

# Cooperating Intelligent Systems

Bayesian networks

Chapter 14, AIMA

# Inference

- Inference in the statistical setting means computing probabilities of different outcomes given the observed information

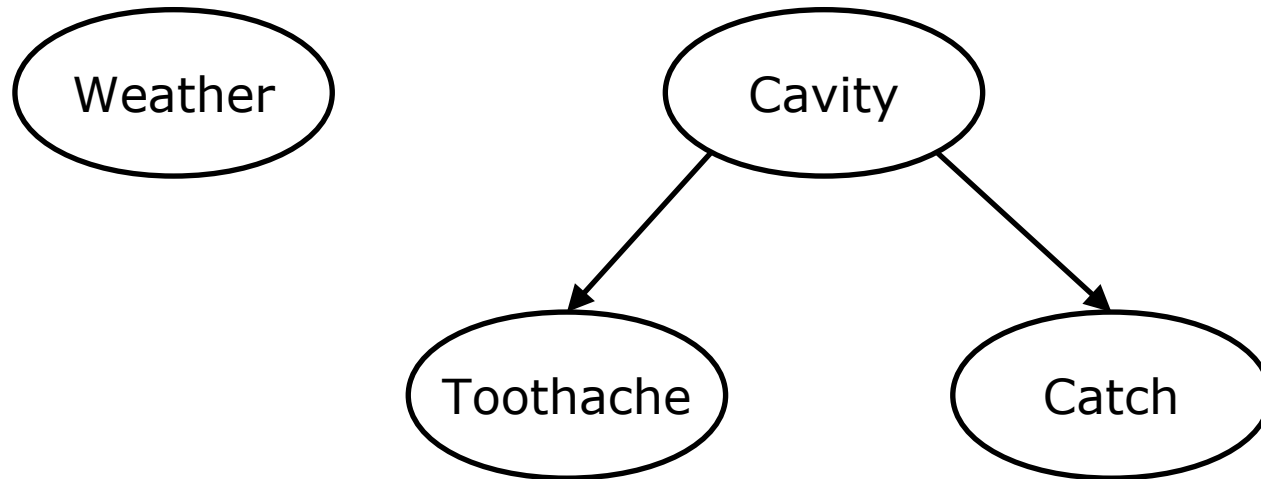
$$\mathbf{P}(\textit{Outcome} \mid \textit{Information})$$

- We need an efficient method for doing this which is more widely applicable than the naïve Bayes model

# Bayesian networks

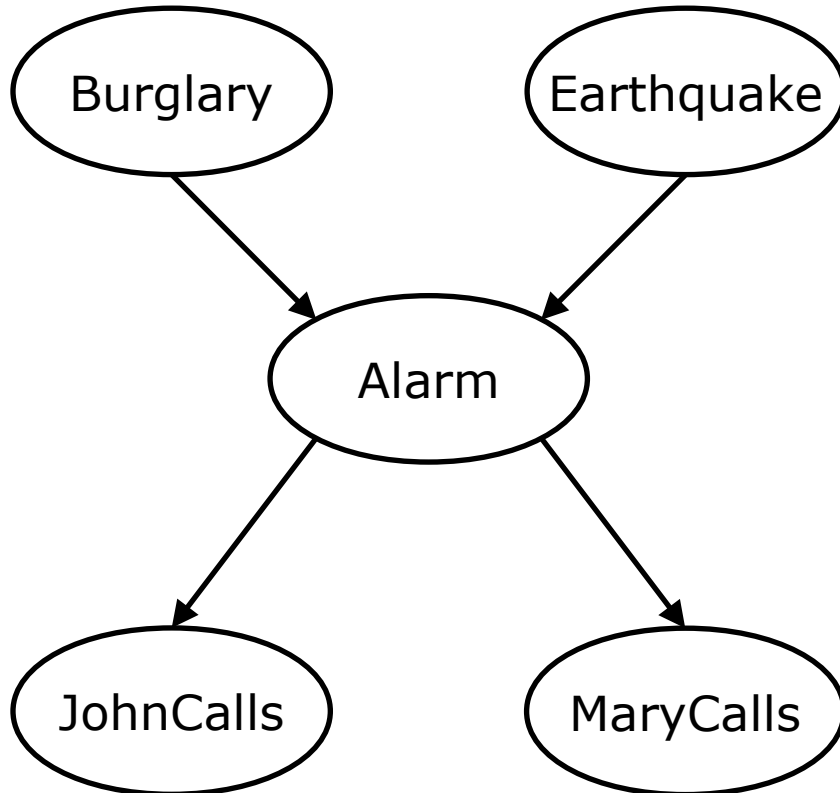
- A **Bayesian network** is a directed graph in which each node is annotated with quantitative probability information:
1. The set of nodes of the network corresponds to a set of random variables  $\{X_1, X_2, X_3, \dots\}$
  2. pairs of nodes can be connected by directed links defining a parent  $\rightarrow$  child relation
  3. Each node  $X_i$  contains a conditional probability distribution  $\mathbf{P}(X_i \mid \textit{Parents}(X_i))$
  4. The graph is a directed acyclic graph (DAG)

# The dentist network



	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

# The alarm network



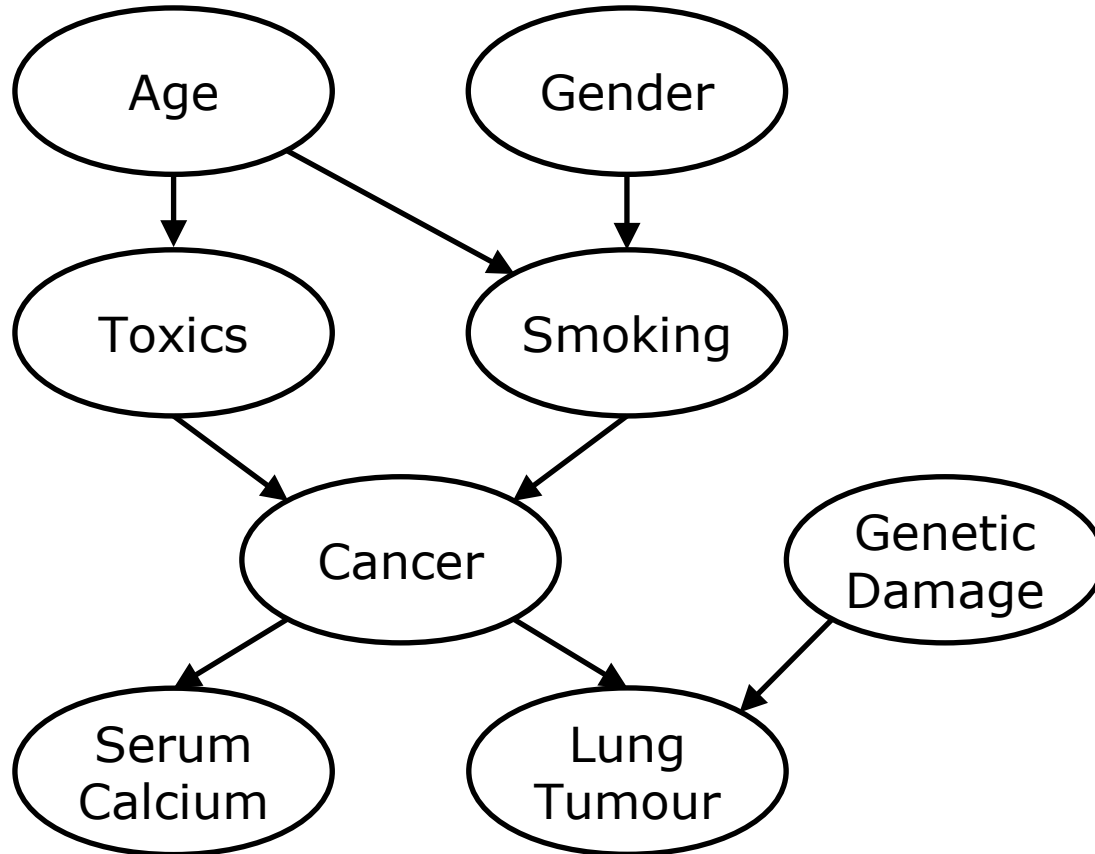
Burglar alarm responds to both earthquakes and burglars.

Two neighbors: John and Mary, who have promised to call you when the alarm goes off.

John always calls when there's an alarm, and sometimes when there's not an alarm.

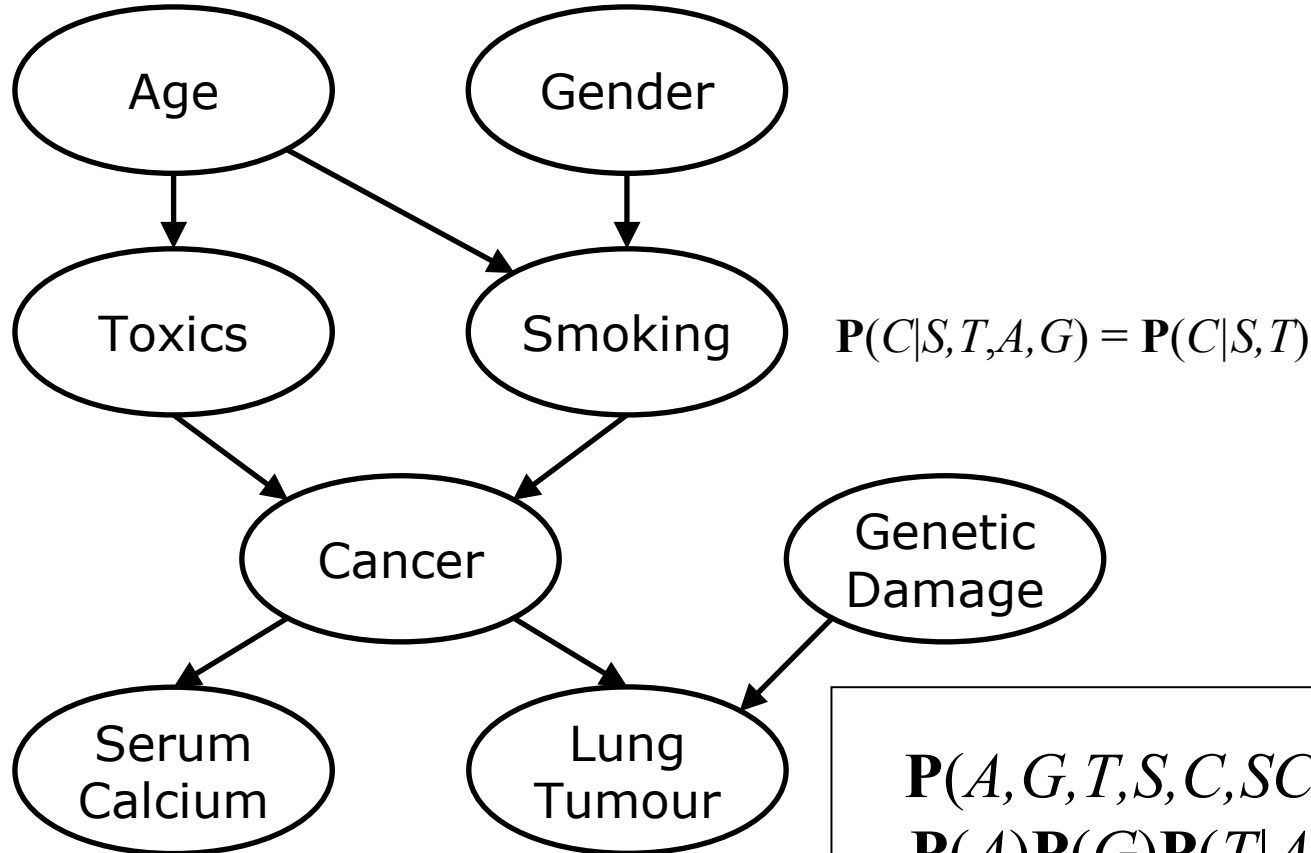
Mary sometimes misses the alarms (she likes loud music).

# The cancer network



# The cancer network

$$P(A, G) = P(A)P(G)$$



$$P(C|S, T, A, G) = P(C|S, T)$$

$$P(SC, C, LT, GD) = P(SC|C)P(LT|C, GD)P(C) P(GD)$$

$$P(A, G, T, S, C, SC, LT, GD) = P(A)P(G)P(T|A)P(S|A, G) \times P(C|T, S)P(GD)P(SC|C) \times P(LT|C, GD)$$

# Meaning of Bayesian network

The general chain rule (always true):

$$\begin{aligned} P(x_1, x_2, \dots, x_n) &= P(x_1 | x_2, x_3, \dots, x_n) P(x_2, x_3, \dots, x_n) = \\ &P(x_1 | x_2, x_3, \dots, x_n) P(x_2 | x_3, x_4, \dots, x_n) P(x_3, x_4, \dots, x_n) = \dots \\ &= \prod_{i=1}^n P(x_i | x_{i+1}, \dots, x_n) \end{aligned}$$

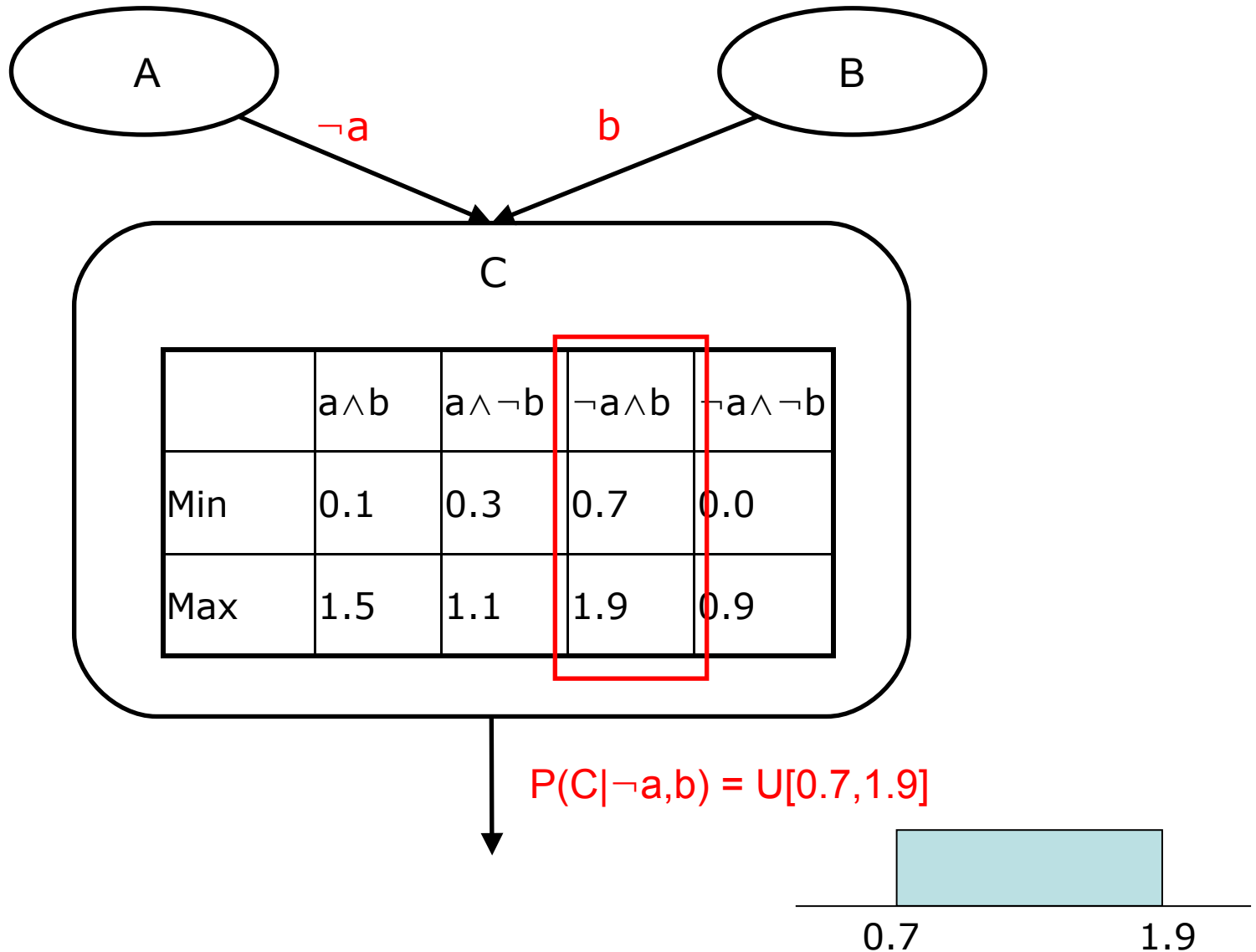
The Bayesian network chain rule:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$

