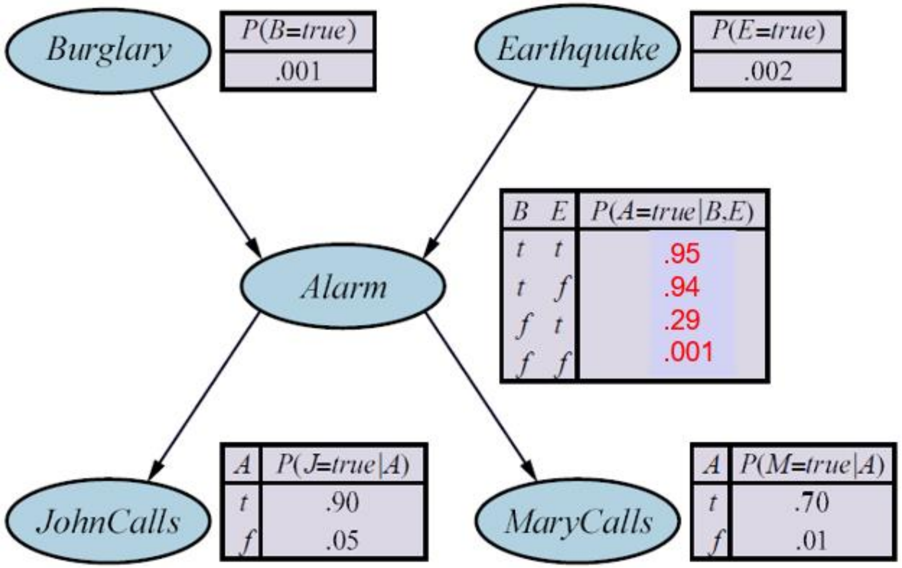
$$P(j, m, a, \neg b, \neg e)$$

$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e)$$

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.



$$P(j, m, a, \neg b, \neg e)$$

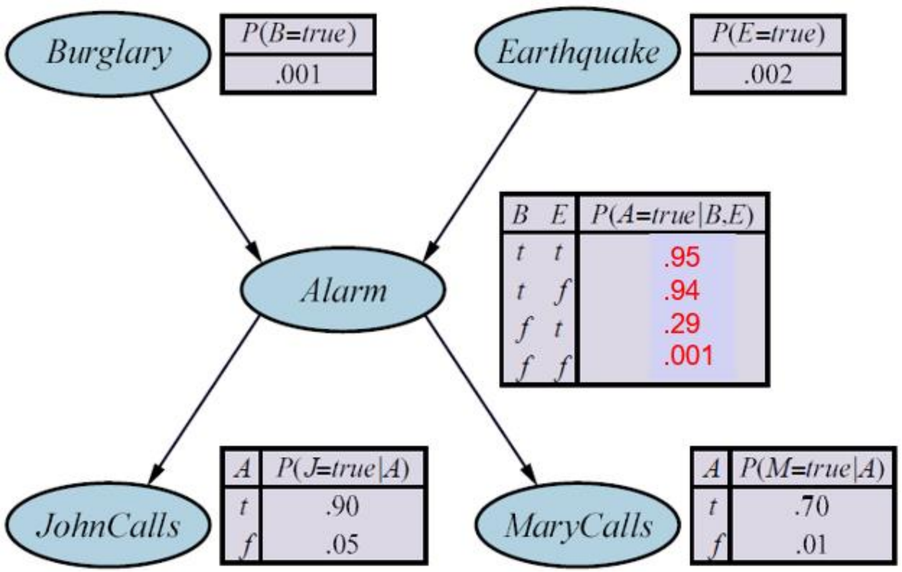
$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e)$$

Alarm is the sole parent of JohnCalls.

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.



$$P(j, m, a, \neg b, \neg e)$$

$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e)$$

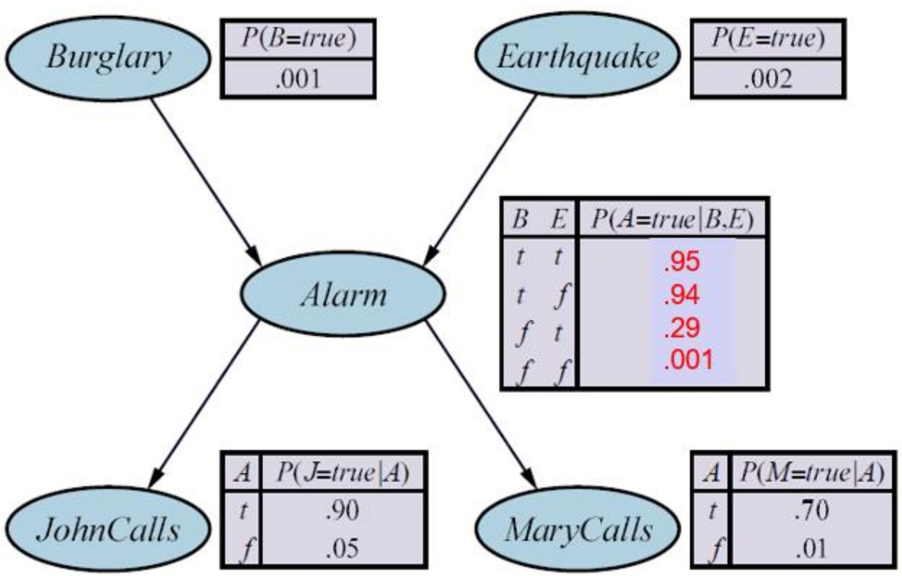
Alarm is the sole parent of JohnCalls.

Burglary and Earthquake are the only two parents of Alarm.

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.



$$P(j, m, a, \neg b, \neg e)$$

$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e)$$

Alarm is the sole parent of JohnCalls.

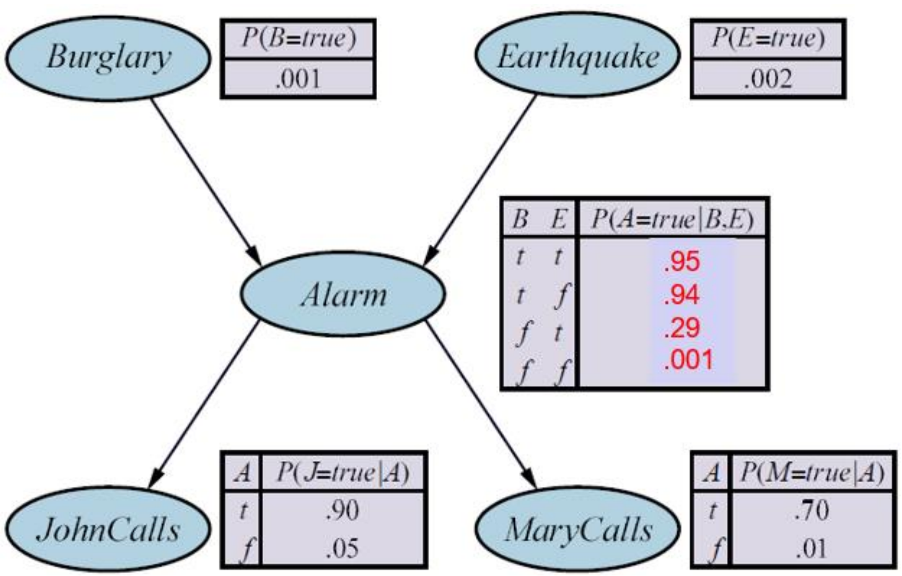
Burglary and Earthquake are the only two parents of Alarm.

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$$

BN as a Knowledge Base

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occurred, and both John and Mary call.



$$P(j, m, a, \neg b, \neg e)$$

$$= P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e) P(\neg b) P(\neg e)$$

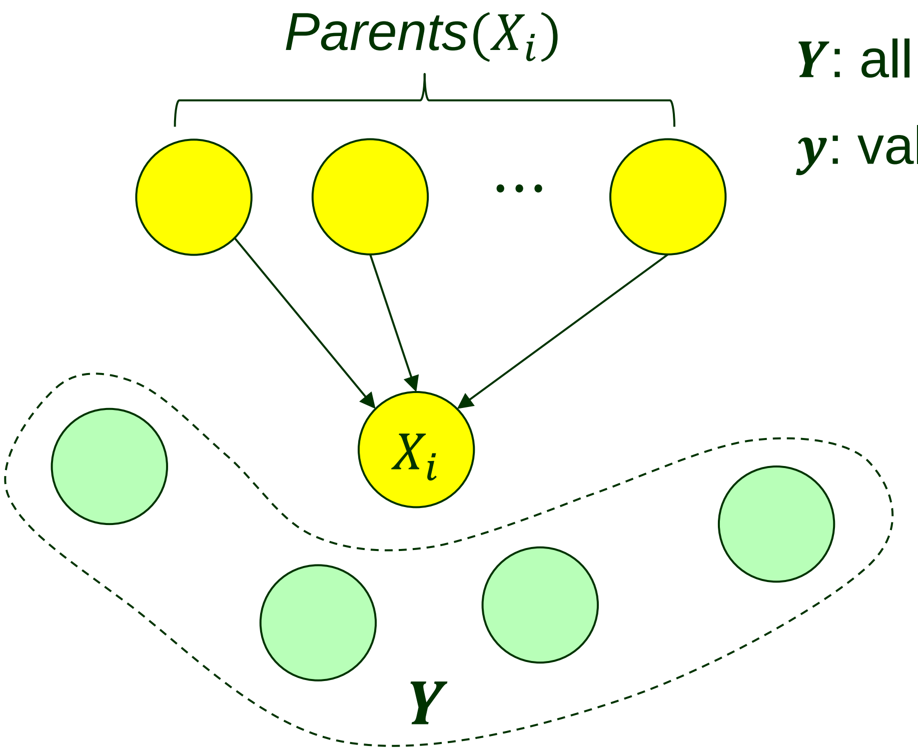
Alarm is the sole parent of JohnCalls.

Burglary and Earthquake are the only two parents of Alarm.