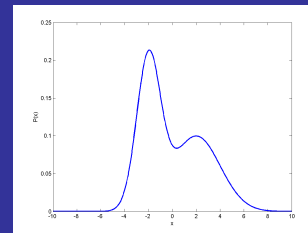
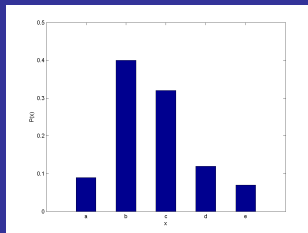
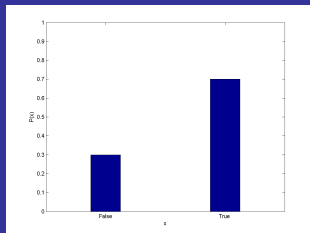
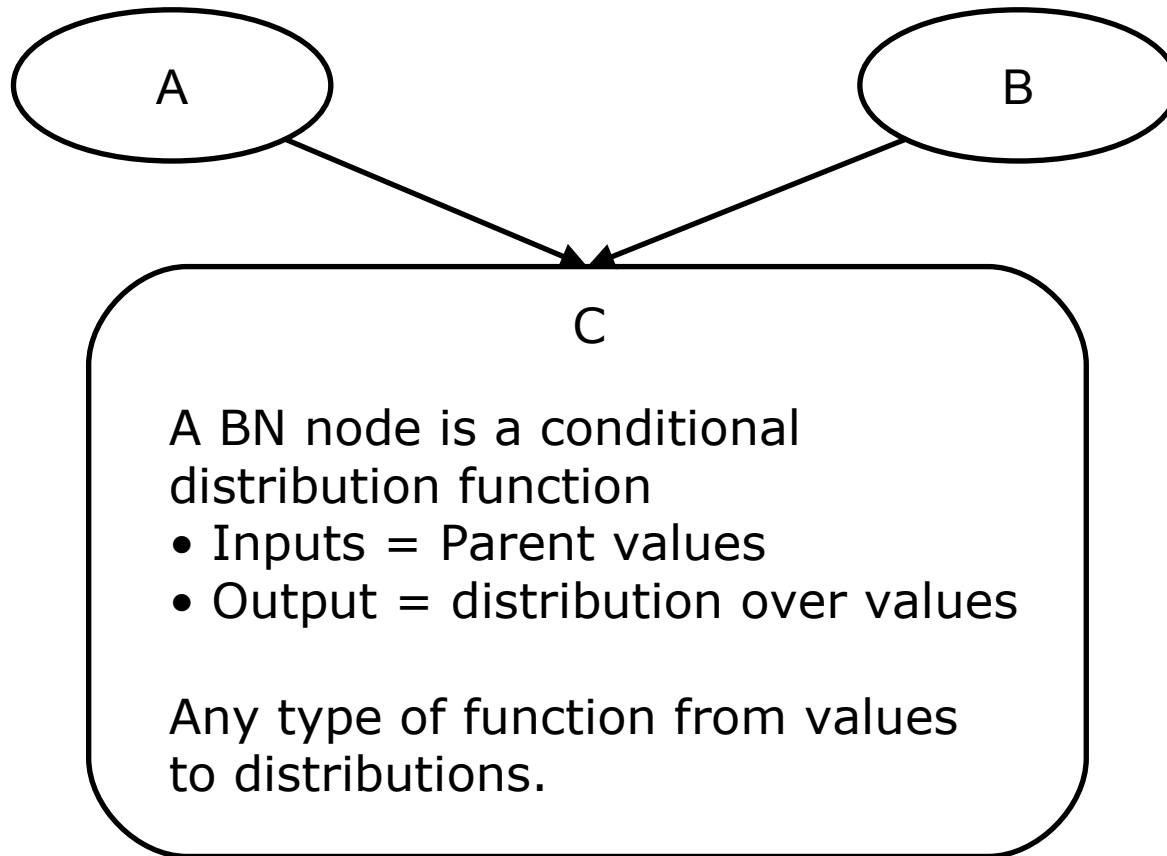
The BN is a correct representation of the domain iff each node is conditionally independent of its predecessors, given its parents.



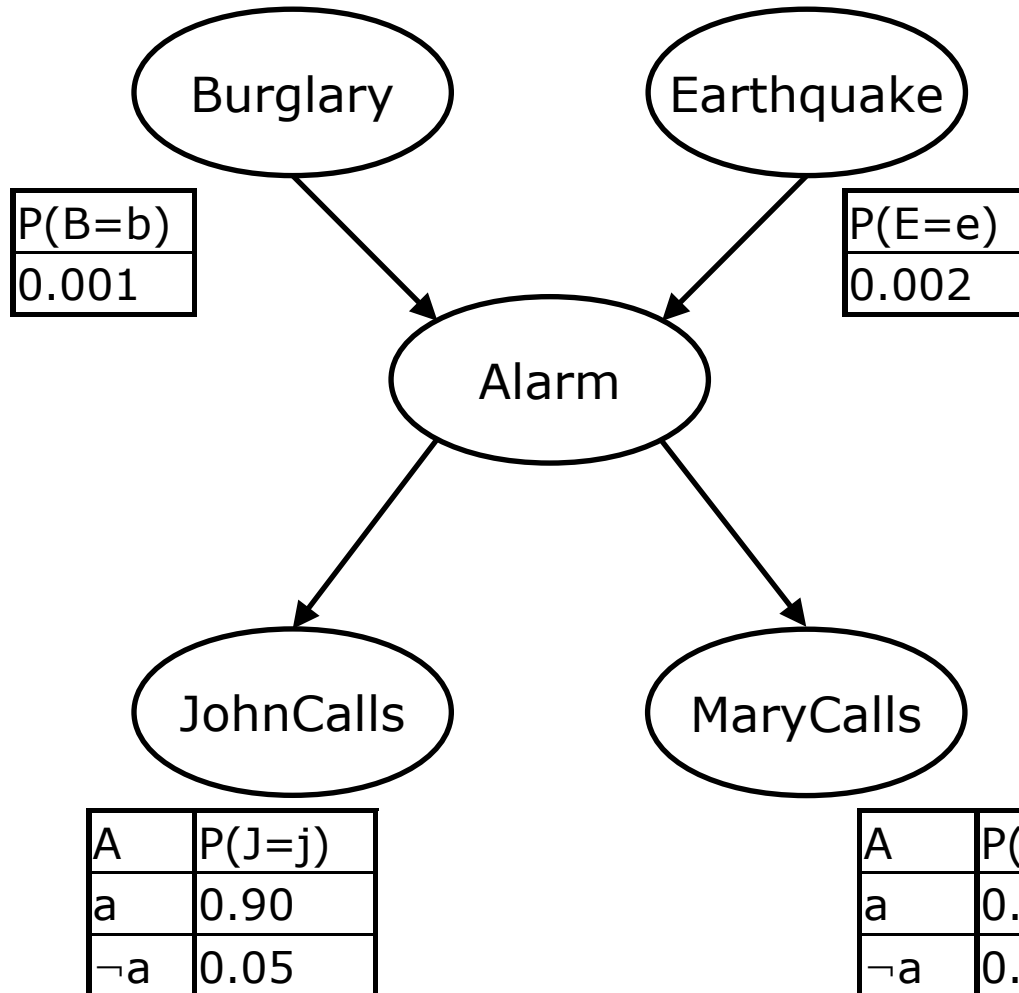
# Bayes network node is a function



# Bayes network node is a function



# Example: The alarm network

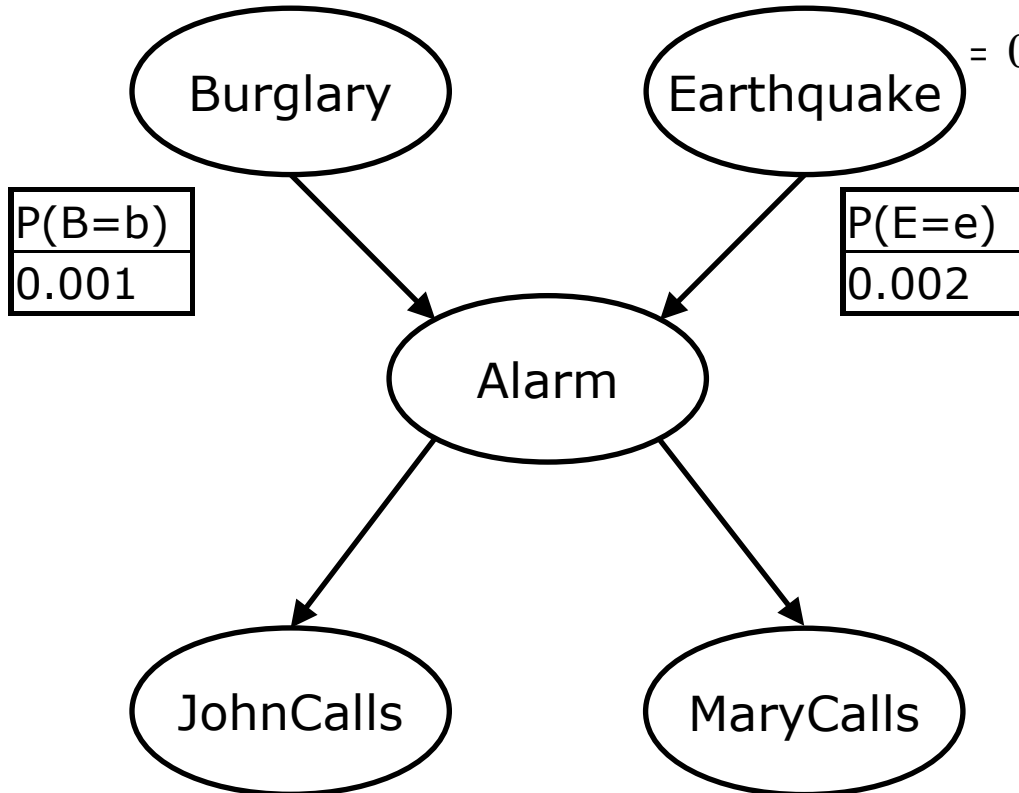


Note: Each number in the tables represents a boolean distribution. Hence there is a distribution output for every input.

B	E	$P(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001

# Example: The alarm network

$$\begin{aligned}
 &P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\
 &= P(\neg b)P(\neg e)P(a | \neg b, \neg e)P(m | a)P(j | a) \\
 &= 0.999 \cdot 0.998 \cdot 0.001 \cdot 0.70 \cdot 0.90 = 0.00063
 \end{aligned}$$



$P(B=b)$
0.001

$P(E=e)$
0.002

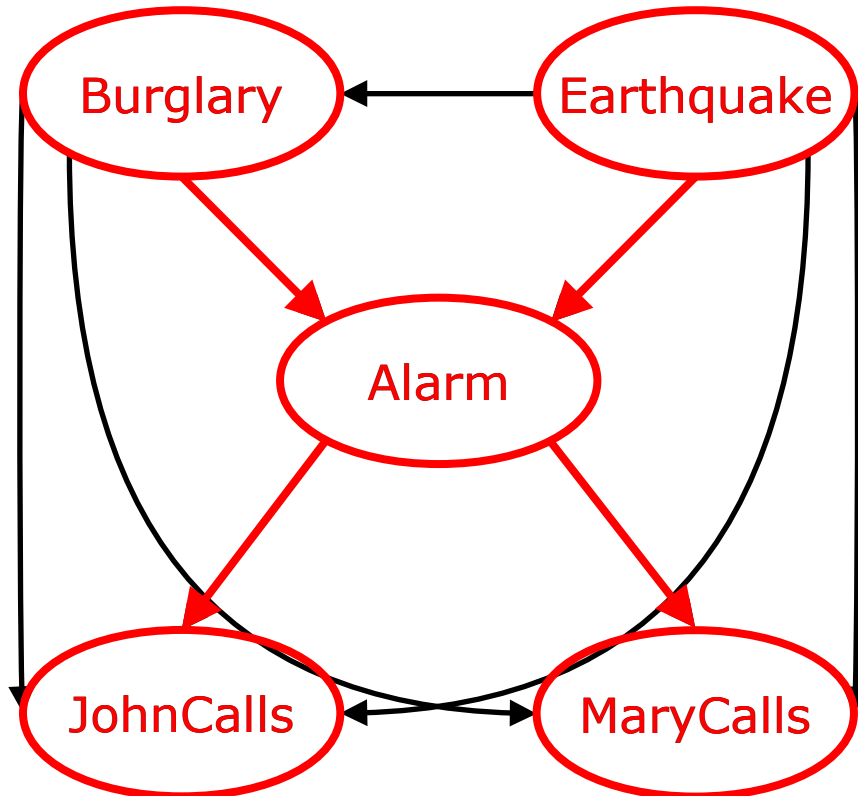
Probability distribution for "no earthquake, no burglary, but alarm, and both Mary and John make the call"

B	E	$P(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001

A	$P(J=j)$
a	0.90
$\neg a$	0.05

A	$P(M=m)$
a	0.70
$\neg a$	0.01

# The alarm network



The fully correct alarm network might look something like the figure.

The **Bayesian network** assumes that some of the variables are independent

- or that the dependencies can be neglected since they are very weak.

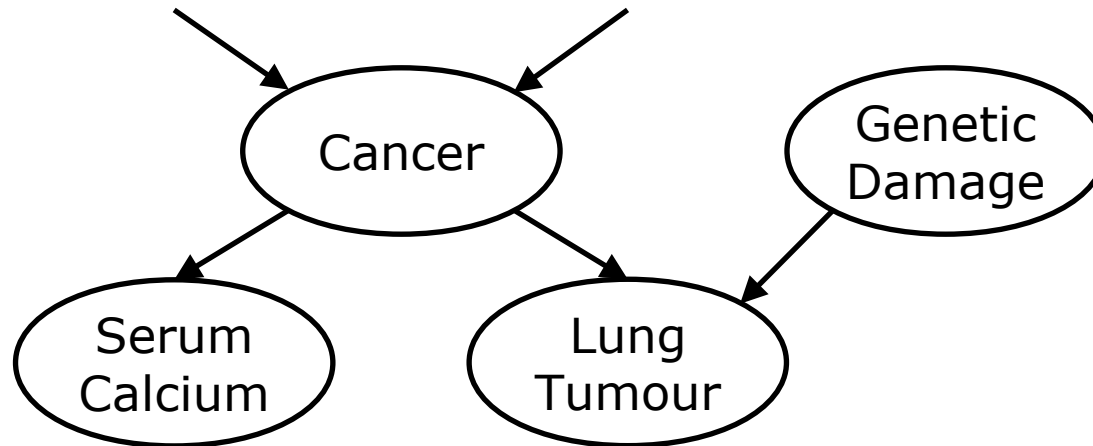
The correctness of the Bayesian network, of course, depends on the validity of these assumptions!