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$$

$$= 0.000628$$

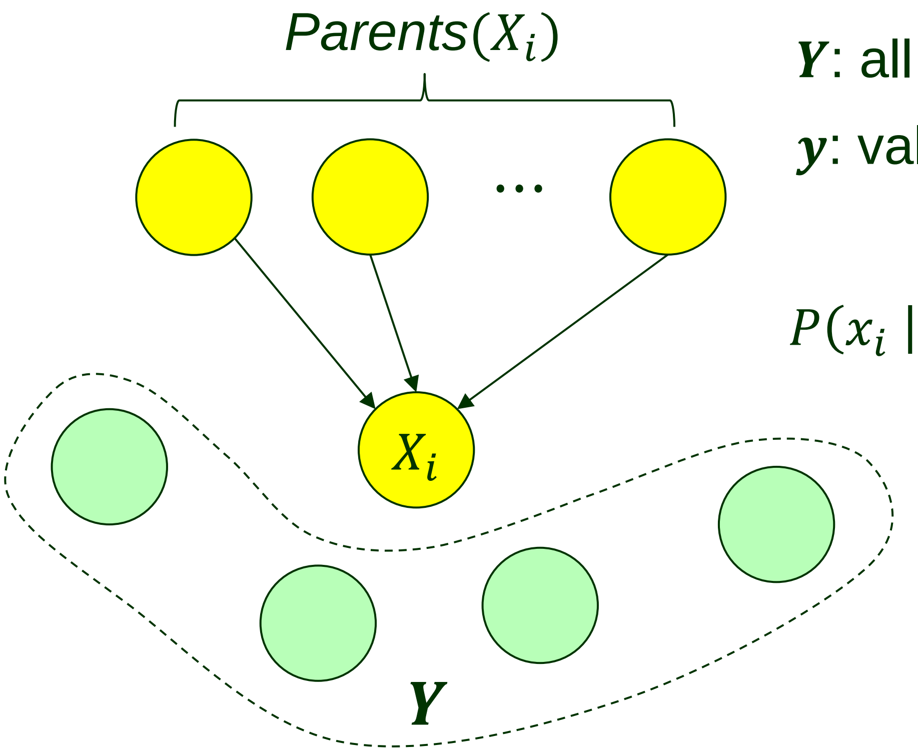
Conditional Probabilities



$\boldsymbol{Y}$: all variables other than X_i and *Parents*(X_i)

$\boldsymbol{y}$: values of $\boldsymbol{Y}$

Conditional Probabilities

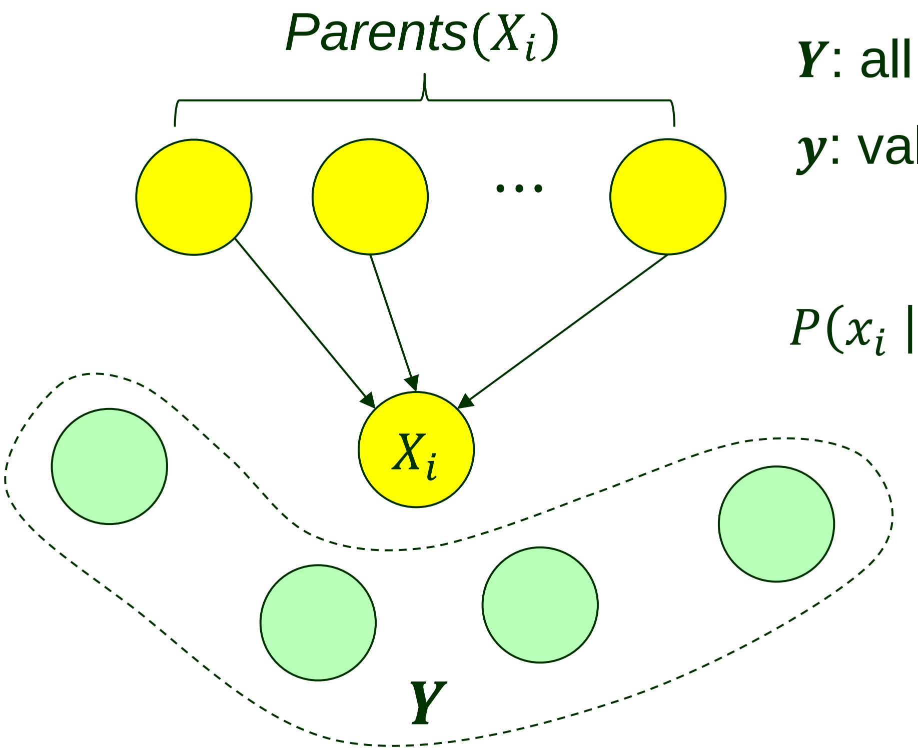


$\boldsymbol{Y}$: all variables other than X_i and $Parents(X_i)$

$\boldsymbol{y}$: values of $\boldsymbol{Y}$

$$P(x_i \mid parents(X_i)) \equiv \frac{P(x_i, parents(X_i))}{P(parents(X_i))}$$

Conditional Probabilities



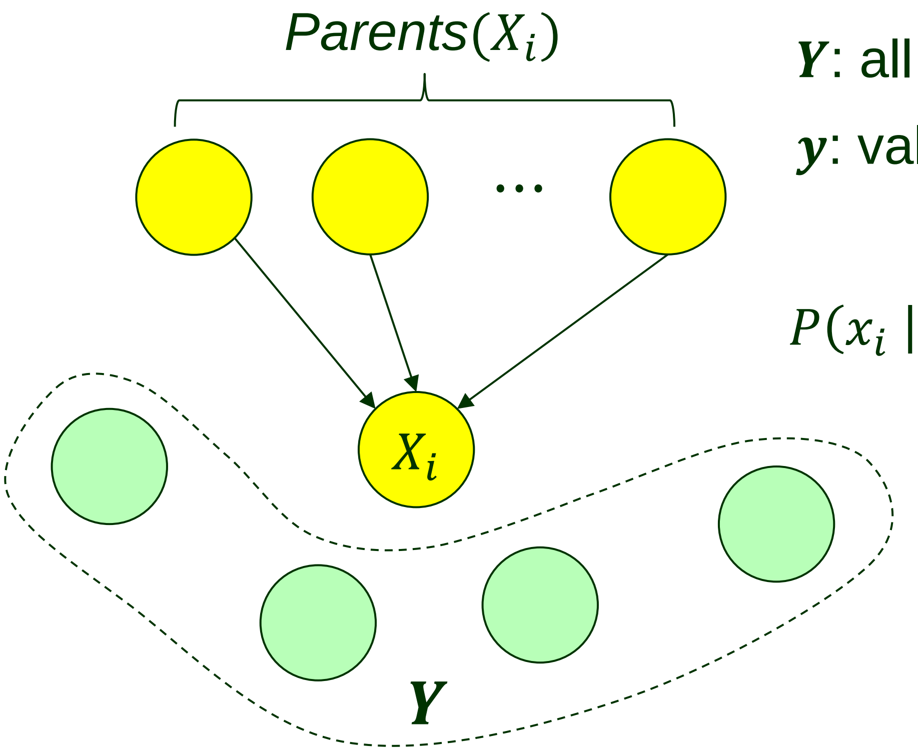
$\boldsymbol{Y}$: all variables other than X_i and $Parents(X_i)$

$\boldsymbol{y}$: values of $\boldsymbol{Y}$

$$P(x_i \mid parents(X_i)) \equiv \frac{P(x_i, parents(X_i))}{P(parents(X_i))}$$

$$= \frac{\sum_{\boldsymbol{y}} P(x_i, parents(X_i), \boldsymbol{y})}{\sum_{x_i', \boldsymbol{y}} P(x_i', parents(X_i), \boldsymbol{y})}$$

Conditional Probabilities



$\boldsymbol{Y}$: all variables other than X_i and $Parents(X_i)$

$\boldsymbol{y}$: values of $\boldsymbol{Y}$

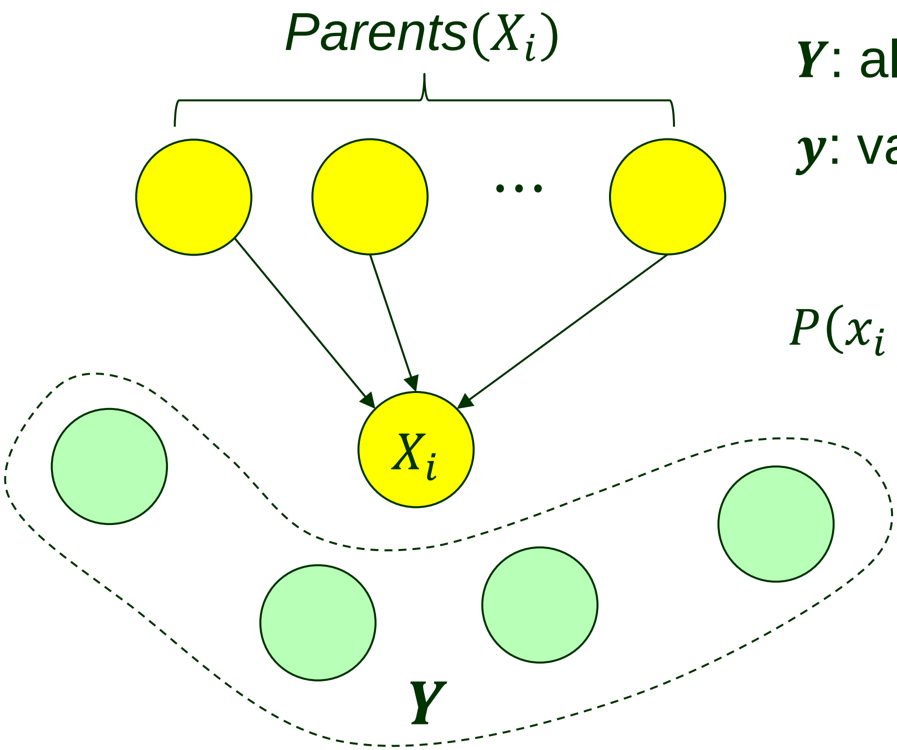
$$P(x_i \mid parents(X_i)) \equiv \frac{P(x_i, parents(X_i))}{P(parents(X_i))}$$

$$= \frac{\sum_{\boldsymbol{y}} P(x_i, parents(X_i), \boldsymbol{y})}{\sum_{x_i', \boldsymbol{y}} P(x_i', parents(X_i), \boldsymbol{y})}$$

$\vdots$ // proof can be derived

$$= \theta(x_i \mid parents(X_i))$$

Conditional Probabilities



$\boldsymbol{Y}$: all variables other than X_i and $Parents(X_i)$

$\boldsymbol{y}$: values of $\boldsymbol{Y}$

$$P(x_i \mid parents(X_i)) \equiv \frac{P(x_i, parents(X_i))}{P(parents(X_i))}$$

$$= \frac{\sum_{\boldsymbol{y}} P(x_i, parents(X_i), \boldsymbol{y})}{\sum_{x_i', \boldsymbol{y}} P(x_i', parents(X_i), \boldsymbol{y})}$$

⋮ // proof can be derived

$$= \theta(x_i \mid parents(X_i))$$

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} \theta(x_i \mid parents(X_i))$$