It is this sparse connection structure that makes the BN approach feasible:  
~linear growth in complexity rather than exponential, in practice

# How to construct a BN?

- Add nodes in causal order
  - “causal” determined from expertise
- Determine conditional independence using either (or all) of the following semantics:
  - Blocking/d-separation rule
  - Non-descendant rule
  - Markov blanket rule
  - Experience/your beliefs

# Path blocking & d-separation



Intuitively, knowledge about *Serum Calcium* influences our belief about *Cancer* – if we don't know the value of *Cancer* – which, in turn, influences our belief about *Lung Tumour*, etc.

However, if we are given the value of *Cancer* (i.e.  $C = \text{true}$  or  $\text{false}$ ), then knowledge of *Serum Calcium* will not tell us anything about *Lung Tumour* that we don't already know

- remember conditional independence from last lecture.

We say that *Cancer* **d-separates** (direction-dependent separates) *Serum Calcium* and *Lung Tumour*.

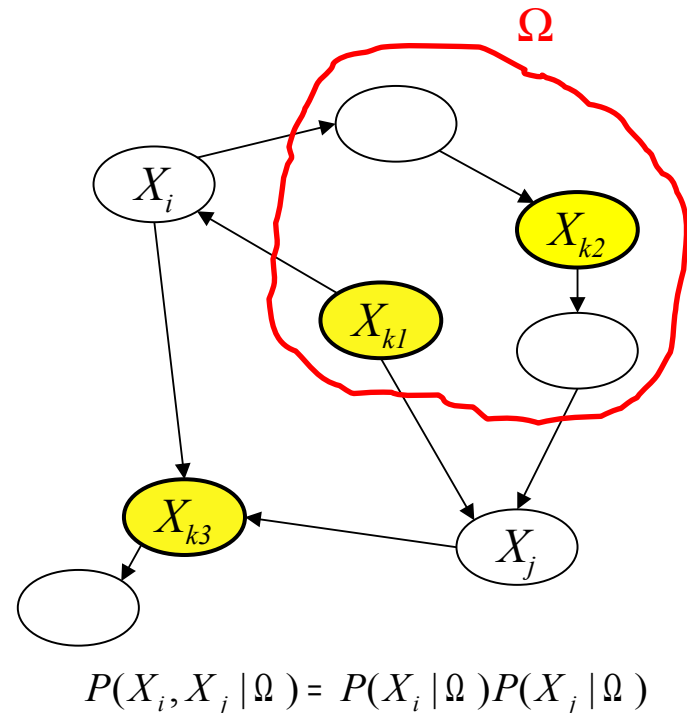
# Path blocking & d-separation

$X_i$  and  $X_j$  are d-separated if all paths between them are blocked

Two nodes  $X_i$  and  $X_j$  are conditionally independent given a set  $\Omega = \{X_1, X_2, X_3, \dots\}$  of nodes if for every undirected path in the BN between  $X_i$  and  $X_j$  there is some node  $X_k$  on the path having one of the following three properties:

1.  $X_k \in \Omega$ , and both arcs on the path lead out of  $X_k$ .
2.  $X_k \in \Omega$ , and one arc on the path leads into  $X_k$  and one arc leads out.
3. Neither  $X_k$  nor any descendant of  $X_k$  is in  $\Omega$ , and both arcs on the path lead into  $X_k$ .

$X_k$  **blocks** the path between  $X_i$  and  $X_j$





# Some definitions of BN

(from Wikipedia)

- 1.  $\mathbf{X}$**  is a Bayesian network with respect to  **$\mathbf{G}$**  if its joint probability density function can be written as a product of the individual density functions, conditional on their parent variables:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents}(X_i))$$

**$\mathbf{X}$**  =  $\{X_1, X_2, \dots, X_N\}$  is a set of random variables

**$\mathbf{G}$**  =  $(V, E)$  is a directed acyclic graph (DAG) of vertices  $(V)$  and edges  $(E)$

# Some definitions of BN

(from Wikipedia)

- 2.  $\mathbf{X}$**  is a Bayesian network with respect to  **$\mathbf{G}$**  if it satisfies the local Markov property: each variable is conditionally independent of its non-descendants given its parent variables:

$$P(x_i \mid \text{non - descendants}(X_i)) = P(x_i \mid \text{parents}(X_i))$$

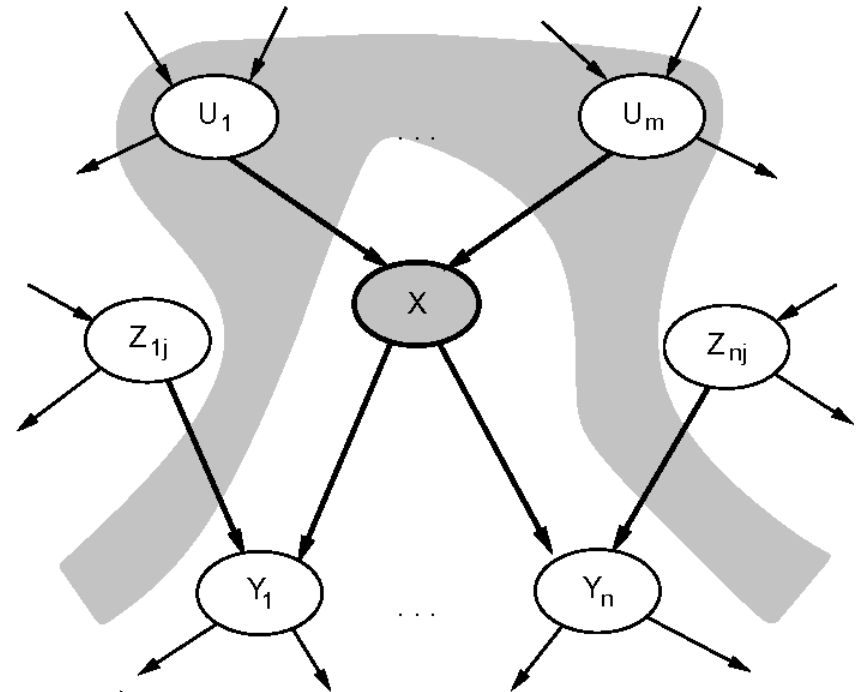
*Note:*  $\text{parents}(X_i) \subseteq \text{non - descendants}(X_i)$

**$\mathbf{X}$**  =  $\{X_1, X_2, \dots, X_N\}$  is a set of random variables

**$\mathbf{G}$**  =  $(V, E)$  is a directed acyclic graph (DAG) of vertices ( $V$ ) and edges ( $E$ )

# Non-descendants

A node is conditionally independent of its non-descendants ( $Z_{ij}$ ), given its parents.



$$P(X, Z_{1j}, \dots, Z_{nj} | U_1, \dots, U_m) =$$

$$P(X | U_1, \dots, U_m) P(Z_{1j}, \dots, Z_{nj} | U_1, \dots, U_m)$$

$$P(X | Z_{1j}, \dots, Z_{nj}, U_1, \dots, U_m) =$$

$$P(X | U_1, \dots, U_m)$$

# Some definitions of BN

(from Wikipedia)

- 3.  $\mathbf{X}$**  is a Bayesian network with respect to  **$\mathbf{G}$**  if every node is conditionally independent of all other nodes in the network, given its Markov blanket.

The *Markov blanket* of a node is its parents, children and children's parents.

$$P(x_i \mid \text{all nodes}) = P(x_i \mid \text{Markov blanket}(X_i))$$

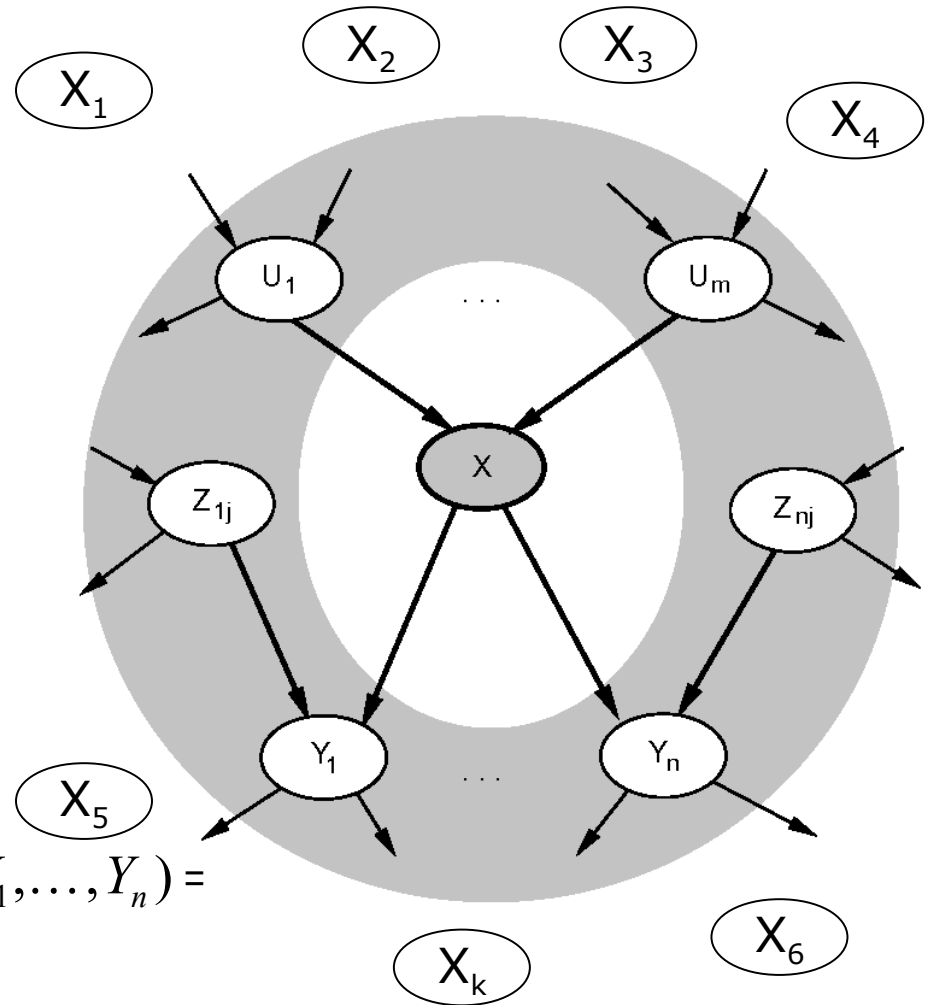
**$\mathbf{X}$**  =  $\{X_1, X_2, \dots, X_N\}$  is a set of random variables

**$\mathbf{G}$**  = (V,E) is a directed acyclic graph (DAG) of vertices (V) and edges (E)

# Markov blanket

A node is conditionally independent of all other nodes in the network, given its parents, children, and children's parents

These constitute the node's *Markov blanket*.



$$P(X | X_1, \dots, X_k, U_1, \dots, U_m, Z_{1j}, \dots, Z_{nj}, Y_1, \dots, Y_n) =$$

$$P(X | U_1, \dots, U_m, Z_{1j}, \dots, Z_{nj}, Y_1, \dots, Y_n)$$

$$P(X, X_1, \dots, X_k | U_1, \dots, U_m, Z_{1j}, \dots, Z_{nj}, Y_1, \dots, Y_n) =$$

$$P(X | U_1, \dots, U_m, Z_{1j}, \dots, Z_{nj}, Y_1, \dots, Y_n) P(X_1, \dots, X_k | U_1, \dots, U_m, Z_{1j}, \dots, Z_{nj}, Y_1, \dots, Y_n)$$

# Some definitions of BN

(from Wikipedia)

**4.  $\mathbf{X}$**  is a Bayesian network with respect to  **$\mathbf{G}$**  if, for any two nodes  $i, j$ :

$$P(x_i, x_j \mid X_1, X_2, \dots, X_N) =$$

$$P(x_i \mid d\text{-separating set}(i, j))P(x_j \mid d\text{-separating set}(i, j))$$

The  $d$ -separating set( $i, j$ ) is the set of nodes that  $d$ -separate node  $i$  and  $j$ .

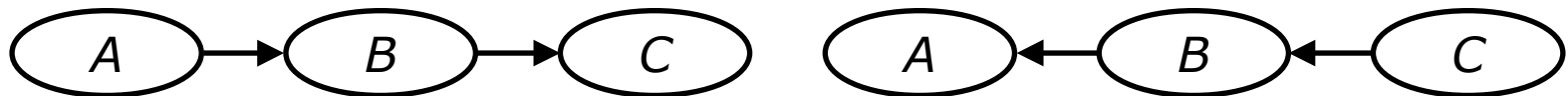
*The Markov blanket of node  $i$  is the minimal set of nodes that  $d$ -separates node  $i$  from all other nodes.*

**$\mathbf{X}$**  =  $\{X_1, X_2, \dots, X_N\}$  is a set of random variables

**$\mathbf{G}$**  =  $(V, E)$  is a directed acyclic graph (DAG) of vertices ( $V$ ) and edges ( $E$ )

# Causal networks

- Bayesian networks are usually used to represent causal relationships. This is, however, not strictly necessary: a directed edge from node  $i$  to node  $j$  does not require that  $X_i$  is causally dependent on  $X_j$ .
  - This is demonstrated by the fact that Bayesian networks on the two graphs:



are equivalent. They impose the same conditional independence requirements.

*A causal network is a Bayesian network with an explicit requirement that the relationships be causal.*

# Causal networks

$$P(A, B, C) = P(A | B)P(B | C)P(C)$$
$$= P(A | B) * \dots ?$$



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

*The equivalence is proved with Bayes theorem...*



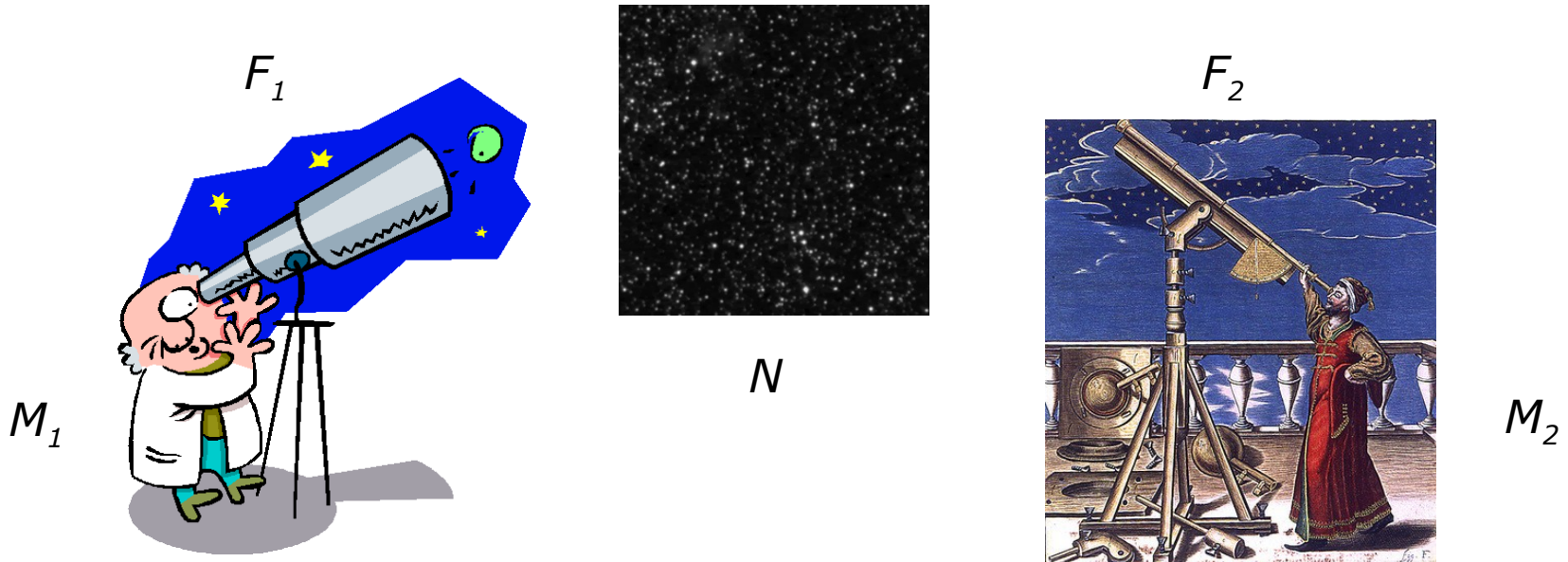
# Exercise 14.12\* (a) in AIMA

- Two astronomers in different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small region of the sky, using their telescopes. Normally there is a small possibility  $e$  of error up to one star in each direction. Each telescope can also (with a much smaller probability  $f$ ) be badly out of focus (events  $F_1$  and  $F_2$ ) in which case the scientist will undercount by three or more stars (or, if  $N$  is less than 3, fail to detect any stars at all). Consider the three networks in Figure 14.22\*.
  - (a) Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?