(by definition of the Bayes net)

Full joint distribution: $P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$

Correct Domain Representation

Chain rule:

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

Correct Domain Representation

Chain rule:

$$
\begin{aligned}
P(x_1, \dots, x_n) &= P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1}, \dots, x_1) \\
&\vdots \\
&= P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1} \mid x_{n-2}, \dots, x_1) \cdots P(x_2 \mid x_1) P(x_1)
\end{aligned}
$$

Correct Domain Representation

Chain rule:

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

Correct Domain Representation

Chain rule:

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

Meanwhile,

$$
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))
$$

Correct Domain Representation

Chain rule:

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

$$\vdots$$

$$= P(x_n \mid x_{n-1}, \ldots, x_1) P(x_{n-1} \mid x_{n-2}, \ldots, x_1) \cdots P(x_2 \mid x_1) P(x_1)$$

$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1)$$

Meanwhile,

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1}, \ldots, x_1) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

Correct Domain Representation

Chain rule:

$$P(x_1, \dots, x_n) = P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1}, \dots, x_1)$$
$$\vdots$$
$$= P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1} \mid x_{n-2}, \dots, x_1) \cdots P(x_2 \mid x_1) P(x_1)$$
$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \dots, x_1)$$

Meanwhile,

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1}, \dots, x_1) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

Joint distribution is specified by the BN if, for every X_i,

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

Correct Domain Representation

Chain rule:

$$P(x_1, \dots, x_n) = P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1}, \dots, x_1)$$
$$\vdots$$
$$= P(x_n \mid x_{n-1}, \dots, x_1) P(x_{n-1} \mid x_{n-2}, \dots, x_1) \cdots P(x_2 \mid x_1) P(x_1)$$
$$= \prod_{i=1}^{n} P(x_i \mid x_{i-1}, \dots, x_1)$$

Meanwhile,

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

$$\prod_{i=1}^{n} P(x_i \mid x_{i-1}, \dots, x_1) = \prod_{i=1}^{n} P(x_i \mid \textit{parents}(X_i))$$

Joint distribution is specified by the BN if, for every X_i,

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

This is guaranteed when $\textit{Parents}(X_i) \subseteq \{X_1, \dots X_{i-1}\}$ for every i.

Topological Order

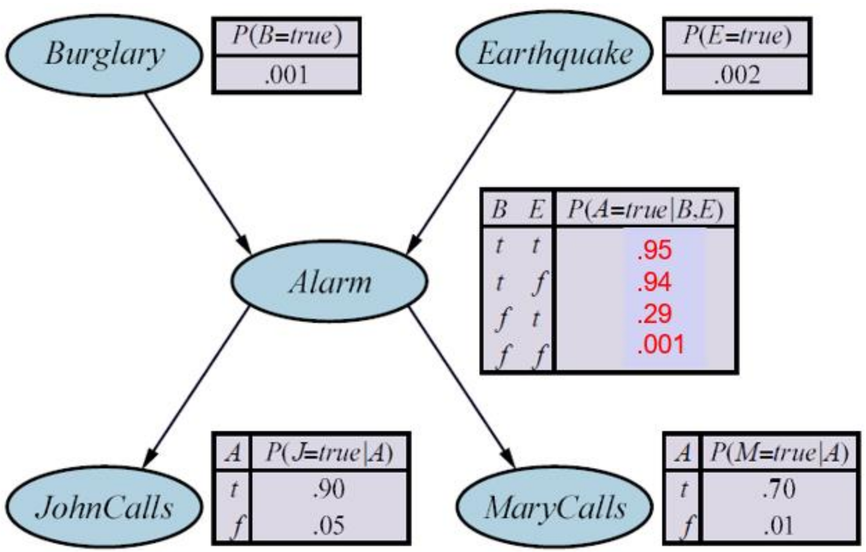
$$\textit{Parents}(X_i) \subseteq \{X_1, \dots X_{i-1}\} \text{ for } i = 2, \dots, n$$

The above is guaranteed if we number the nodes in *topological order* (which exists since the Bayesian network is a DAG).

Topological Order

$Parents(X_i) \subseteq \{X_1, \dots X_{i-1}\}$ for $i = 2, \dots, n$

The above is guaranteed if we number the nodes in *topological order* (which exists since the Bayesian network is a DAG).



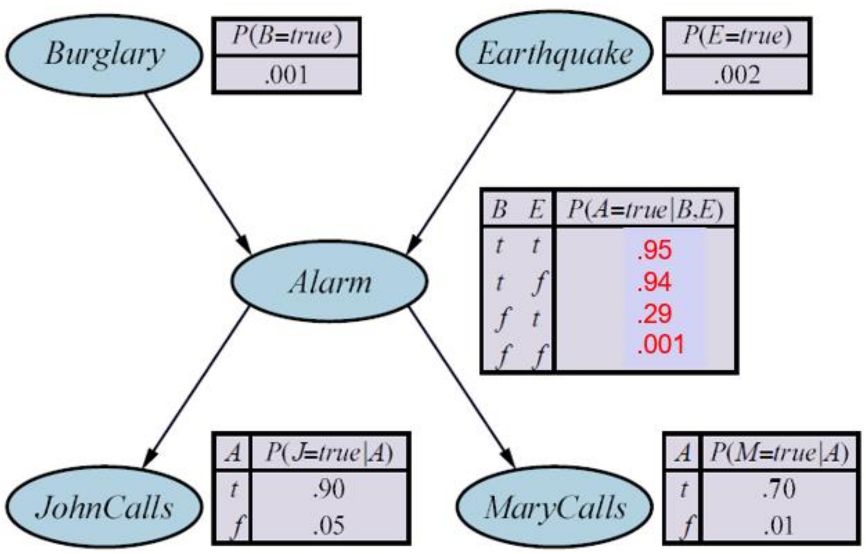
Four topological orders:

B, E, A, J, M
B, E, A, M, J
E, B, A, J, M
E, B, A, M, J

Topological Order

$Parents(X_i) \subseteq \{X_1, \ldots X_{i-1}\}$ for $i = 2, \ldots, n$

The above is guaranteed if we number the nodes in *topological order* (which exists since the Bayesian network is a DAG).



Four topological orders:

B, E, A, J, M
B, E, A, M, J
E, B, A, J, M
E, B, A, M, J

Any one of the four suffices.

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \ldots, X_1) = \boldsymbol{P}(X_i \mid \textit{Parents}(X_i)) \quad \text{for } i = 2, \ldots, n$$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \leq j \leq i-1$, such that $X_j \notin \textit{Parents}(X_i)$.

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \dots, X_1) = \boldsymbol{P}(X_i \mid \textit{Parents}(X_i)) \quad \text{for } i = 2, \dots, n$$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \leq j \leq i-1$, such that $X_j \notin \textit{Parents}(X_i)$.

Construction algorithm

1. Determine the set of variables that are required to model the domain.

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \ldots, X_1) = \boldsymbol{P}(X_i \mid \mathit{Parents}(X_i)) \quad \text{for } i = 2, \ldots, n$$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \leq j \leq i-1$, such that $X_j \notin \mathit{Parents}(X_i)$.

Construction algorithm

1. Determine the set of variables that are required to model the domain.

2. Order them as $X_1, X_2, \ldots, X_n$.

   Any order works, although network compactness depends on how much the order – whether causes precede effects – is respected.

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \ldots, X_1) = \boldsymbol{P}(X_i \mid \textit{Parents}(X_i)) \quad \text{for } i = 2, \ldots, n$$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \leq j \leq i-1$, such that $X_j \notin \textit{Parents}(X_i)$.

Construction algorithm

1. Determine the set of variables that are required to model the domain.

2. Order them as $X_1, X_2, \ldots, X_n$.

   Any order works, although network compactness depends on how much the order – whether causes precede effects – is respected.

3. For $i = 1$ to n do

   a) Choose a minimal set of parents for X_i from $X_1, X_2, \ldots, X_{i-1}$ such that

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

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \ldots, X_1) = \boldsymbol{P}(X_i \mid \mathit{Parents}(X_i)) \quad \text{for } i = 2, \ldots, n$$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \le j \le i-1$, such that $X_j \notin \mathit{Parents}(X_i)$.

Construction algorithm

1. Determine the set of variables that are required to model the domain.
2. Order them as $X_1, X_2, \ldots, X_n$.

   Any order works, although network compactness depends on how much the order – whether causes precede effects – is respected.

3. For $i = 1$ to n do

   a) Choose a minimal set of parents for X_i from $X_1, X_2, \ldots, X_{i-1}$ such that

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

   b) Add a directed edge from every parent to X_i.

II. Construction of the Bayesian Network

$$\boldsymbol{P}(X_i \mid X_{i-1}, \dots, X_1) = \boldsymbol{P}(X_i \mid \textit{Parents}(X_i))$$ for $i = 2, \dots, n$

The Bayesian network is correct only if X_i is conditionally independent of any X_j, $1 \le j \le i-1$, such that $X_j \notin \textit{Parents}(X_i)$.

Construction algorithm

1. Determine the set of variables that are required to model the domain.

2. Order them as $X_1, X_2, \dots, X_n$.

   Any order works, although network compactness depends on how much the order – whether causes precede effects – is respected.

3. For $i = 1$ to n do

   a) Choose a minimal set of parents for X_i from $X_1, X_2, \dots, X_{i-1}$ such that

   $$\boldsymbol{P}(X_i \mid X_{i-1}, \dots, X_1) = \boldsymbol{P}(X_i \mid \textit{Parents}(X_i))$$

   b) Add a directed edge from every parent to X_i.

   c) Write down the conditional probability table (CPT), $\boldsymbol{P}(X_i \mid \textit{Parents}(X_i))$.

Construction (cont'd)

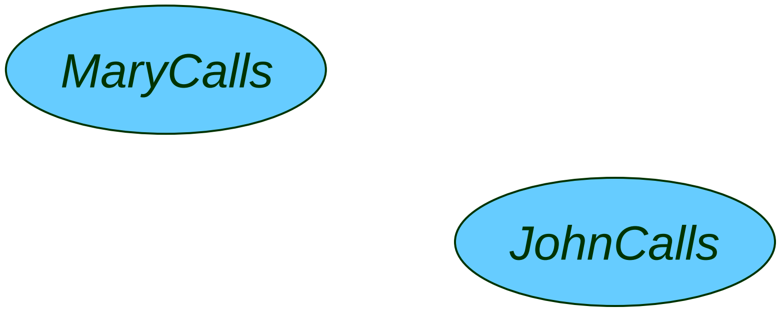
Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

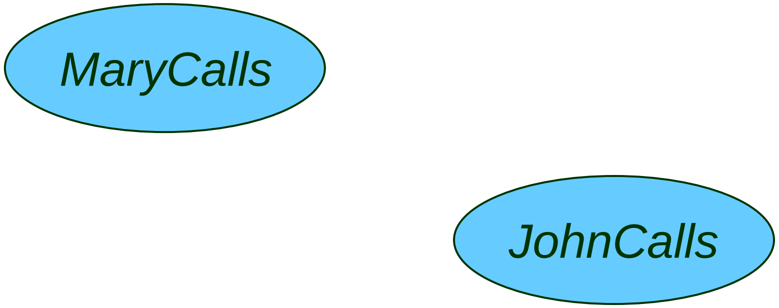
Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



Construction (cont'd)

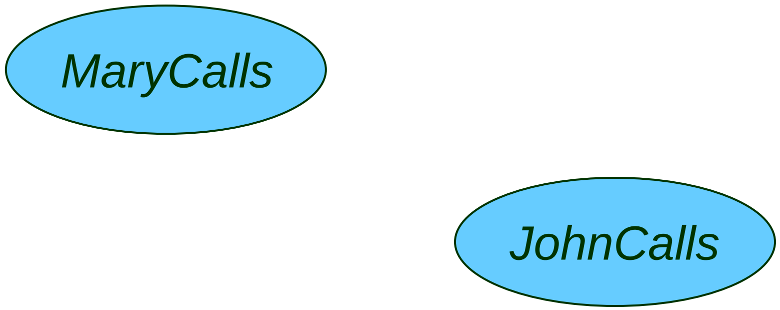
Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$P(j \mid m)$ $P(j)$

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

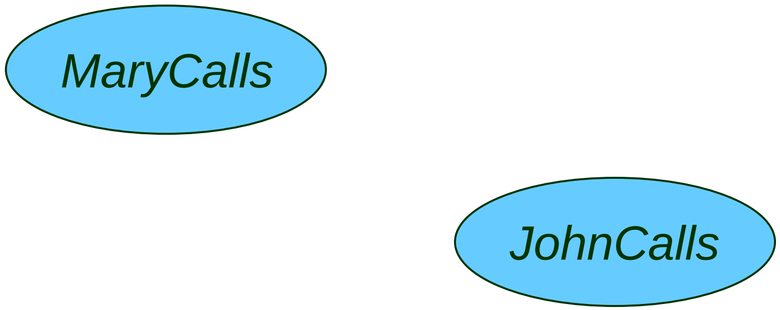


$P(j \mid m)$ $P(j)$

// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

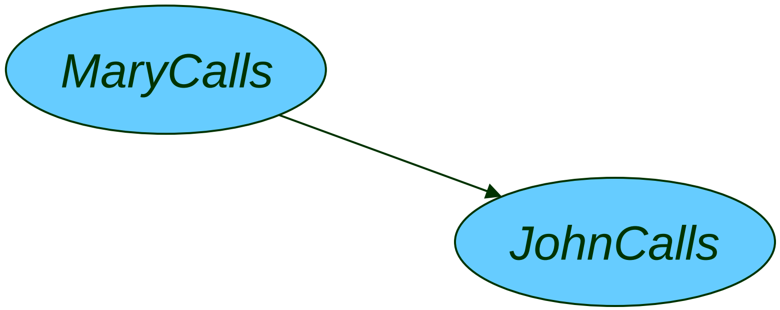


$$P(j \mid m) > P(j)$$

// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

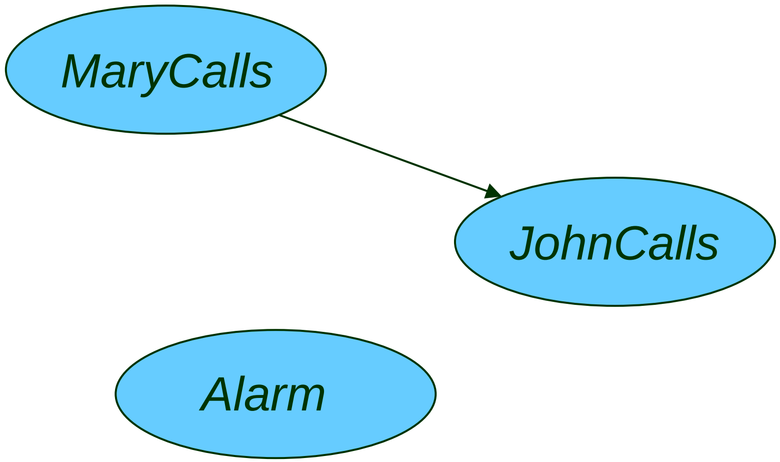


$$P(j \mid m) > P(j)$$

// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

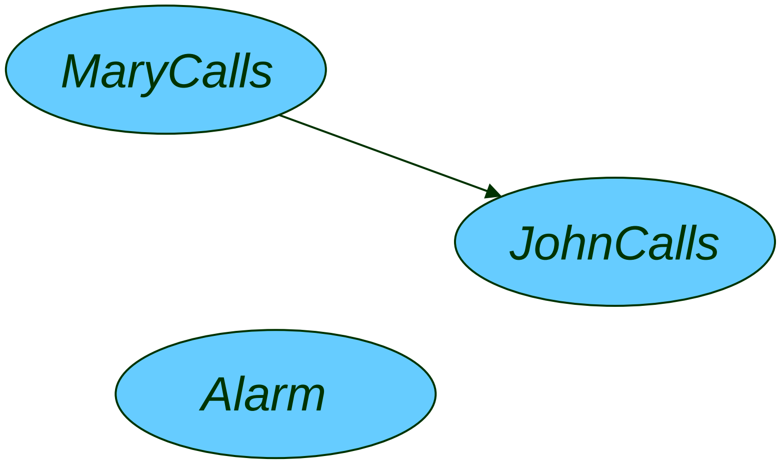


$$P(j \mid m) > P(j)$$

// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



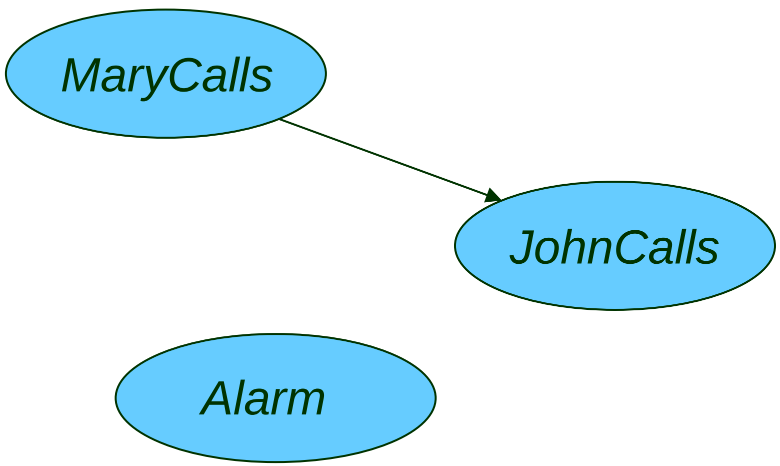
$P(j \mid m) > P(j)$

// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

$P(a \mid m, j) \quad P(a \mid j), P(a \mid m), P(a)$

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$P(j \mid m) > P(j)$

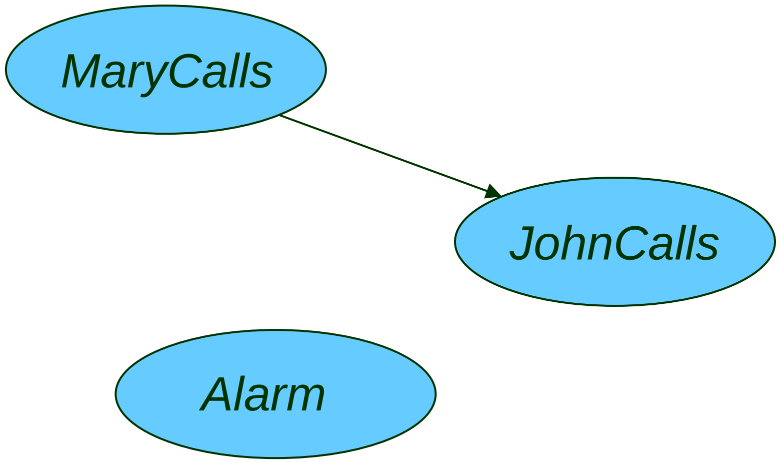
// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

$P(a \mid m, j) \quad P(a \mid j), P(a \mid m), P(a)$

// If both Mary and John call, the alarm
// is more likely to go off than if just
// one calls.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$$P(j \mid m) > P(j)$$

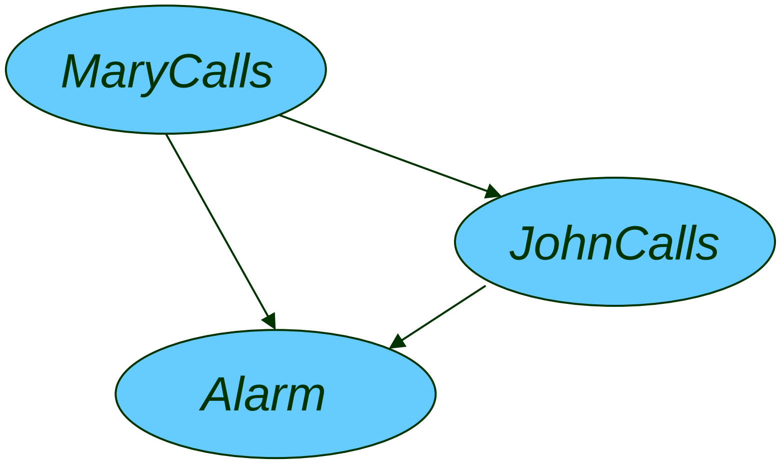
// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

$$P(a \mid m, j) > P(a \mid j), P(a \mid m), P(a)$$

// If both Mary and John call, the alarm
// is more likely to go off than if just
// one calls.