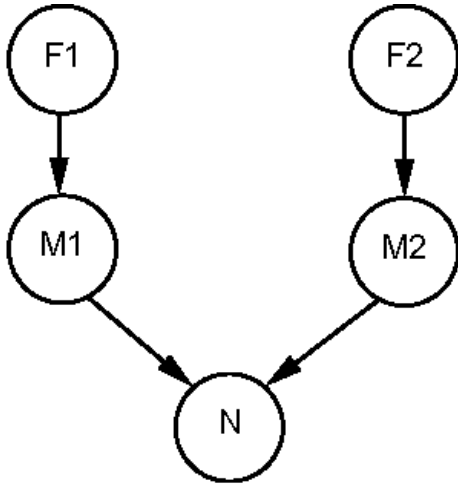
\*In the 2:nd edition is this exercise 14.3 and the figure is 14.19.

# Exercise 14.12 (a) in AIMA

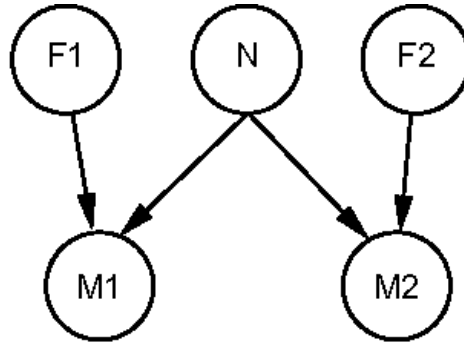
- Two astronomers in different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small region of the sky, using their telescopes. Normally there is a small possibility  $e$  of error up to one star in each direction. Each telescope can also (with a much smaller probability  $f$ ) be badly out of focus (events  $F_1$  and  $F_2$ ) in which case the scientist will undercount by three or more stars (or, if  $N$  is less than 3, fail to detect any stars at all). Consider the three networks in Figure 14.22.
  - (a) Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?



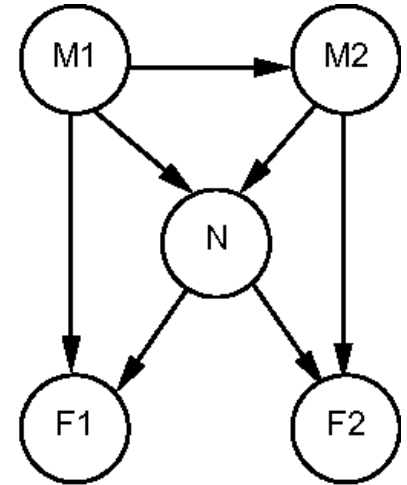
# Exercise 14.12 (a) in AIMA



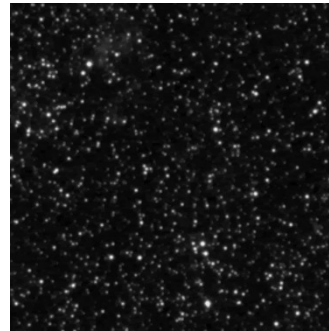
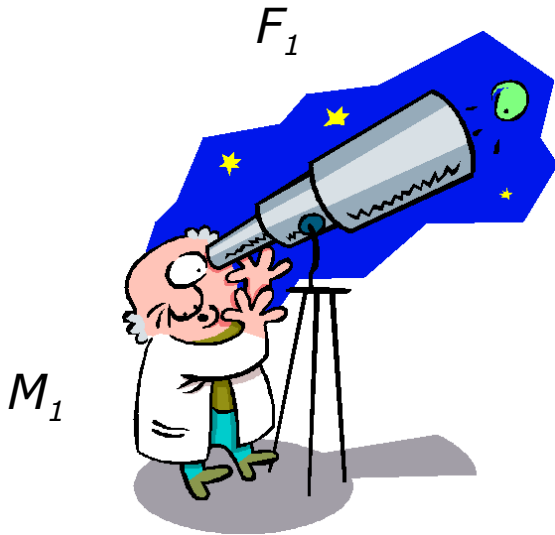
(i)



(ii)

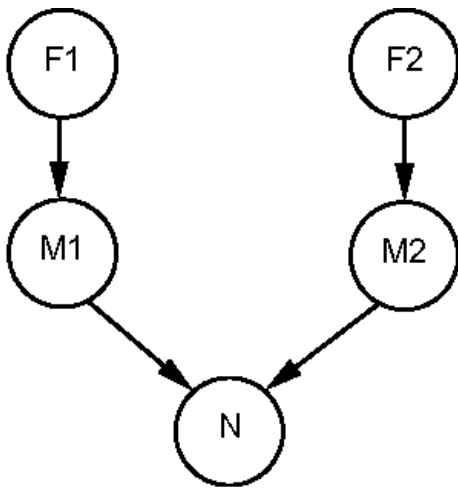


(iii)

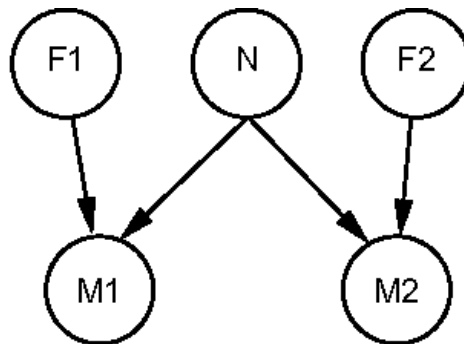


# Exercise 14.12 (a) in AIMA

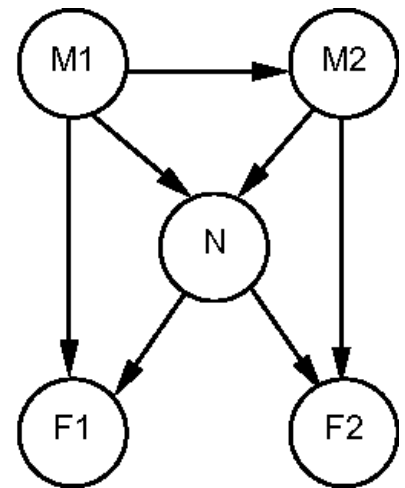
- Two astronomers in different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small region of the sky, using their telescopes. Normally there is a small possibility  $e$  of error up to one star in each direction. Each telescope can also (with a much smaller probability  $f$ ) be badly out of focus (events  $F_1$  and  $F_2$ ) in which case the scientist will undercount by three or more stars (or, if  $N$  is less than 3, fail to detect any stars at all). Consider the three networks in Figure 14.22.
  - (a) Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?



(i)



(ii)

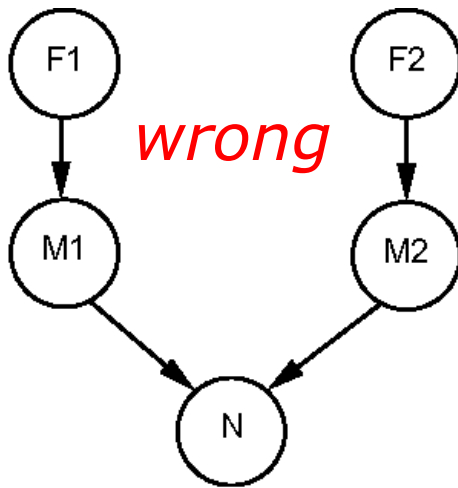


(iii)

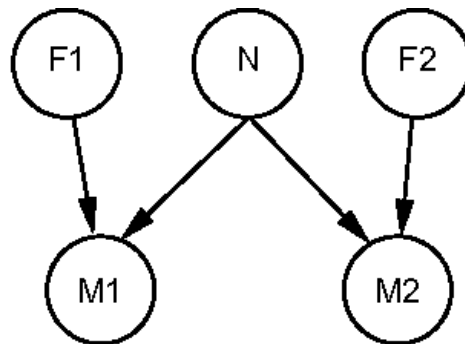
# Exercise 14.12 (a) in AIMA

- (i) must be incorrect –  $N$  is d-separated from  $F_1$  (or  $F_2$ ), relative to  $\{M_1\}$  (or  $\{M_2\}$ ) i.e. knowing the focus states  $F$  would not affect  $N$  if we know  $M$ :

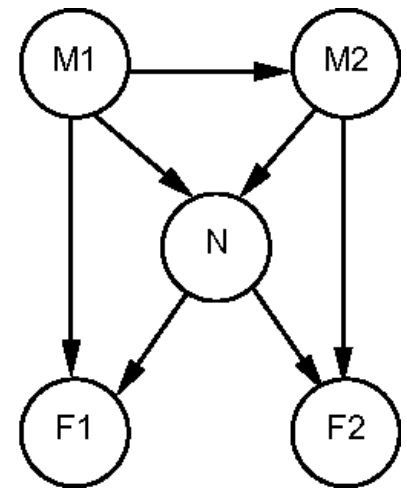
This cannot be correct!



(i)



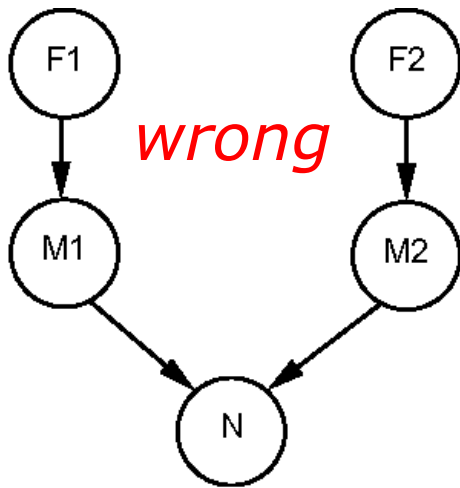
(ii)



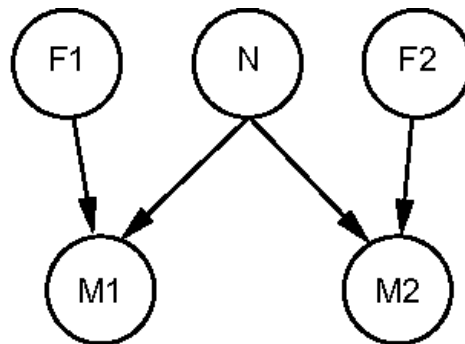
(iii)

# Exercise 14.12 (a) in AIMA

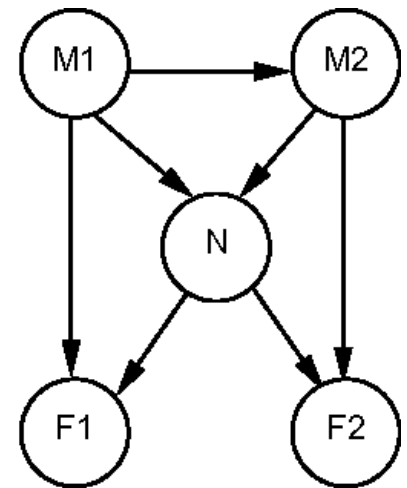
- Two astronomers in different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small region of the sky, using their telescopes. Normally there is a small possibility  $e$  of error up to one star in each direction. Each telescope can also (with a much smaller probability  $f$ ) be badly out of focus (events  $F_1$  and  $F_2$ ) in which case the scientist will undercount by three or more stars (or, if  $N$  is less than 3, fail to detect any stars at all). Consider the three networks in Figure 14.22.
  - (a) Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?



(i)



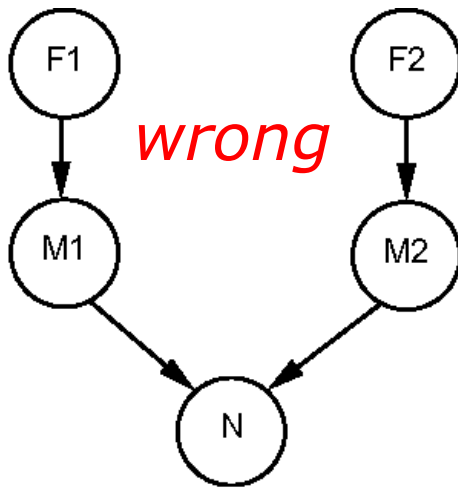
(ii)



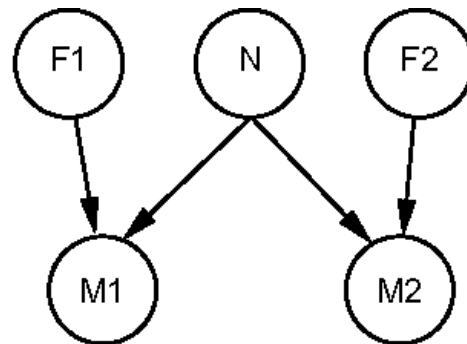
(iii)

# Exercise 14.12 (a) in AIMA

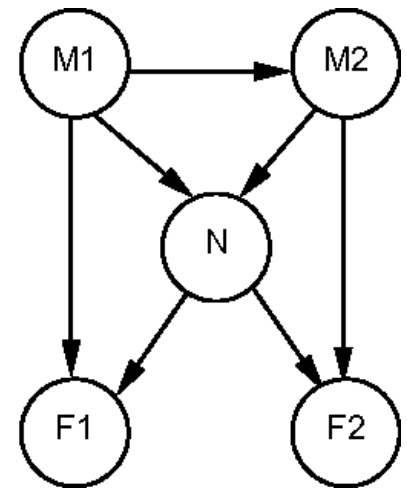
- (ii) is correct – it describes the causal relationships. It is a causal network.



(i)



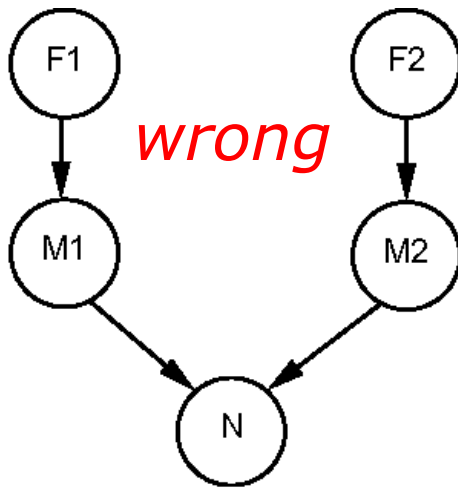
(ii)



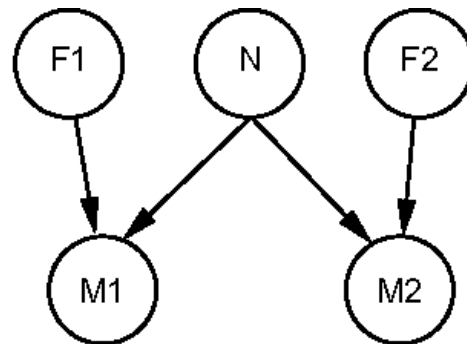
(iii)

# Exercise 14.12 (a) in AIMA

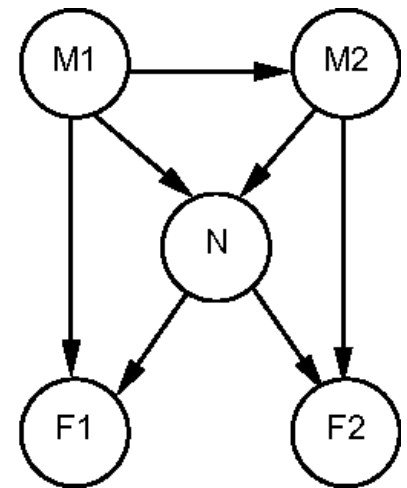
- Two astronomers in different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small region of the sky, using their telescopes. Normally there is a small possibility  $e$  of error up to one star in each direction. Each telescope can also (with a much smaller probability  $f$ ) be badly out of focus (events  $F_1$  and  $F_2$ ) in which case the scientist will undercount by three or more stars (or, if  $N$  is less than 3, fail to detect any stars at all). Consider the three networks in Figure 14.22.
  - (a) Which of these Bayesian networks are correct (but not necessarily efficient) representations of the preceding information?