Construction (cont'd)

Chosen order: *MaryCalls*, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$P(j \mid m) > P(j)$

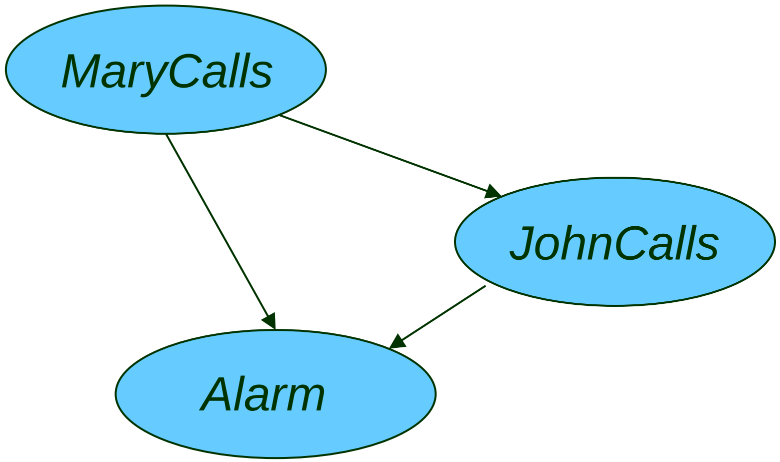
// If May calls, that probably means
// the alarm has gone off, which
// makes John more likely to call.

$P(a \mid m, j) > P(a \mid j), P(a \mid m), P(a)$

// If both Mary and John call, the alarm
// is more likely to go off than if just
// one calls.

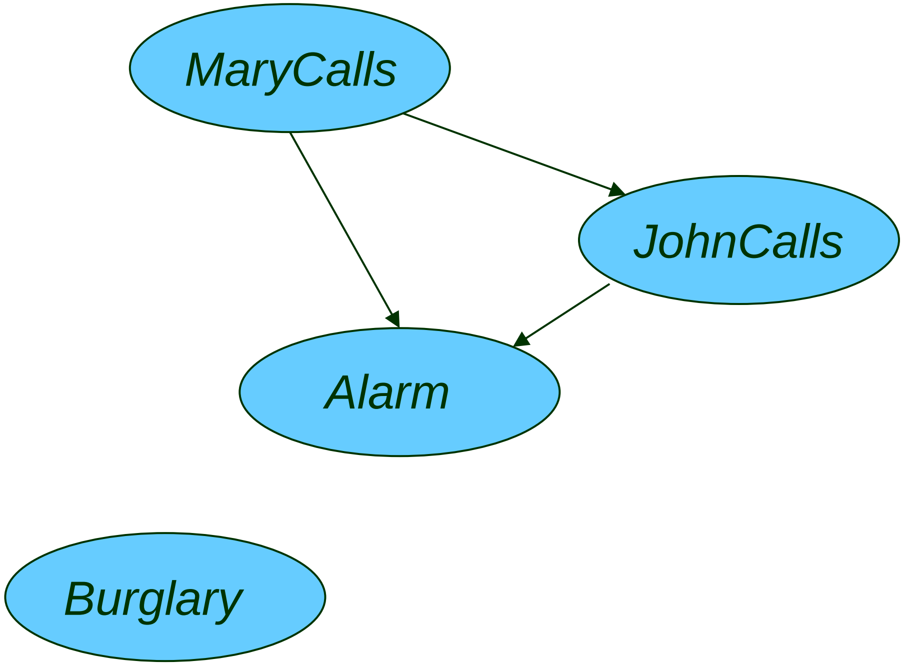
Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



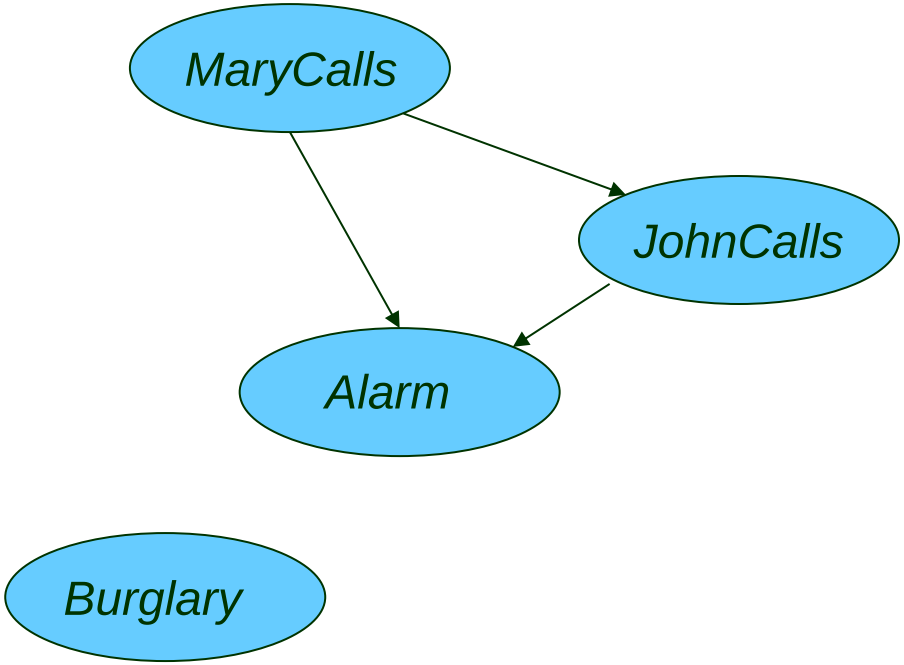
Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

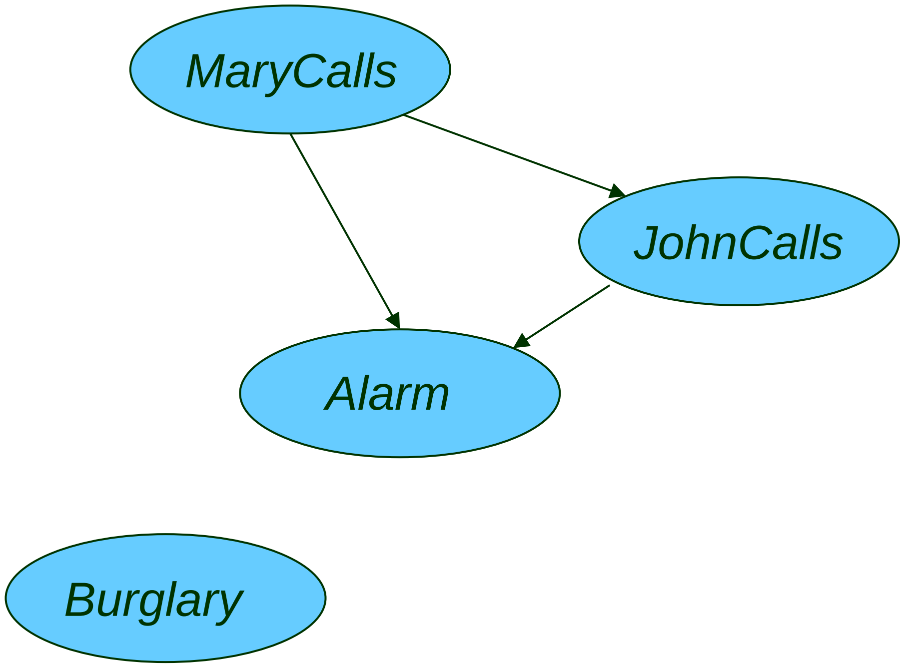


$\boldsymbol{P}(B \mid A, J, M)$ $\boldsymbol{P}(B \mid A)$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

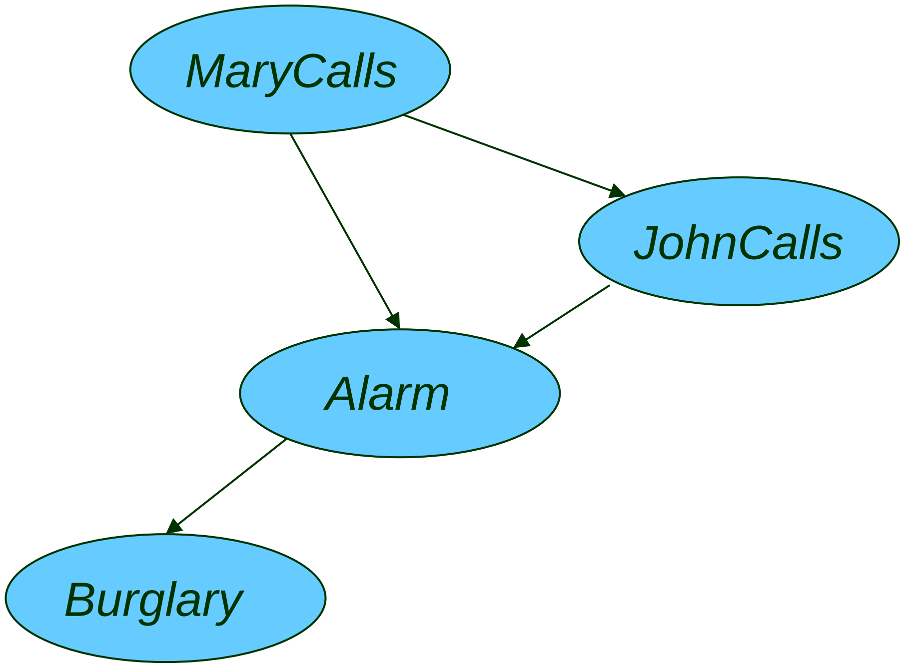


$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

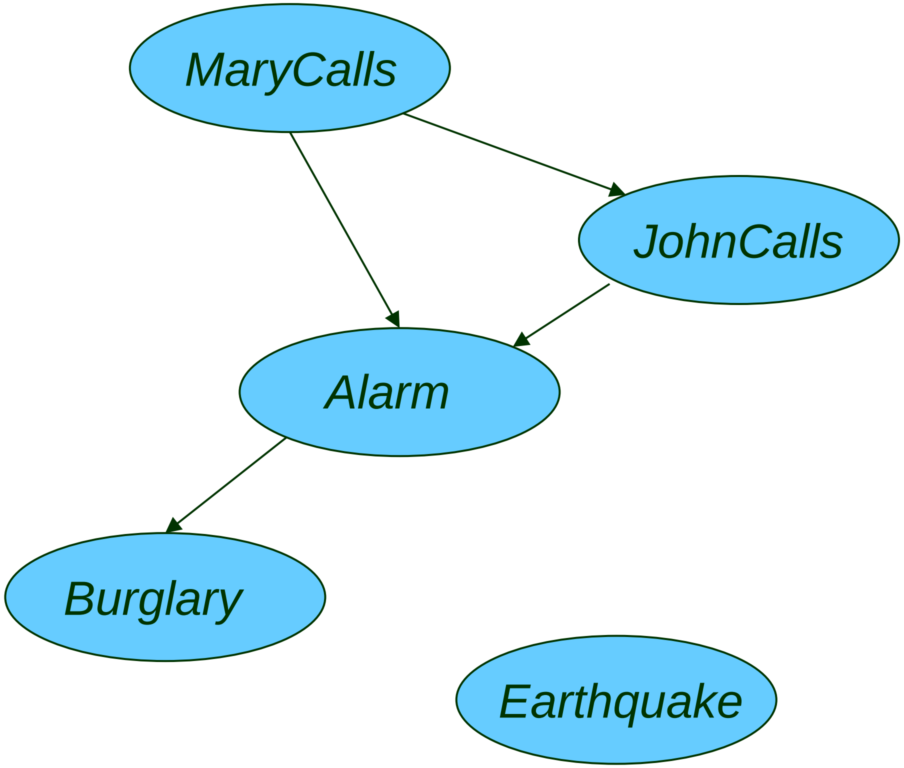


$$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.

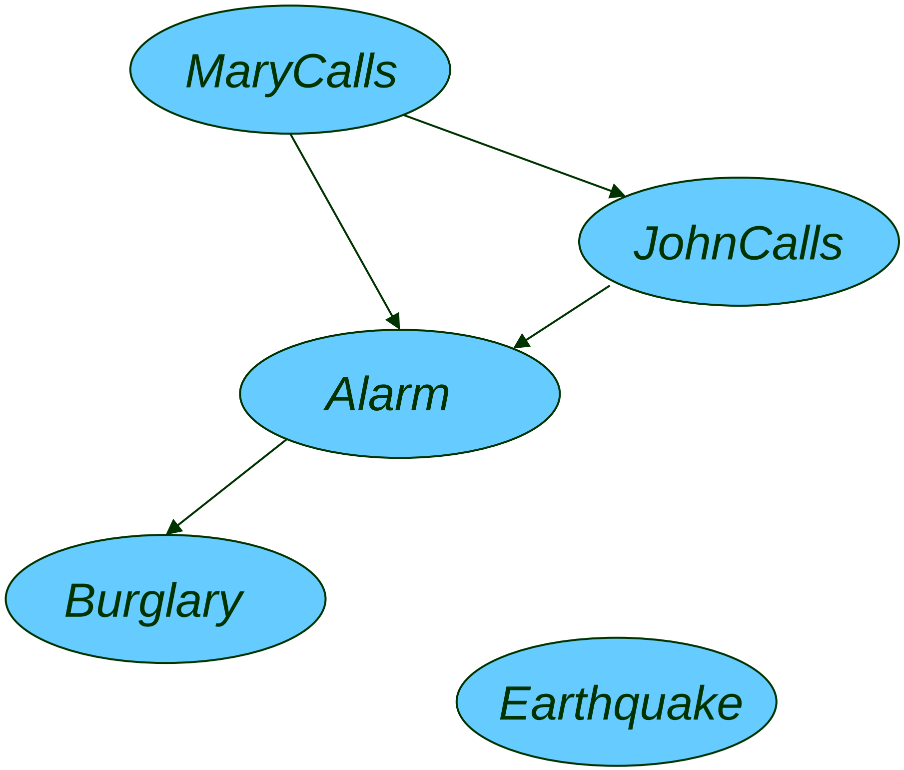


$$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



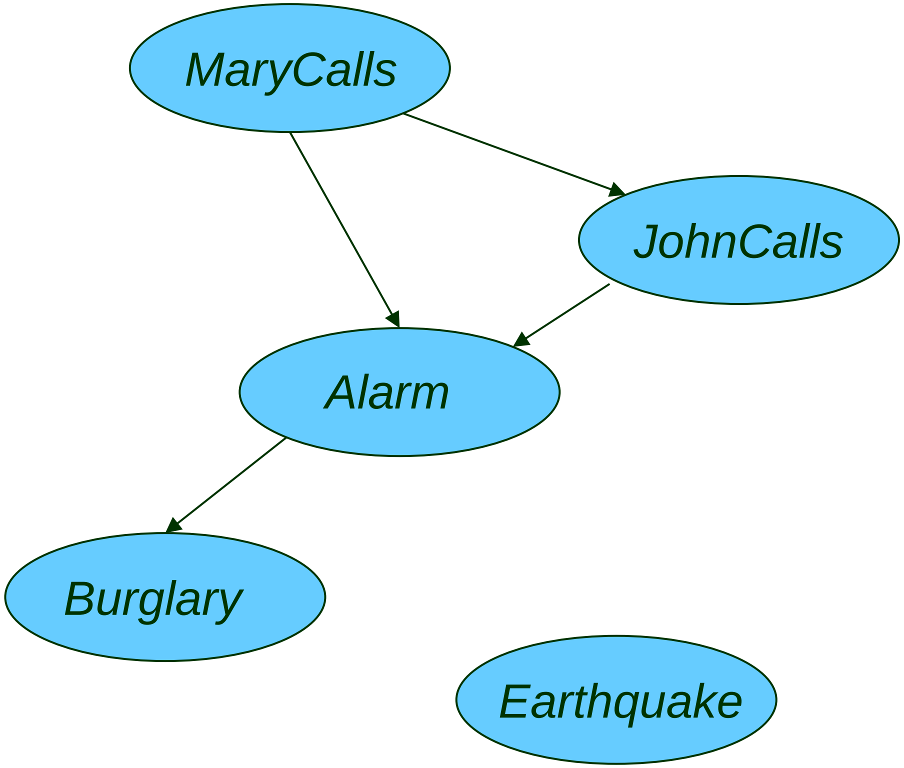
$$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

$P(e \mid a, b) \qquad P(e \mid a), P(e \mid b)$

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$$

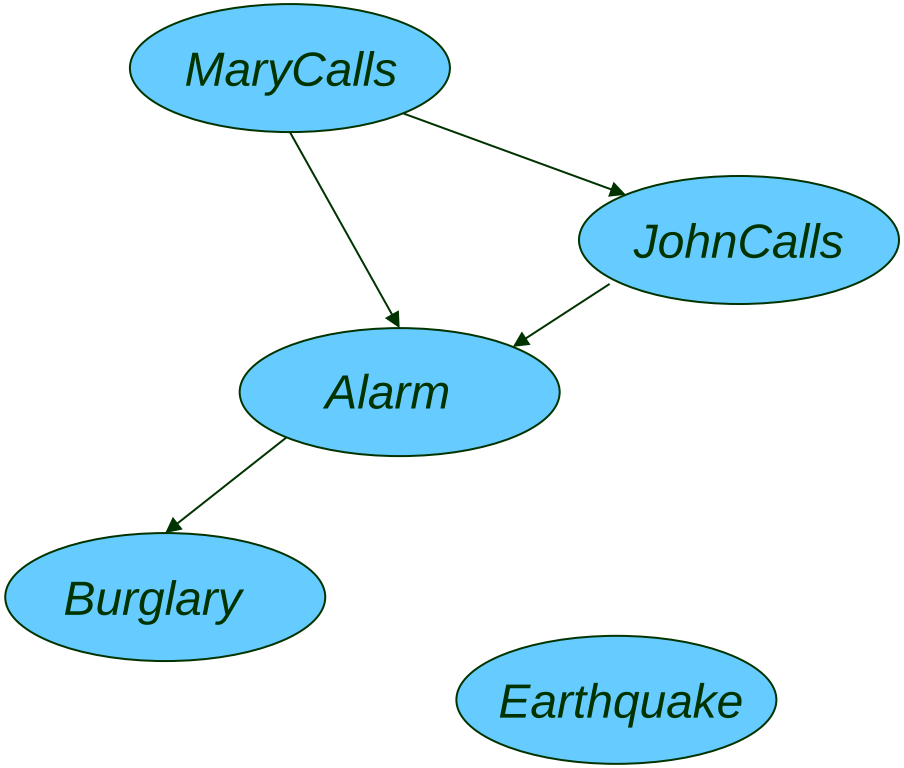
// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

$$P(e \mid a, b) \qquad P(e \mid a), P(e \mid b)$$

// If the alarm is on, it is more likely that
// there has been earthquake. If there
// has been a burglary, it is slightly more
// likely that it happened after an
// earthquake. In the occurrences of
// both events, the chance of earthquake
// occurrence is even higher.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



$$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$$

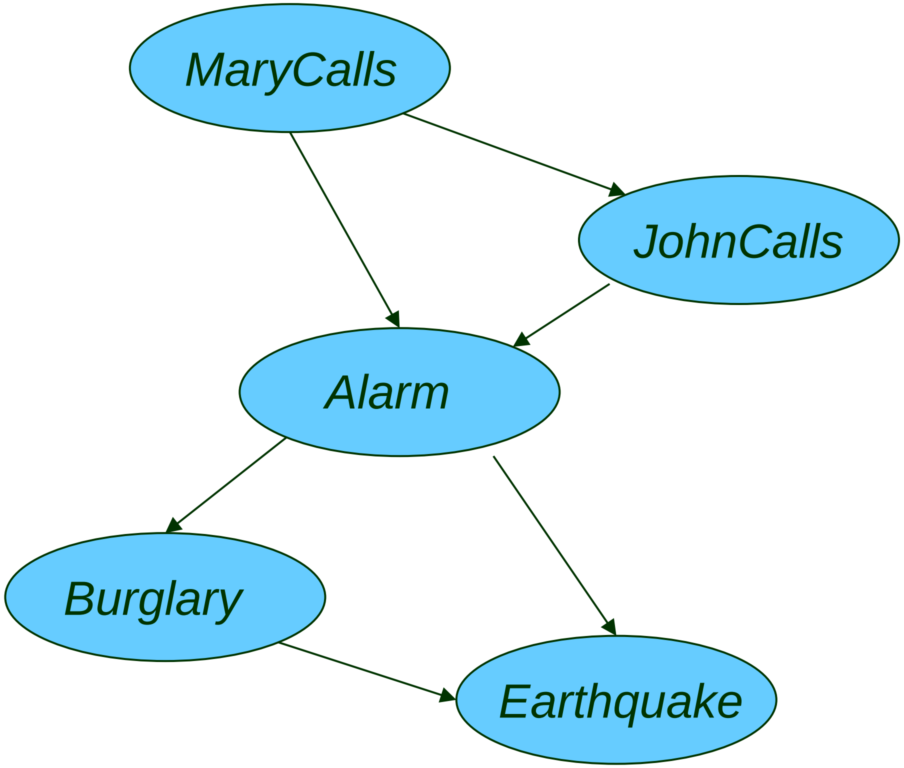
// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

$$P(e \mid a, b) > P(e \mid a), P(e \mid b)$$

// If the alarm is on, it is more likely that
// there has been earthquake. If there
// has been a burglary, it is slightly more
// likely that it happened after an
// earthquake. In the occurrences of
// both events, the chance of earthquake
// occurrence is even higher.