(i)



(ii)

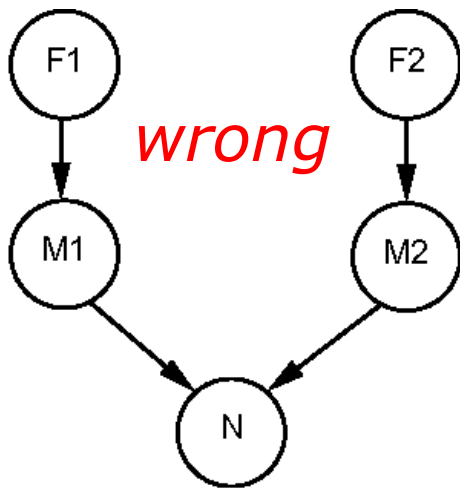


(iii)

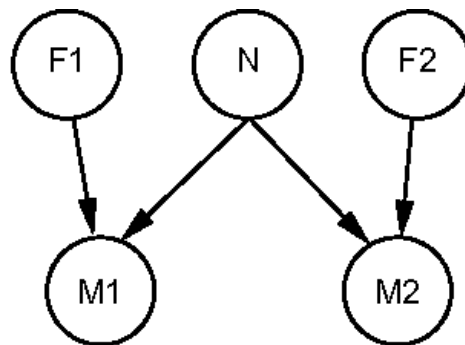


# Exercise 14.12 (a) in AIMA

- (iii) is also ok – a fully connected graph would be correct (but not efficient). (iii) has all connections except  $M_i - F_j$  and  $F_i - F_j$ . (iii) is not causal and not efficient.

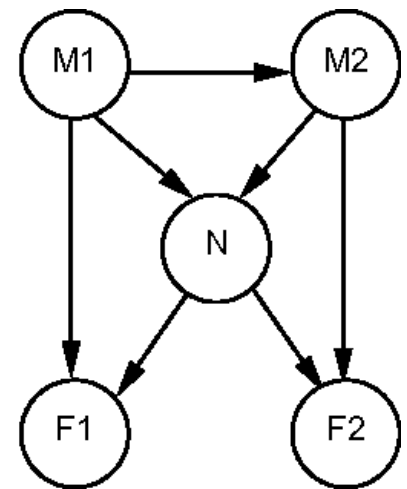


(i)



*ok*

(ii)



(iii)

*ok – but not good*

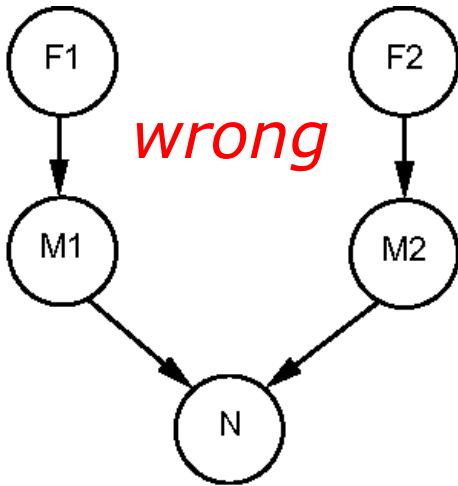
# Exercise 14.12 (a) in AIMA

(ii) says:

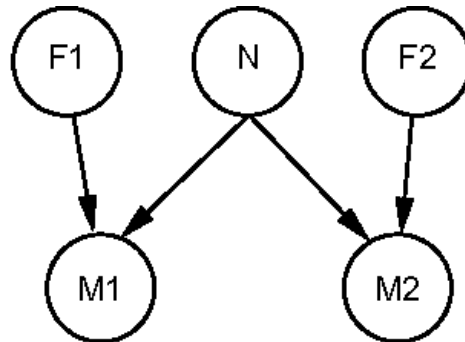
$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(N)P(M1 | F1, N)P(M2 | F2, N)$$

(iii) says:

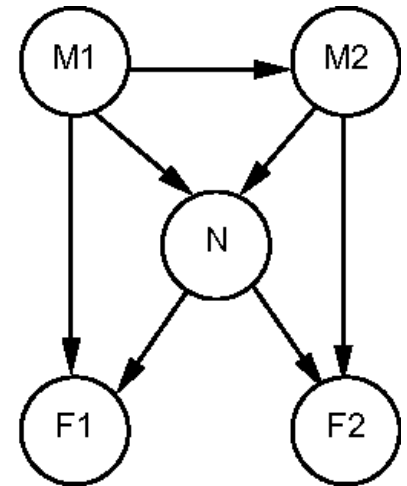
$$P(F1, F2, N, M1, M2) = P(M1)P(F1 | N, M1)P(M2 | M1)P(N | M1, M2)P(F2 | N, M2)$$



(i)



(ii)



(iii)

*ok - but not good*

# Exercise 14.12 (a) in AIMA

(ii) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(N)P(M1 | F1, N)P(M2 | F2, N)$$

The full correct expression (one version) is:

$$\begin{aligned} P(F1, F2, N, M1, M2) &= P(M1 | F1, F2, N, M2)P(F1, F2, N, M2) = \\ &= P(M1 | F1, F2, N, M2)P(M2 | F1, F2, N)P(F1, F2, N) = \\ &= P(M1 | F1, F2, N, M2)P(M2 | F1, F2, N)P(F1 | F2, N)P(F2, N) = \\ &= \underbrace{P(M1 | F1, F2, N, M2)}_{\approx P(M1 | F1, N)} \underbrace{P(M2 | F1, F2, N)}_{\approx P(M2 | F2, N)} \underbrace{P(F1 | F2, N)}_{\approx P(F1)} \underbrace{P(F2 | N)}_{\approx P(F2)} P(N) \end{aligned}$$

*ok*

# Exercise 14.12 (a) in AIMA

This is not as efficient as (ii). This requires more conditional probabilities.

(iii) says:

$$P(F1, F2, N, M1, M2) = P(M1)P(F1 | N, M1)P(M2 | M1)P(N | M1, M2)P(F2 | N, M2)$$

The full correct expression (another version) is:

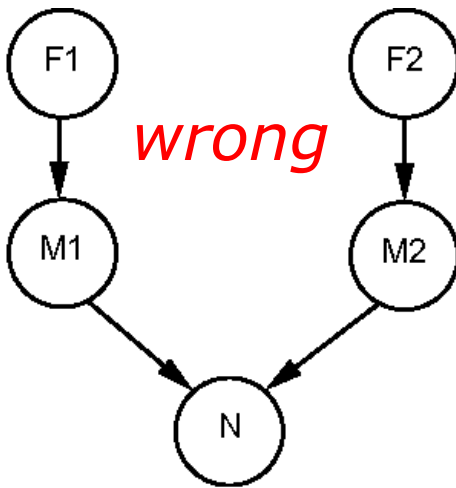
$$\begin{aligned} P(F1, F2, N, M1, M2) &= P(F1 | F2, N, M1, M2)P(F2, N, M1, M2) = \\ &P(F1 | F2, N, M1, M2)P(F2 | N, M1, M2)P(N, M1, M2) = \\ &P(F1 | F2, N, M1, M2)P(F2 | N, M1, M2)P(N | M1, M2)P(M1, M2) = \\ &\underbrace{P(F1 | F2, N, M1, M2)}_{\approx P(F1 | N, M1)} \underbrace{P(F2 | N, M1, M2)}_{\approx P(F2 | N, M2)} P(N | M1, M2)P(M2 | M1)P(M1) \end{aligned}$$

ok

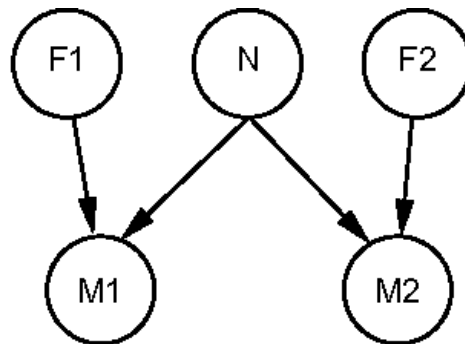
# Exercise 14.12 (a) in AIMA

(i) says:

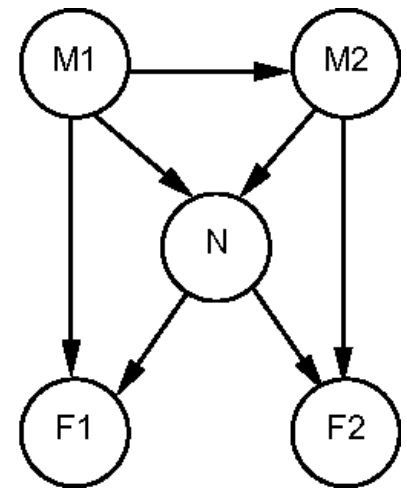
$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(M1 | F1)P(M2 | F2)P(N | M1, M2)$$



(i)



(ii)



(iii)

*ok – but not good*

# Exercise 14.12 (a) in AIMA

(i) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(M1 | F1)P(M2 | F2)P(N | M1, M2)$$

The full correct expression (a third version) is:

$$\begin{aligned} P(F1, F2, N, M1, M2) &= P(N | M1, M2, F1, F2)P(M1, M2, F1, F2) = \\ &= P(N | M1, M2, F1, F2)P(M1 | M2, F1, F2)P(M2, F1, F2) = \\ &= P(N | M1, M2, F1, F2)P(M1 | M2, F1, F2)P(M2 | F1, F2)P(F1, F2) = \\ &= \underbrace{P(N | M1, M2, F1, F2)}_{\approx P(N | M1, M2)} \underbrace{P(M1 | M2, F1, F2)}_{\approx P(M1 | F1)} \underbrace{P(M2 | F1, F2)}_{\approx P(M2 | F2)} \underbrace{P(F1 | F2)P(F2)}_{\approx P(F1)} \end{aligned}$$

# Exercise 14.12 (a) in AIMA

(i) says:

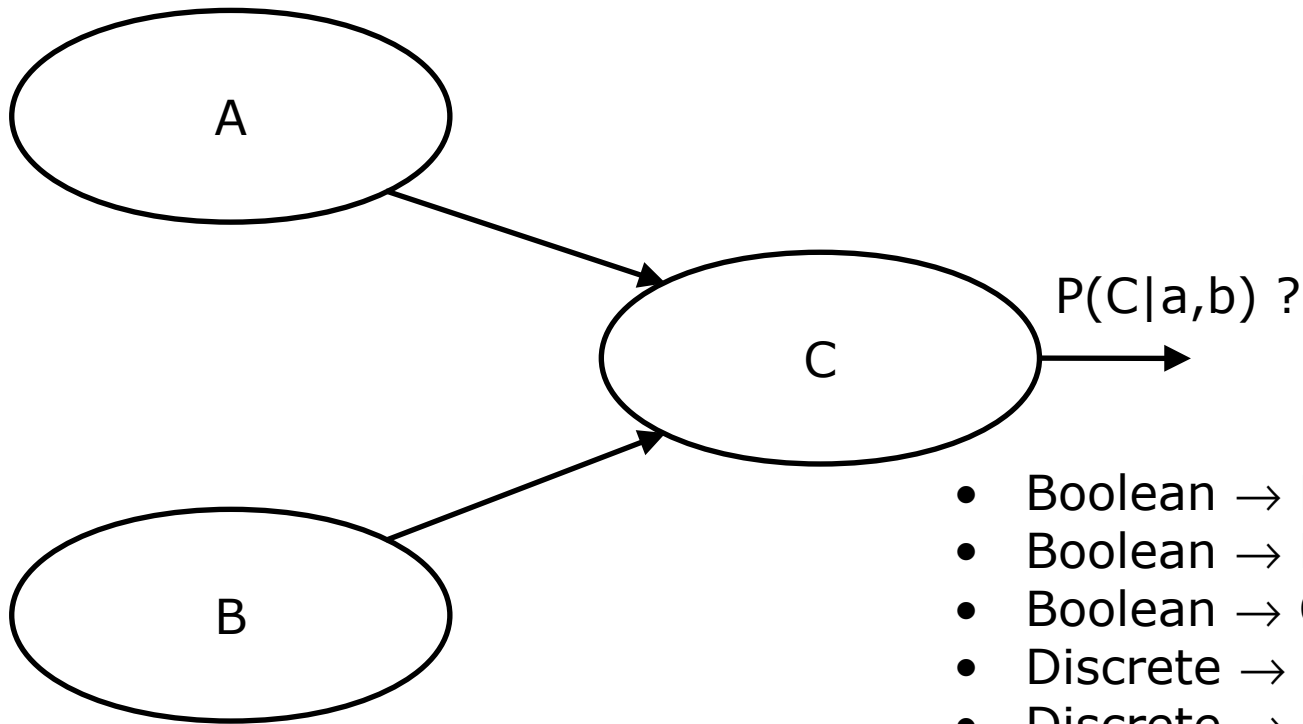
$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(M1 | F1)P(M2 | F2)P(N | M1, M2)$$

This is an unreasonable approximation. The rest is ok.