Construction for the Burglary Example

MaryCalls, *JohnCalls*, *Alarm*, *Burglary*, *Earthquake*.



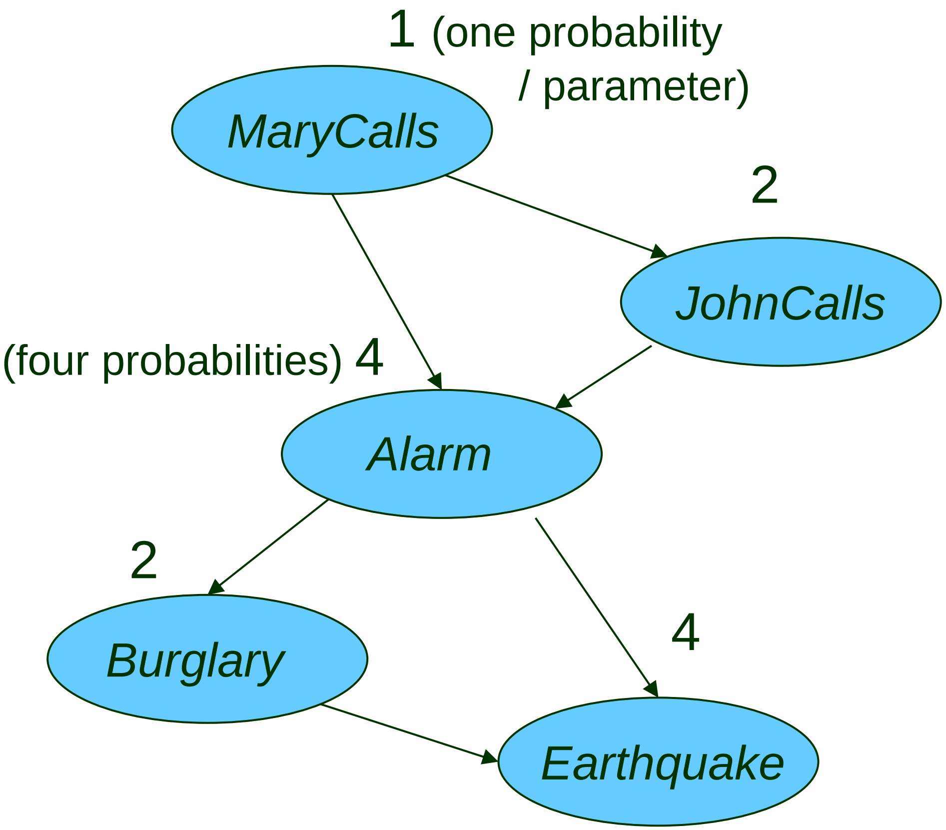
$$\boldsymbol{P}(B \mid A, J, M) = \boldsymbol{P}(B \mid A)$$

// If the value of A (either a or $\neg a$) is
// known, then the call from John or
// Mary does not add any information
// about burglary.

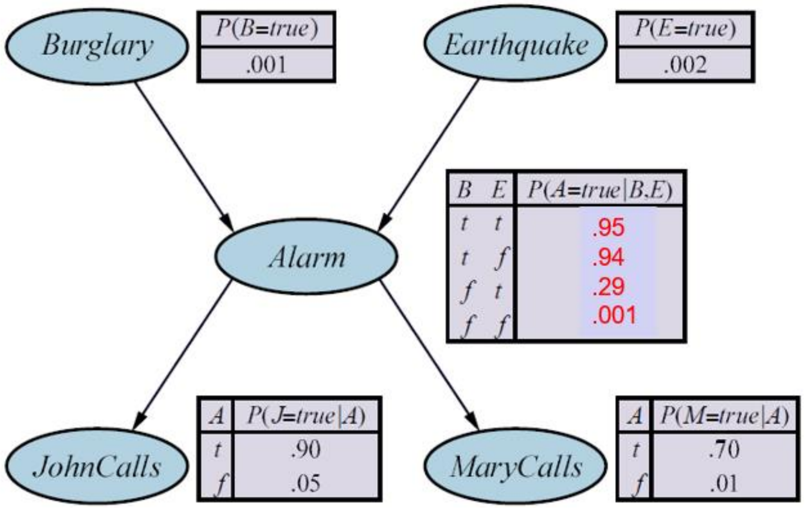
$$P(e \mid a, b) > P(e \mid a), P(e \mid b)$$

// If the alarm is on, it is more likely that
// there has been earthquake. If there
// has been a burglary, it is slightly more
// likely that it happened after an
// earthquake. In the occurrences of
// both events, the chance of earthquake
// occurrence is even higher.

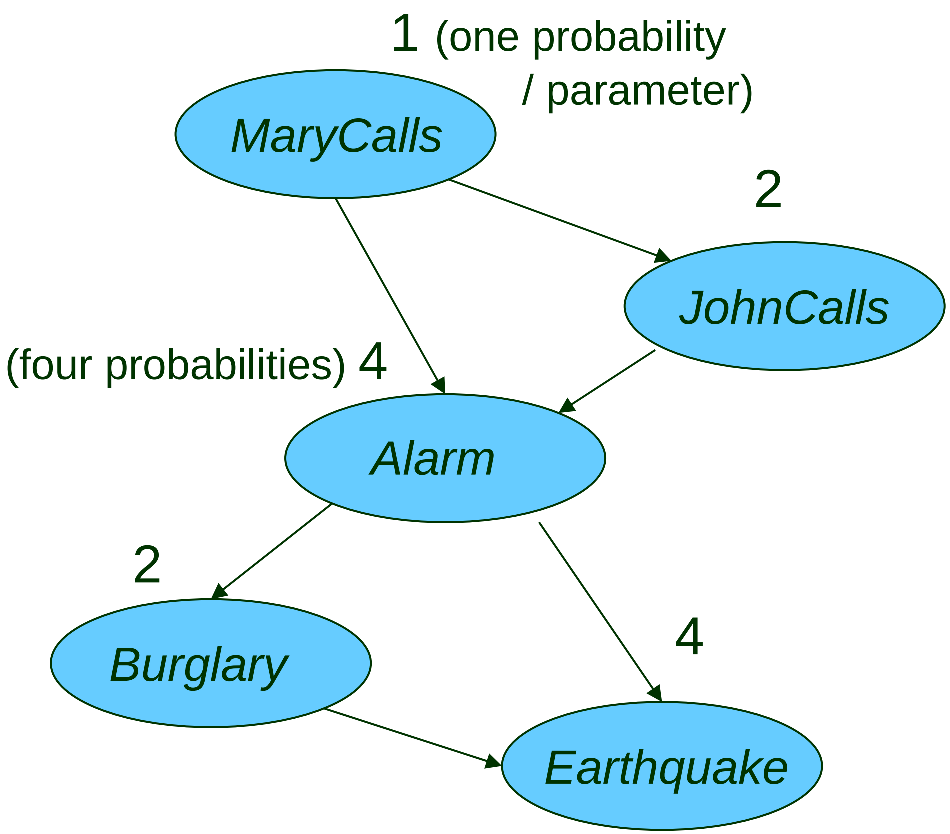
Node Ordering Matters



$1 + 2 + 4 + 2 + 4 = 13$
conditional probabilities



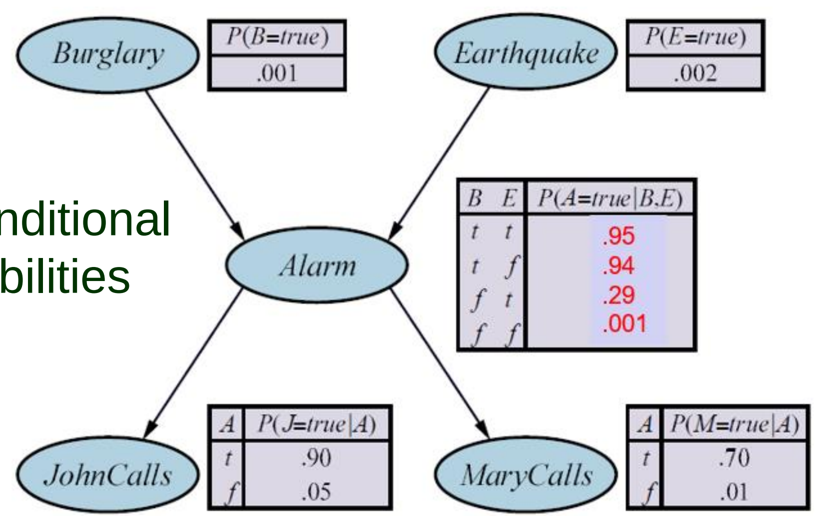
Node Ordering Matters



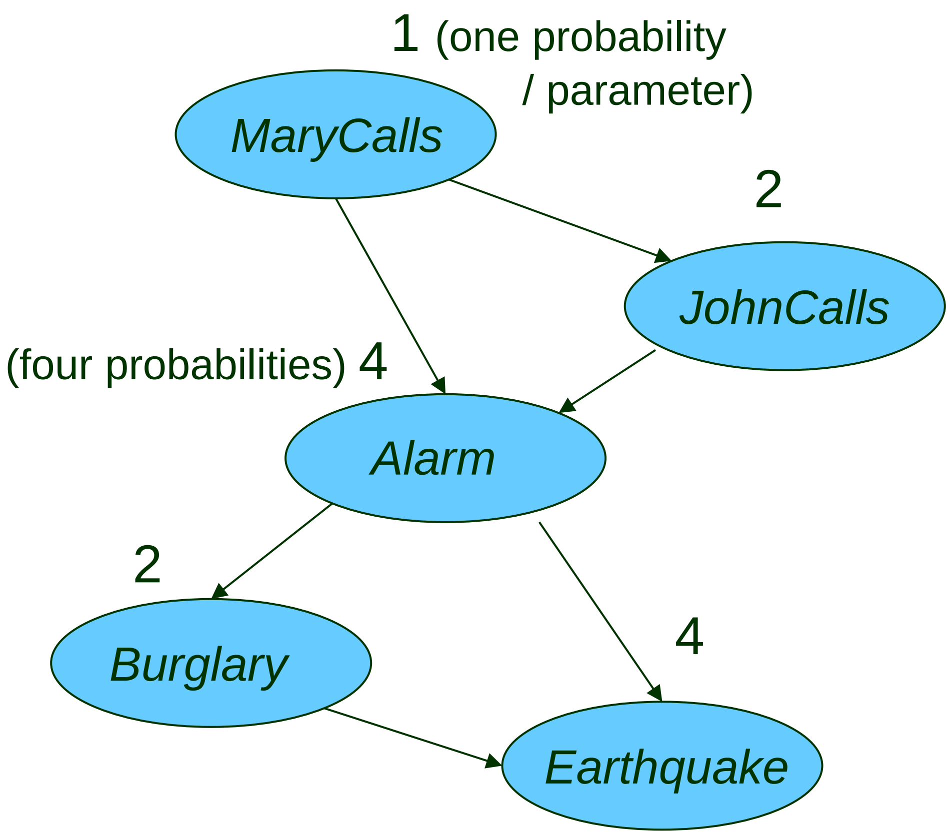
$1 + 2 + 4 + 2 + 4 = 13$
conditional probabilities

- More conditional probabilities than needed.

10 conditional probabilities



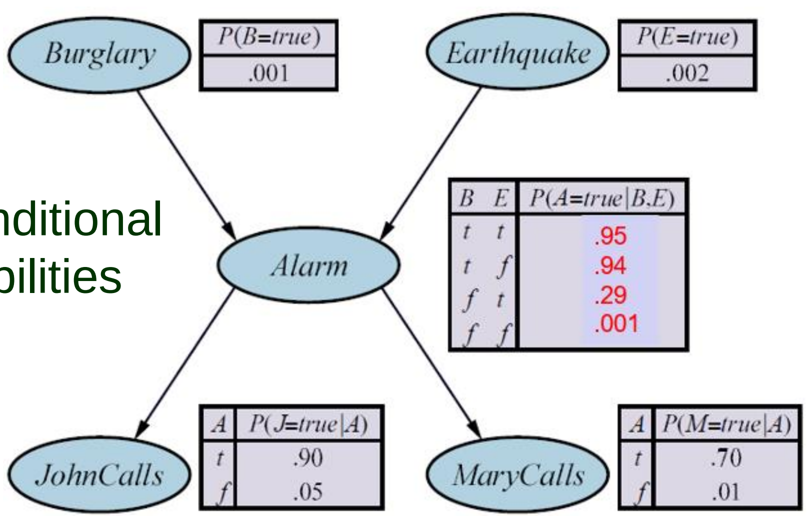
Node Ordering Matters



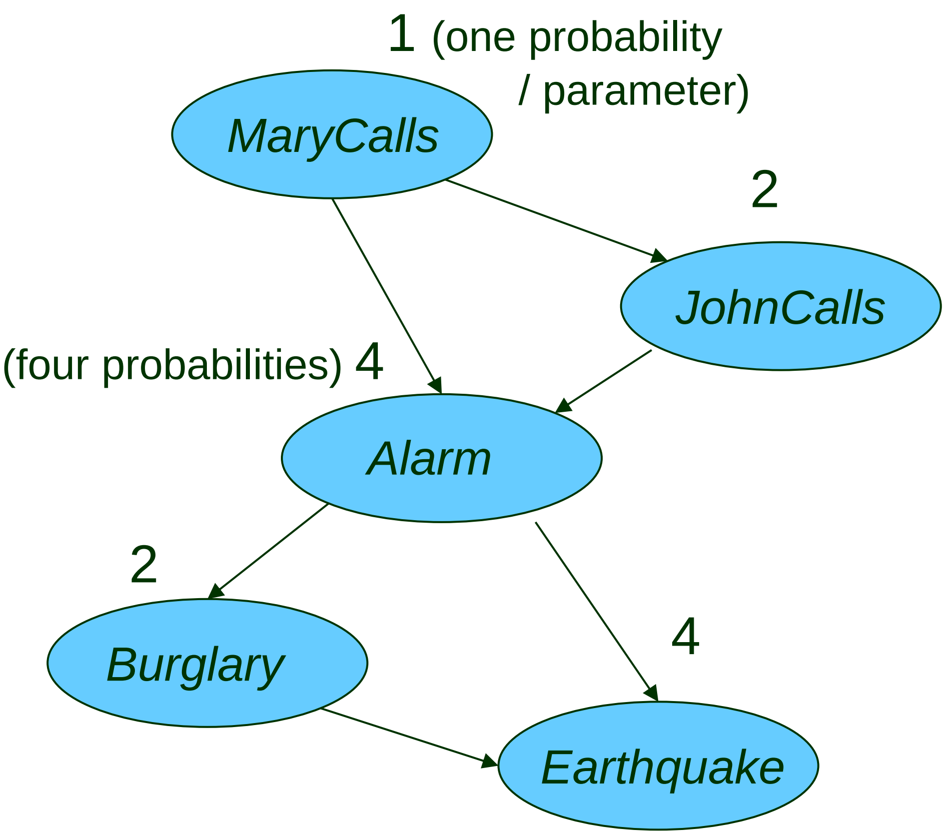
$1 + 2 + 4 + 2 + 4 = 13$ conditional probabilities

- More conditional probabilities than needed.
- Assessment of unnatural probabilities, e.g., $\boldsymbol{P}$(*Earthquake* | *Burglary, Alarm*}.

10 conditional probabilities



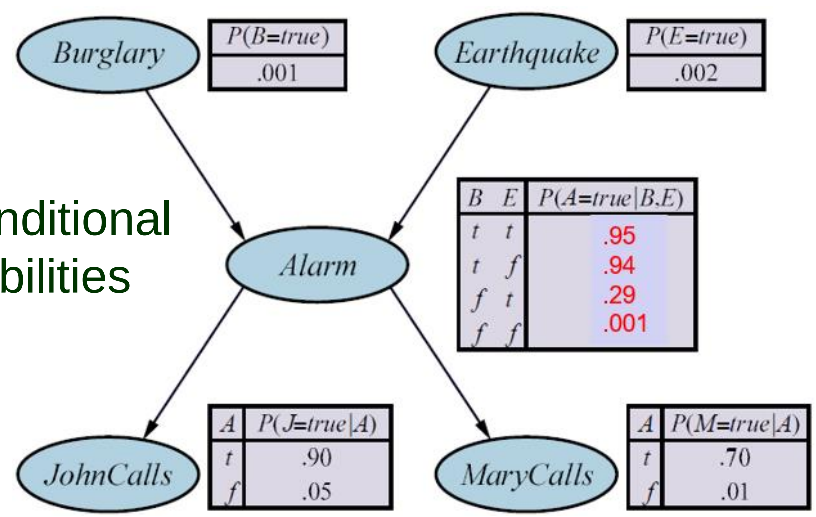
Node Ordering Matters



$1 + 2 + 4 + 2 + 4 = 13$
conditional probabilities

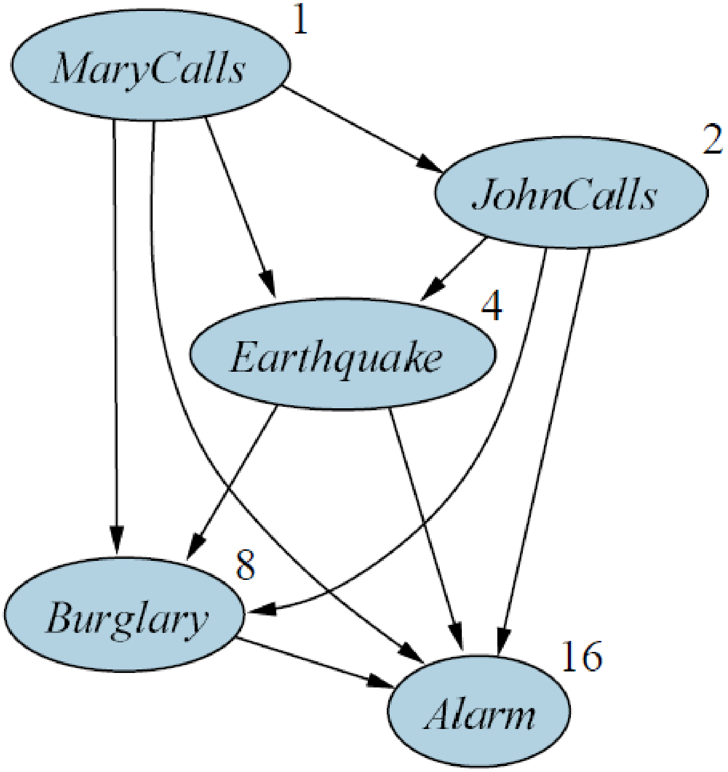
- ♣ More conditional probabilities than needed.
- ♣ Assessment of unnatural probabilities, e.g., $\boldsymbol{P}$(*Earthquake* | *Burglary, Alarm*}.
- ♦ Sticking to a causal model results in fewer probabilities that are also easier to come up with.

10 conditional probabilities



Bad Node Ordering

MaryCalls, *JohnCalls*, *Earthquake, Burglary*, *Alarm*.



$1 + 2 + 4 + 8 + 16 = 31$ distinct probabilities (exactly the same as the full joint distribution)!

Roles of Casualty

- Deciding conditional independence is hard in noncausal directions.
  (Causal models and conditional independence seem hardwired for humans!)

- Assessing conditional probabilities is hard in noncausal directions.

- The interpretation of directed acyclic graphs as carriers of independence assumptions does not necessarily imply causation.

- The ubiquity of DAG models in statistical and AI applications stems (often unwittingly) primarily from their causal interpretation.

- In practice, DAG models are rarely used in any variable ordering other than those which respect the direction of time and causation.

Compactness of Bayes Nets

- The full joint distribution contains 2^n numbers.

Compactness of Bayes Nets

- ♠ The full joint distribution contains 2^n numbers.

- It is reasonable to assume that each random variable is directly influenced by $\leq k$ others (i.e., every node has $\leq k$ parents in a BN).

Compactness of Bayes Nets

♠ The full joint distribution contains 2^n numbers.

- It is reasonable to assume that each random variable is directly influenced by $\leq k$ others (i.e., every node has $\leq k$ parents in a BN).

- The conditional probability table (CPT) for each node has size $\leq 2^k$.

Compactness of Bayes Nets

- ♠ The full joint distribution contains 2^n numbers.

- It is reasonable to assume that each random variable is directly influenced by $\leq k$ others (i.e., every node has $\leq k$ parents in a BN).

- The conditional probability table (CPT) for each node has size $\leq 2^k$.

- With n Boolean variables, the network has $\leq n \cdot 2^k$ numbers.

Compactness of Bayes Nets

- ♠ The full joint distribution contains 2^n numbers.

- It is reasonable to assume that each random variable is directly influenced by $\leq k$ others (i.e., every node has $\leq k$ parents in a BN).

- The conditional probability table (CPT) for each node has size $\leq 2^k$.

- With n Boolean variables, the network has $\leq n \cdot 2^k$ numbers.

- ♦ To avoid a fully connected network, leave out links that represent slight dependencies.