The full correct expression (a third version) is:

$$\begin{aligned} P(F1, F2, N, M1, M2) &= P(N | M1, M2, F1, F2)P(M1, M2, F1, F2) = \\ &= P(N | M1, M2, F1, F2)P(M1 | M2, F1, F2)P(M2, F1, F2) = \\ &= P(N | M1, M2, F1, F2)P(M1 | M2, F1, F2)P(M2 | F1, F2)P(F1, F2) = \\ &= \underbrace{P(N | M1, M2, F1, F2)}_{\approx P(N | M1, M2)} \underbrace{P(M1 | M2, F1, F2)}_{\approx P(M1 | F1)} \underbrace{P(M2 | F1, F2)}_{\approx P(M2 | F2)} \underbrace{P(F1, F2)}_{\approx P(F1)} \end{aligned}$$

# Efficient representation of PDs



- Boolean  $\rightarrow$  Boolean
- Boolean  $\rightarrow$  Discrete
- Boolean  $\rightarrow$  Continuous
- Discrete  $\rightarrow$  Boolean
- Discrete  $\rightarrow$  Discrete
- Discrete  $\rightarrow$  Continuous
- Continuous  $\rightarrow$  Boolean
- Continuous  $\rightarrow$  Discrete
- Continuous  $\rightarrow$  Continuous



# Efficient representation of PDs

Boolean → Boolean:

Noisy-OR, Noisy-AND

Boolean/Discrete → Discrete:

Noisy-MAX

Bool./Discr./Cont. → Continuous:

Parametric distribution (e.g. Gaussian)

Continuous → Boolean:

Logit/Probit

# Noisy-OR example

Boolean  $\rightarrow$  Boolean

P(E C <sub>1</sub> ,C <sub>2</sub> ,C <sub>3</sub> )								
C <sub>1</sub>	0	1	0	0	1	1	0	1
C <sub>2</sub>	0	0	1	0	1	0	1	1
C <sub>3</sub>	0	0	0	1	0	1	1	1
P(E=0)	1	0.1	0.1	0.1	0.01	0.01	0.01	0.001
P(E=1)	0	0.9	0.9	0.9	0.99	0.99	0.99	0.999

The effect (E) is off (false) when none of the causes are true. The probability for the effect increases with the number of true causes.

$$P(E = 0) = 10^{-(\#True)} \quad (\text{for this example})$$

# Noisy-OR general case

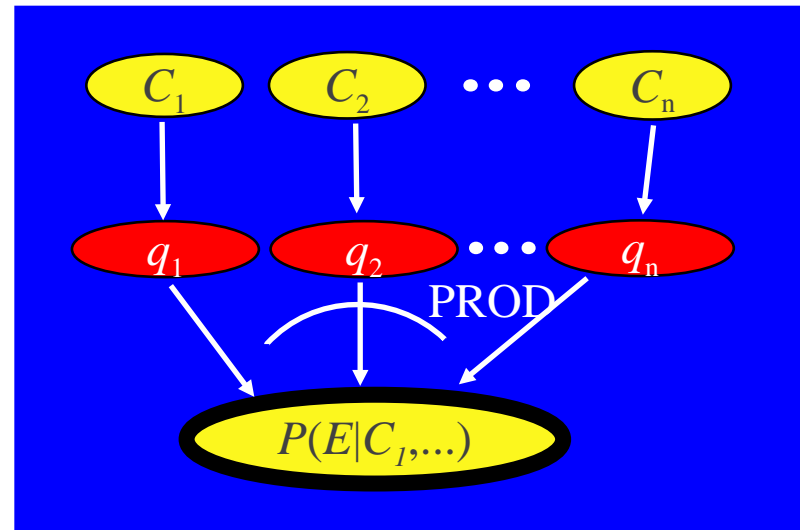
Boolean  $\rightarrow$  Boolean

$$P(E = 0 | C_1, C_2, \dots, C_n) = \prod_{i=1}^n q_i^{C_i}$$

$$C_i = \begin{cases} 1 & \text{if true} \\ 0 & \text{if false} \end{cases}$$

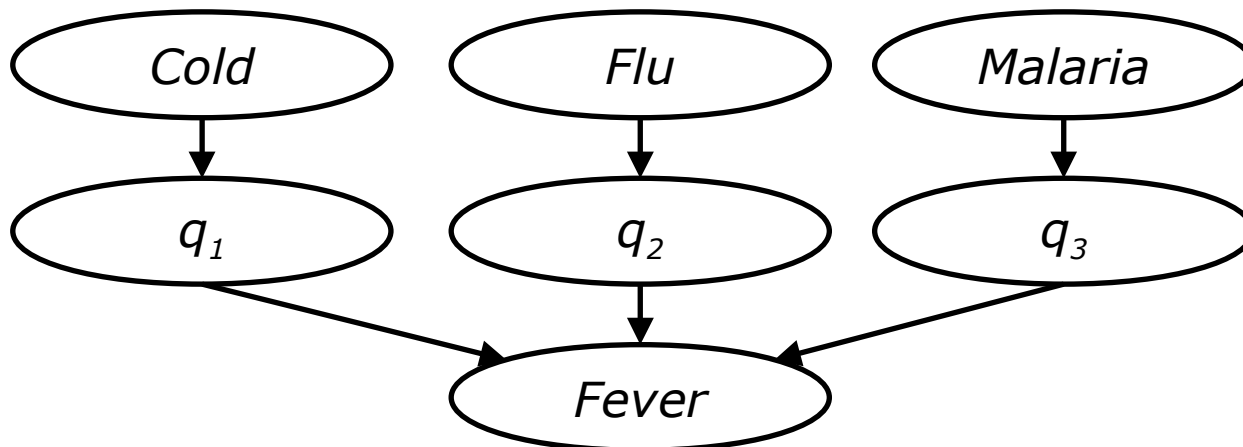
Example on previous slide used  
 $q_i = 0.1$  for all  $i$ .

*Needs only  $n$  parameters,  
not  $2^n$  parameters.*



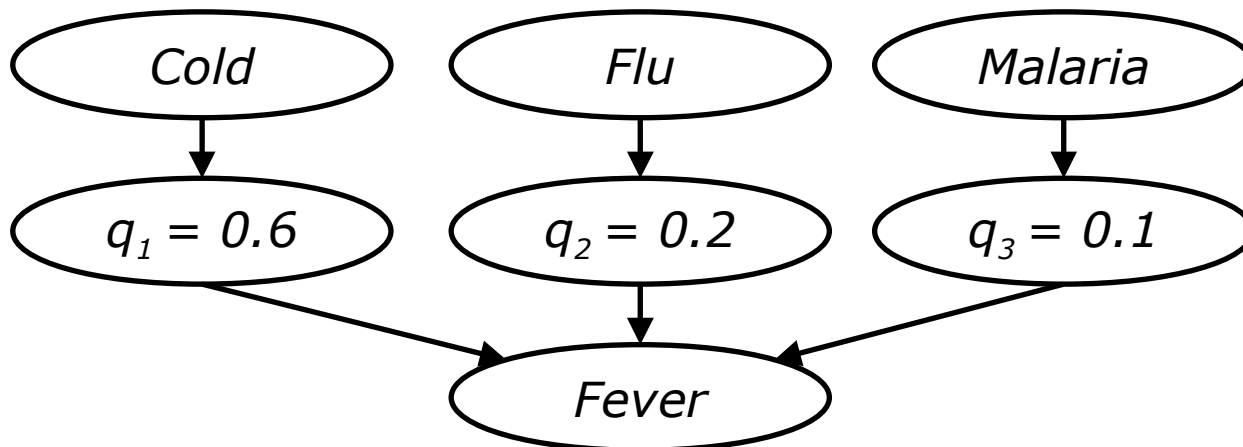
# Noisy-OR example (II)

- *Fever* is True if and only if *Cold*, *Flu* or *Malaria* is True.
- each cause has an independent chance of causing the effect.
  - all possible causes are listed
  - inhibitors are independent



# Noisy-OR example (II)

- $P(\text{Fever} \mid \text{Cold}) = 0.4 \Rightarrow q_1 = 0.6$
- $P(\text{Fever} \mid \text{Flu}) = 0.8 \Rightarrow q_2 = 0.2$
- $P(\text{Fever} \mid \text{Malaria}) = 0.9 \Rightarrow q_3 = 0.1$

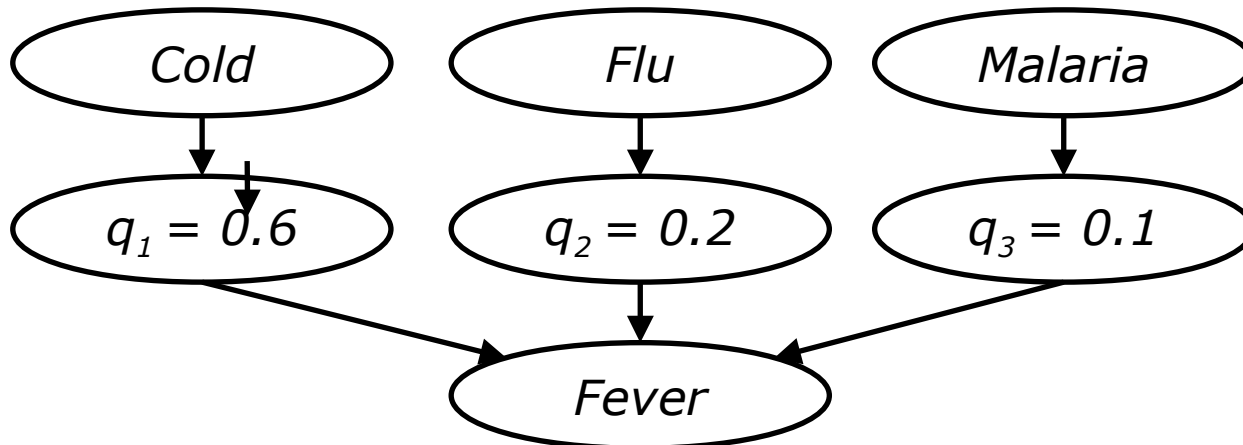


# Noisy-OR example (II)

- $P(\text{Fever} \mid \text{Cold}) = 0.4 \Rightarrow q_1 = 0.6$
- $P(\text{Fever} \mid \text{Flu}) = 0.8 \Rightarrow q_2 = 0.2$
- $P(\text{Fever} \mid \text{Malaria}) = 0.9 \Rightarrow q_3 = 0.1$

$$P(\neg \text{Fever} \mid \neg \text{Cold}, \neg \text{Flu}, \neg \text{Malaria}) = 0.6^0 \times 0.2^0 \times 0.1^0 = 1$$

$$P(\text{Fever} \mid \neg \text{Cold}, \neg \text{Flu}, \neg \text{Malaria}) = 1 - 1 = 0$$

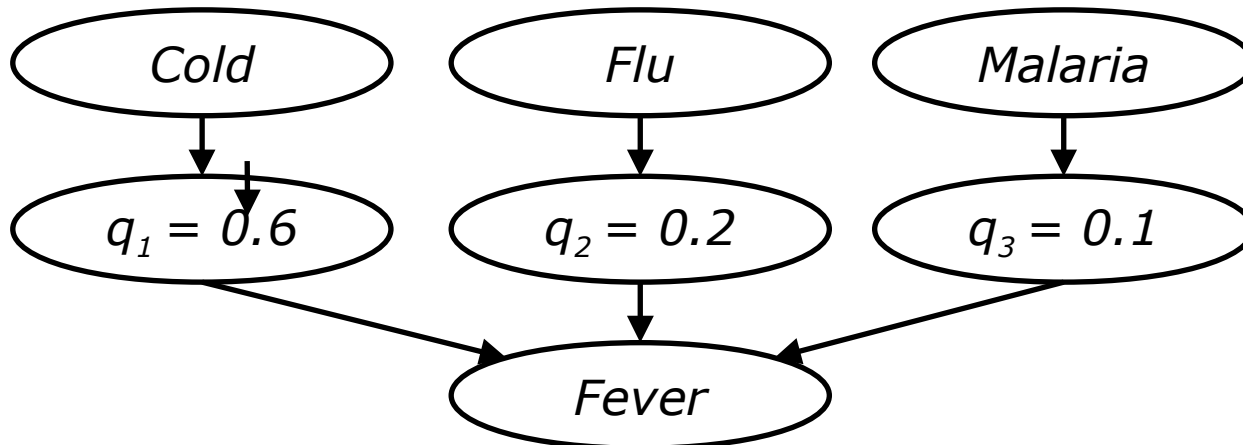


# Noisy-OR example (II)

- $P(\text{Fever} \mid \text{Cold}) = 0.4 \Rightarrow q_1 = 0.6$
- $P(\text{Fever} \mid \text{Flu}) = 0.8 \Rightarrow q_2 = 0.2$
- $P(\text{Fever} \mid \text{Malaria}) = 0.9 \Rightarrow q_3 = 0.1$

$$P(\neg \text{Fever} \mid \neg \text{Cold}, \neg \text{Flu}, \text{Malaria}) = 0.6^0 \times 0.2^0 \times 0.1^1 = 0.1$$

$$P(\text{Fever} \mid \neg \text{Cold}, \neg \text{Flu}, \text{Malaria}) = 1 - 0.1 = 0.9$$

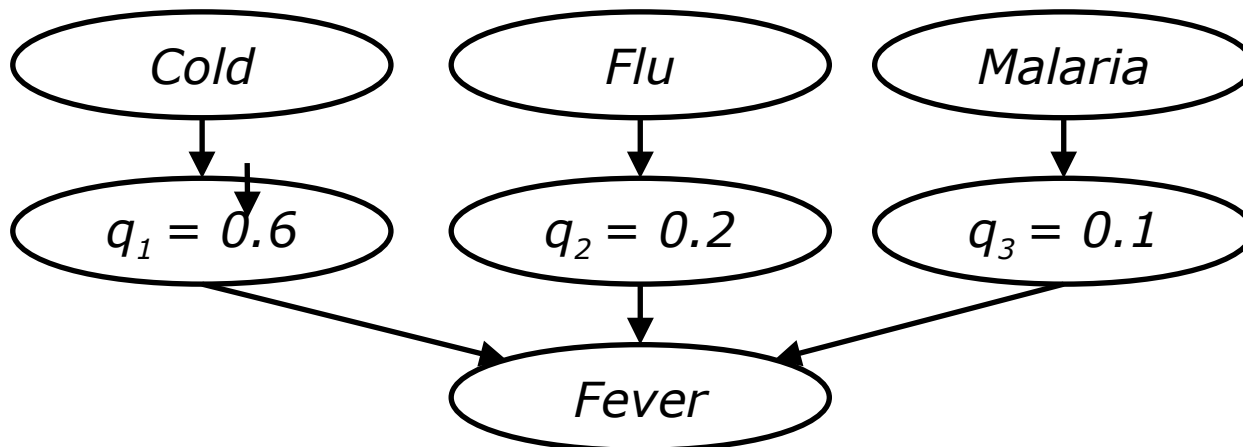


# Noisy-OR example (II)

- $P(\text{Fever} \mid \text{Cold}) = 0.4 \Rightarrow q_1 = 0.6$
- $P(\text{Fever} \mid \text{Flu}) = 0.8 \Rightarrow q_2 = 0.2$
- $P(\text{Fever} \mid \text{Malaria}) = 0.9 \Rightarrow q_3 = 0.1$

$$P(\neg \text{Fever} \mid \text{Cold}, \text{Flu}, \neg \text{Malaria}) = 0.6^1 \times 0.2^1 \times 0.1^0 = 0.12$$

$$P(\text{Fever} \mid \text{Cold}, \text{Flu}, \neg \text{Malaria}) = 1 - 0.12 = 0.88$$





# Parametric probability densities

Boolean/Discr./Continuous → Continuous

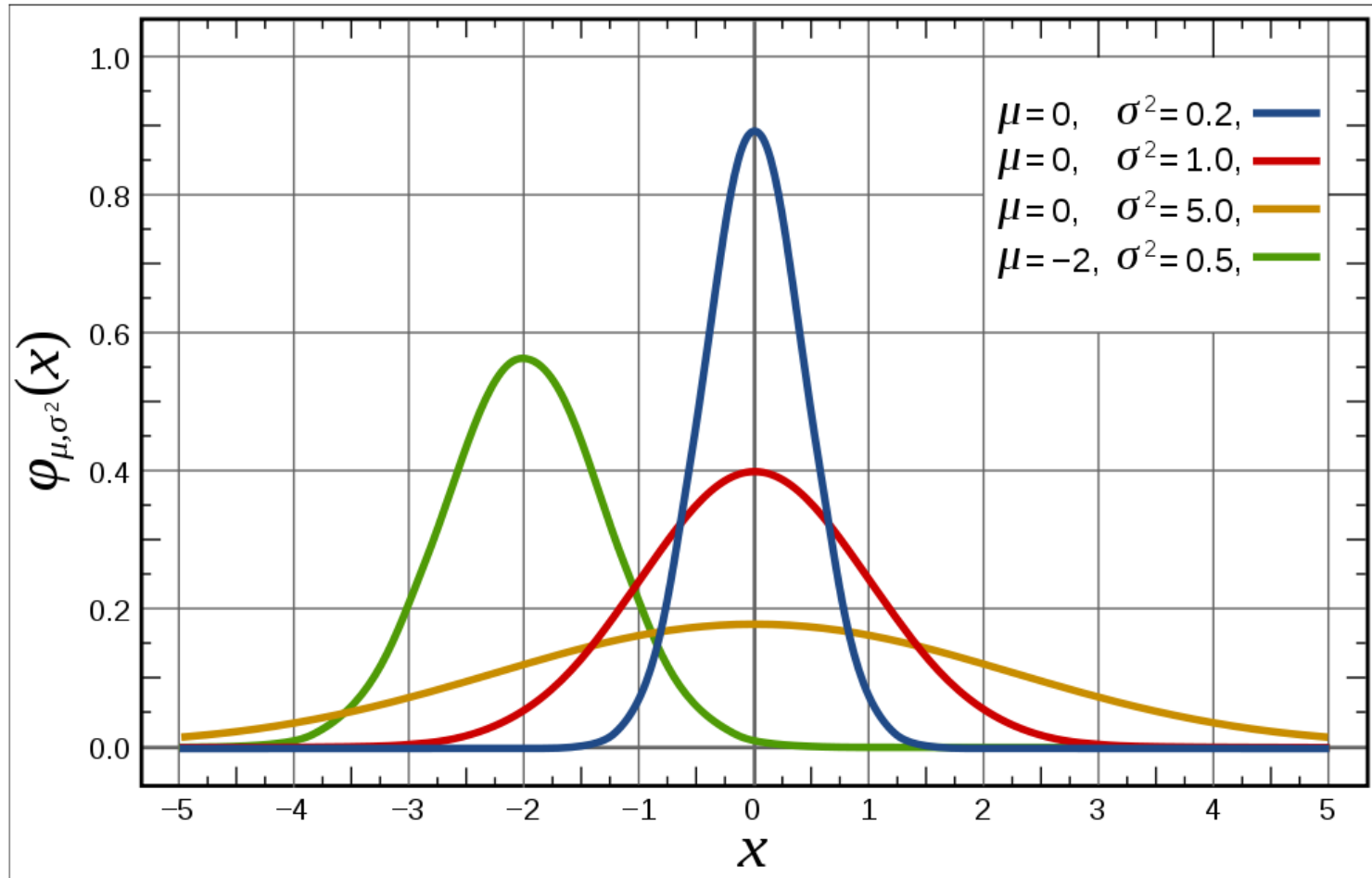
Use parametric probability densities, e.g.,  
the normal distribution

$$P(X) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) = N(\mu, \sigma)$$

Gaussian networks ( $a$  = input to the node)

$$P(X) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[\frac{-(x - \alpha - \beta a)^2}{2\sigma^2}\right]$$

# Normal Distribution



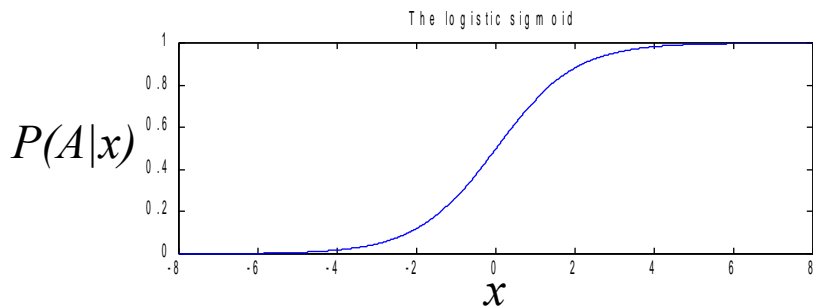
# Probit & Logit

Discrete → Boolean

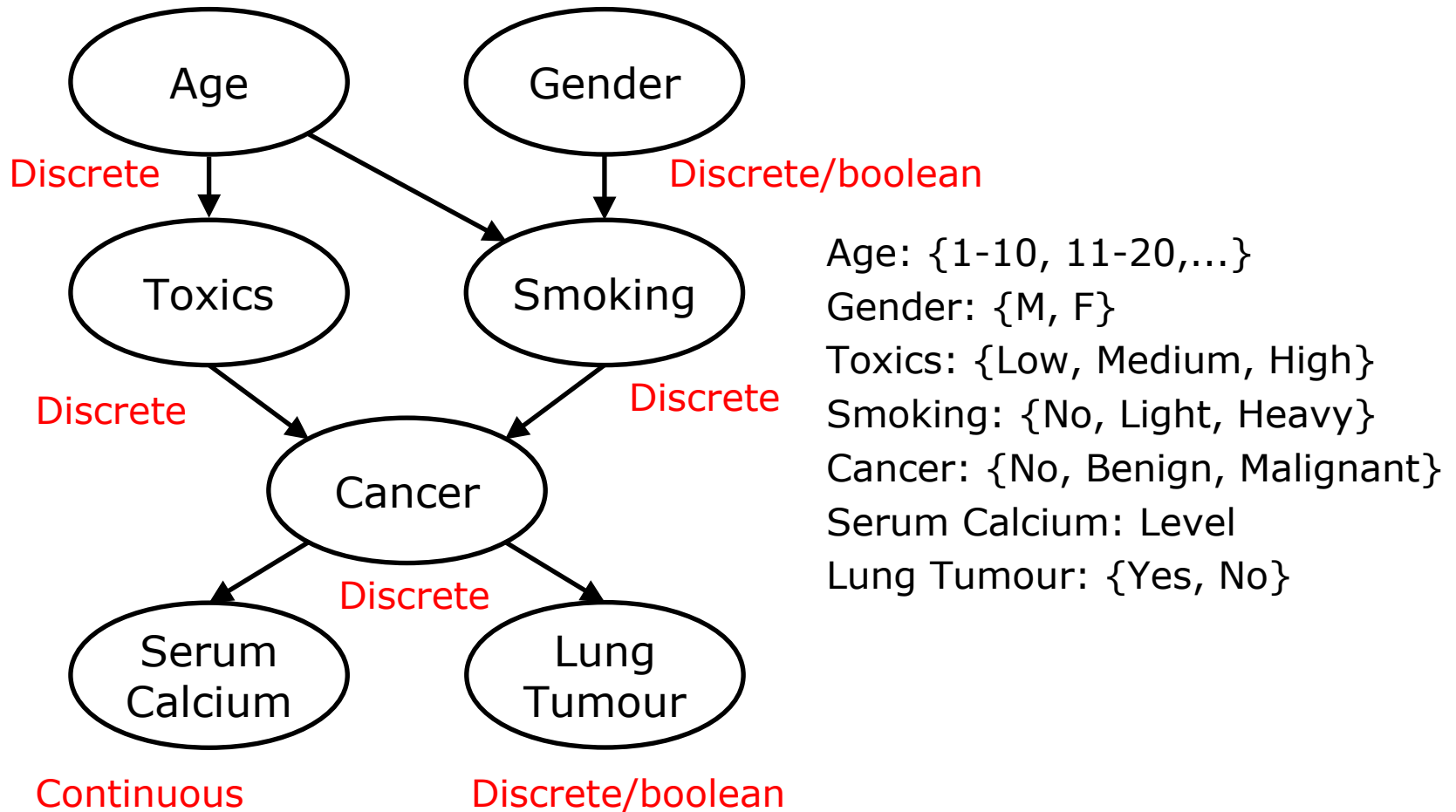
If the input is continuous but output is boolean, use probit or logit

$$\text{Logit: } P(A = a | x) = \frac{1}{1 + \exp[-2(\mu - x) / \sigma]}$$

$$\text{Probit: } P(A = a | x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-(x - \mu)^2 / \sigma^2) dx$$



# The cancer network



# Inference in BN

Inference means computing  $\mathbf{P}(X|\mathbf{e})$ , where  $X$  is a query (variable) and  $\mathbf{e}$  is a set of evidence variables (for which we know the values).

Examples:

$\mathbf{P}(\text{Burglary} \mid \text{john\_calls}, \text{mary\_calls})$

$\mathbf{P}(\text{Cancer} \mid \text{age}, \text{gender}, \text{smoking}, \text{serum\_calcium})$

$\mathbf{P}(\text{Cavity} \mid \text{toothache}, \text{catch})$

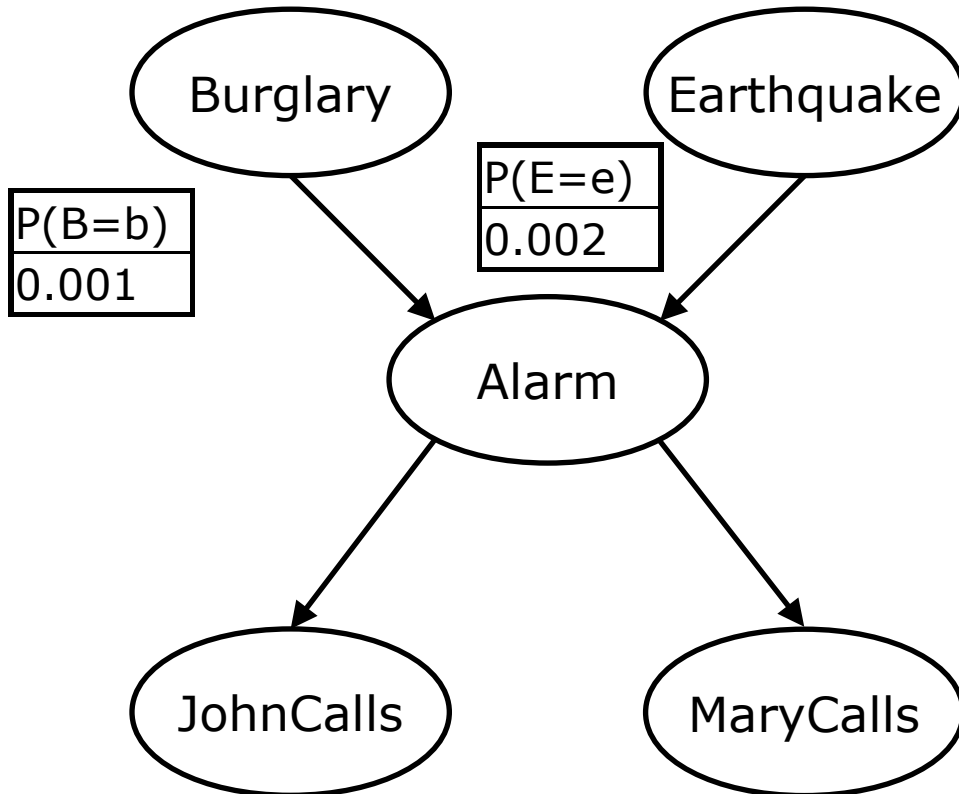
# Exact inference in BN

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

“Doable” for boolean variables: Look up entries in conditional probability tables (CPTs).

# Example: The alarm network

What is the probability for a burglary if both John and Mary call?  $P(B | j, m) = \alpha \sum_{E=\{e, \neg e\}} \sum_{A=\{a, \neg a\}} P(B, E, A, j, m)$



Evidence variables = {J,M}  
 Query variable = B

$P(B=b)$
0.001

$P(E=e)$
0.002

A	$P(J=j)$
a	0.90
$\neg a$	0.05

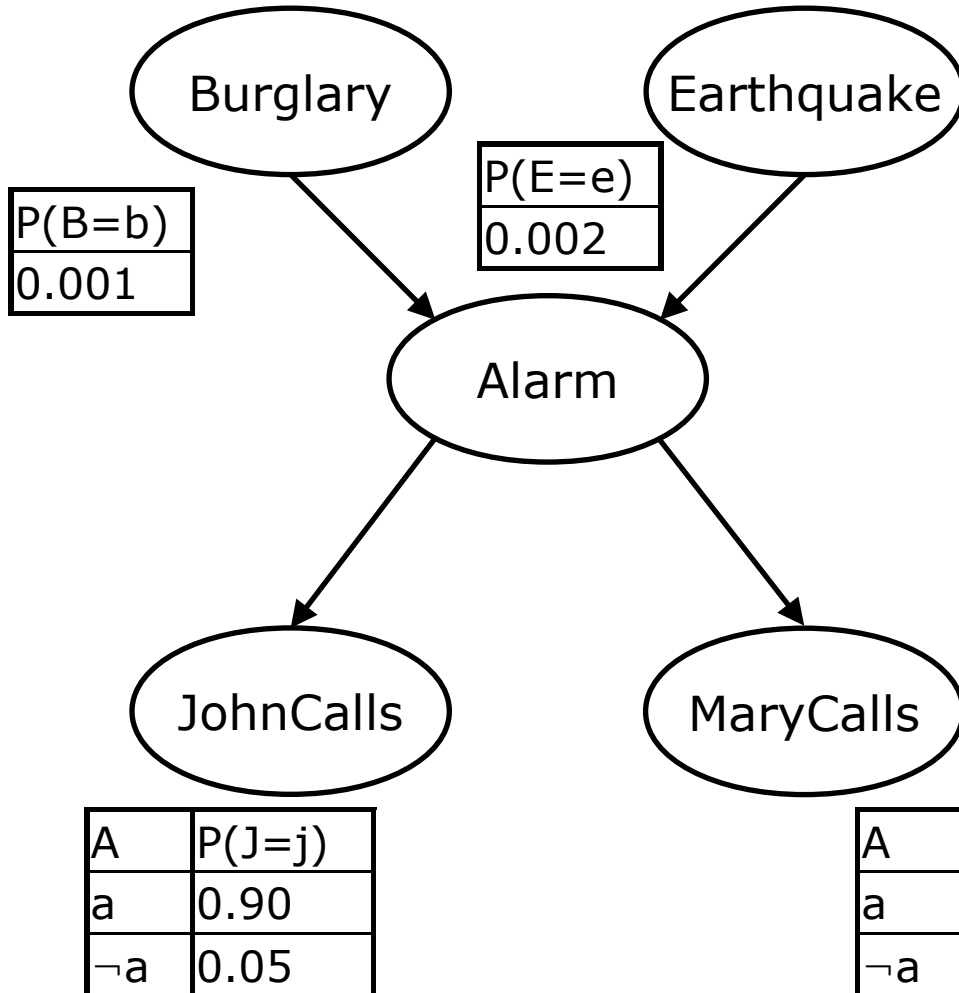
A	$P(M=m)$
a	0.70
$\neg a$	0.01

B	E	$P(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001

# Example: The alarm network

What is the probability for a burglary if both John and Mary call?

$$\mathbf{P}(B | j, m) = \alpha \sum_{E = \{e, \neg e\}} \sum_{A = \{a, \neg a\}} \mathbf{P}(B, E, A, j, m)$$



$$\mathbf{P}(B = b, E, A, j, m) =$$

$$\mathbf{P}(j, m | b, E, A) \mathbf{P}(b, E, A) =$$

$$P(j | A) P(m | A) \mathbf{P}(b, E, A) =$$

$$P(j | A) P(m | A) P(A | b, E) \mathbf{P}(b, E) =$$

$$P(j | A) P(m | A) P(A | b, E) \underbrace{P(b) P(E)}_{= 0.001 = 10^{-3}}$$

$$= 0.001 = 10^{-3}$$

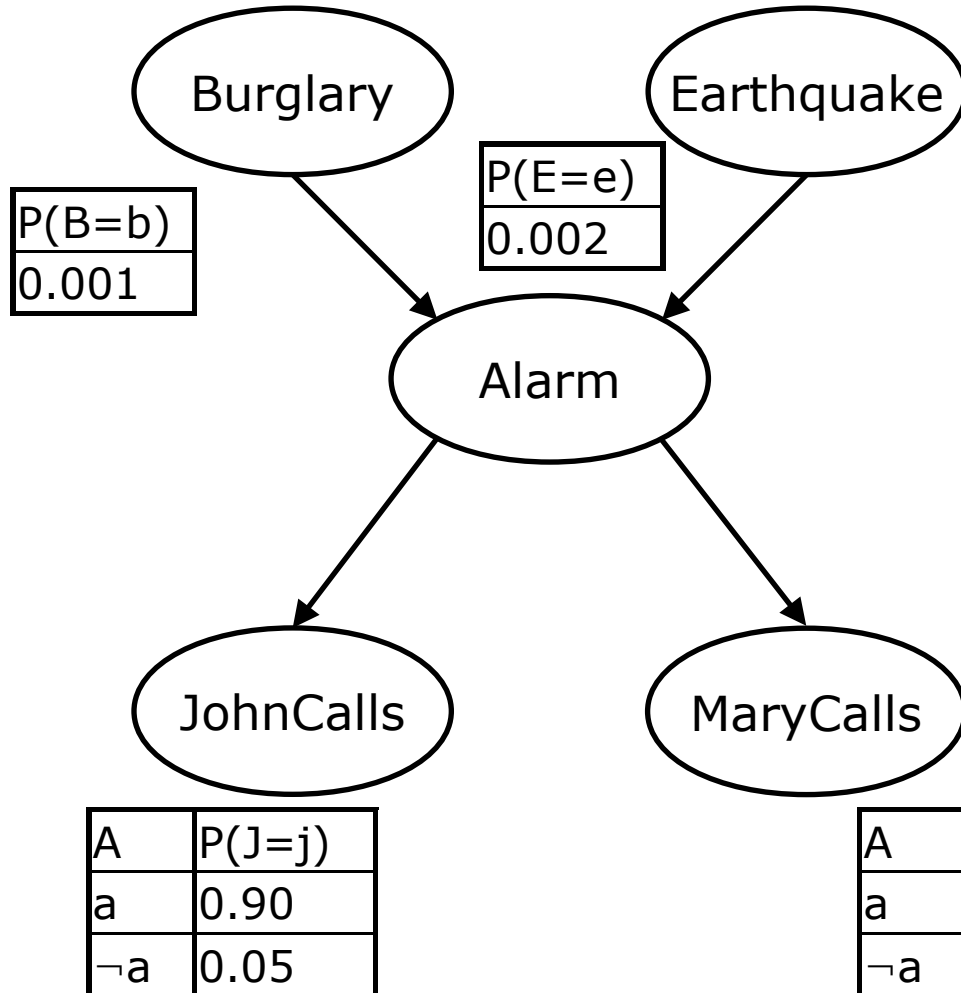
$$10^{-3} \times P(j | A) P(m | A) P(A | b, E) P(E)$$

B	E	$P(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001



# Example: The alarm network

What is the probability for a burglary if both John and Mary call?



$$P(B | j, m) = \alpha \sum_{E=\{e, \neg e\}} \sum_{A=\{a, \neg a\}} P(B, E, A, j, m)$$

$$P(b, j, m) = 10^{-3} \sum_{\substack{A=\{a, \neg a\} \\ E=\{e, \neg e\}}} P(j | A) P(m | A) P(A | b, E) P(E) = 10^{-3} [P(j | a) P(m | a) P(a | b, e) P(e) + P(j | a) P(m | a) P(a | b, \neg e) P(\neg e) + P(j | \neg a) P(m | \neg a) P(\neg a | b, e) P(e) + P(j | \neg a) P(m | \neg a) P(\neg a | b, \neg e) P(\neg e)] = 0.5923 \times 10^{-3}$$

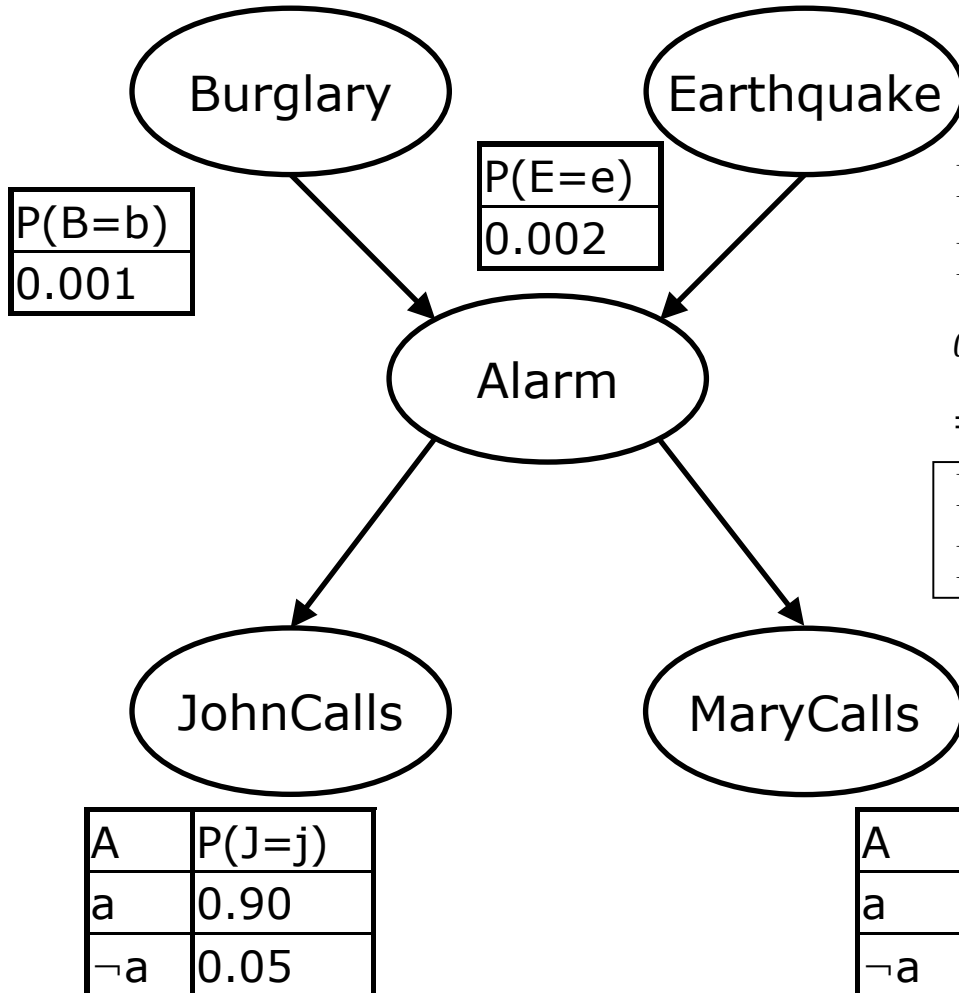
$$P(\neg b, j, m) = 1.491 \times 10^{-3}$$

B	E	$P(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001

# Example: The alarm network

What is the probability for a burglary if both John and Mary call?

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



$$\mathbf{P}(b, j, m) = 0.5923 \times 10^{-3}$$

$$\mathbf{P}(\neg b, j, m) = 1.491 \times 10^{-3}$$

$$\alpha = \mathbf{P}(j, m)^{-1} = [\mathbf{P}(b, j, m) + \mathbf{P}(\neg b, j, m)]^{-1} = [2.083 \times 10^{-3}]^{-1}$$

$$\mathbf{P}(b \mid j, m) = \alpha \mathbf{P}(b, j, m) = 0.284$$

$$\mathbf{P}(\neg b \mid j, m) = \alpha \mathbf{P}(\neg b, j, m) = 0.716$$

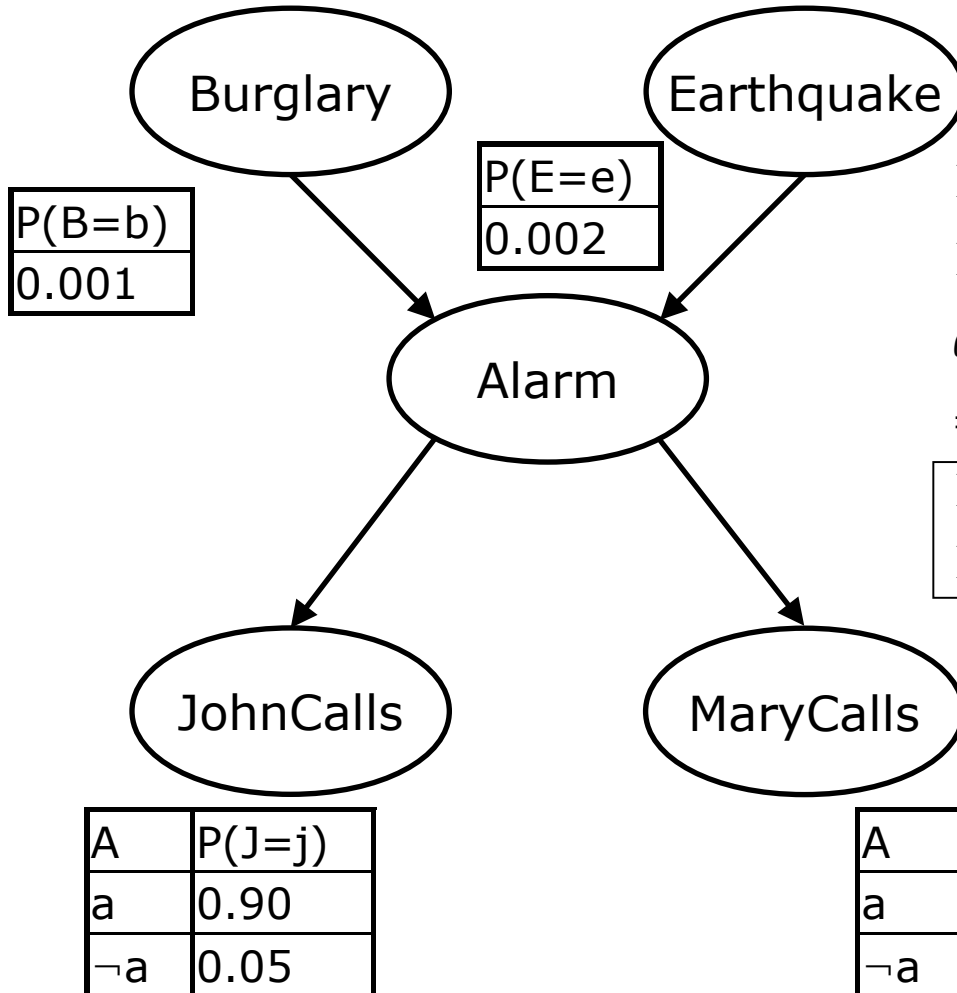
B	E	$\mathbf{P}(A=a)$
b	e	0.95
b	$\neg e$	0.94
$\neg b$	e	0.29
$\neg b$	$\neg e$	0.001

# Example: The alarm network

What is the probability for a burglary if both John and Mary call?

**Answer: 28%**

$$P(B | j, m) = \alpha \sum_{E=\{e, \neg e\}} \sum_{A=\{a, \neg a\}} P(B, E, A, j, m)$$



$$P(b, j, m) = 0.5923 \times 10^{-3}$$

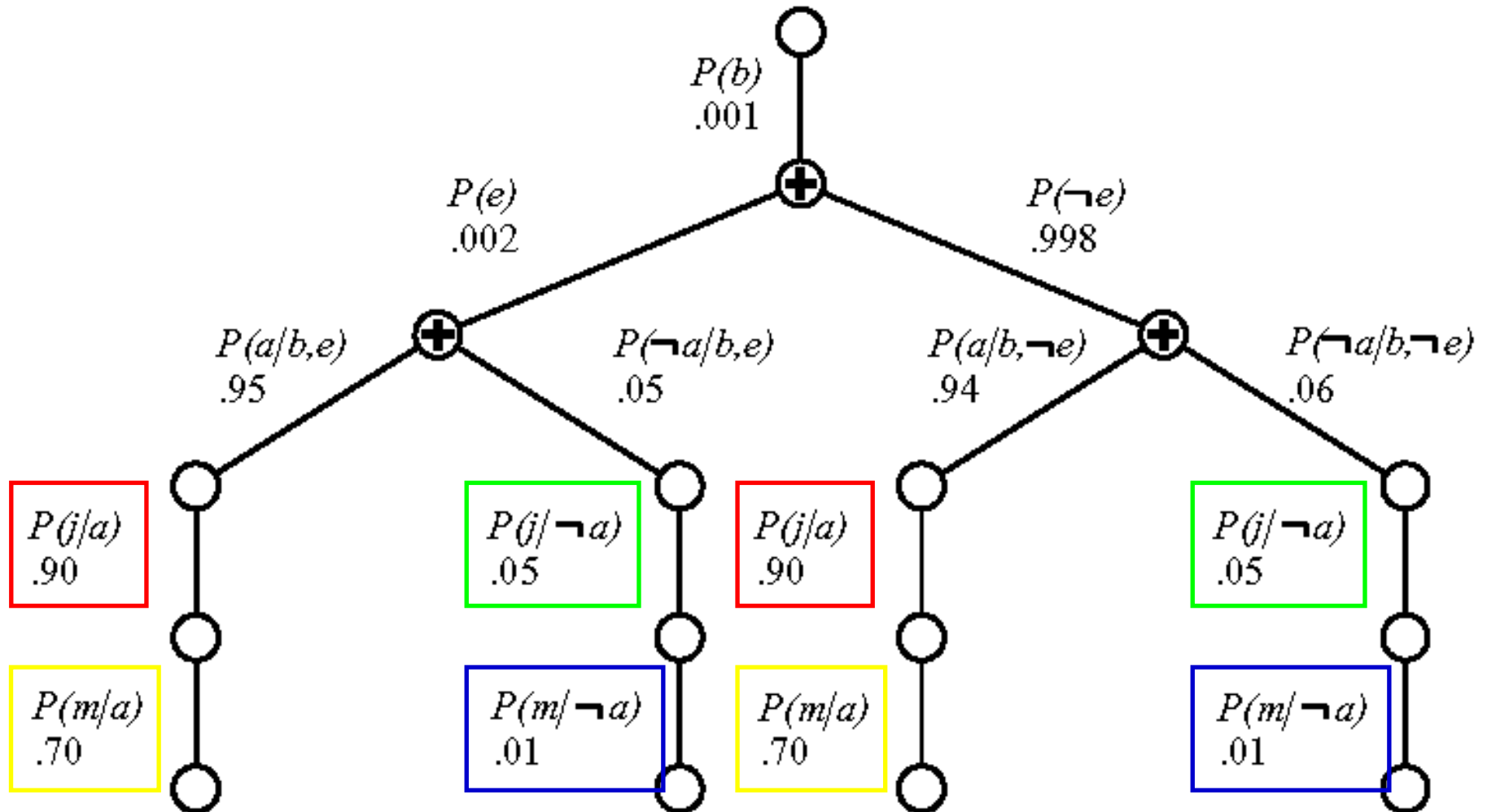
$$P(\neg b, j, m) = 1.491 \times 10^{-3}$$

$$\alpha = P(j, m)^{-1} = [P(b, j, m) + P(\neg b, j, m)]^{-1} = [2.083 \times 10^{-3}]^{-1}$$

$$P(b | j, m) = \alpha P(b, j, m) = 0.284$$

$$P(\neg b | j, m) = \alpha P(\neg b, j, m) = 0.716$$

# Use depth-first search



A lot of unnecessary repeated computation...

# Complexity of exact inference

- By eliminating repeated calculation & uninteresting paths we can speed up the inference a lot.
- Linear time complexity for singly connected networks (polytrees).
- Exponential for multiply connected networks.
  - Clustering can improve this

# Approximate inference in BN

- Exact inference is intractable in large multiply connected BNs  $\Rightarrow$  use approximate inference:  
Monte Carlo methods (random sampling).
  - Direct sampling
  - Rejection sampling
  - Likelihood weighting
  - Markov chain Monte Carlo

# Markov chain Monte Carlo

1. Fix the evidence variables ( $E_1, E_2, \dots$ ) at their given values.
2. Initialize the network with values for all other variables, including the query variable.
3. Repeat the following many, many, many times:
  - a. Pick a non-evidence variable at random (query  $X_i$  or hidden  $Y_j$ )
  - b. Select a new value for this variable, conditioned on the current values in the variable's Markov blanket.

Monitor the values of the query variables.