Cooperating Intelligent Systems

Bayesian networks

Chapter 14, AIMA

Inference

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

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

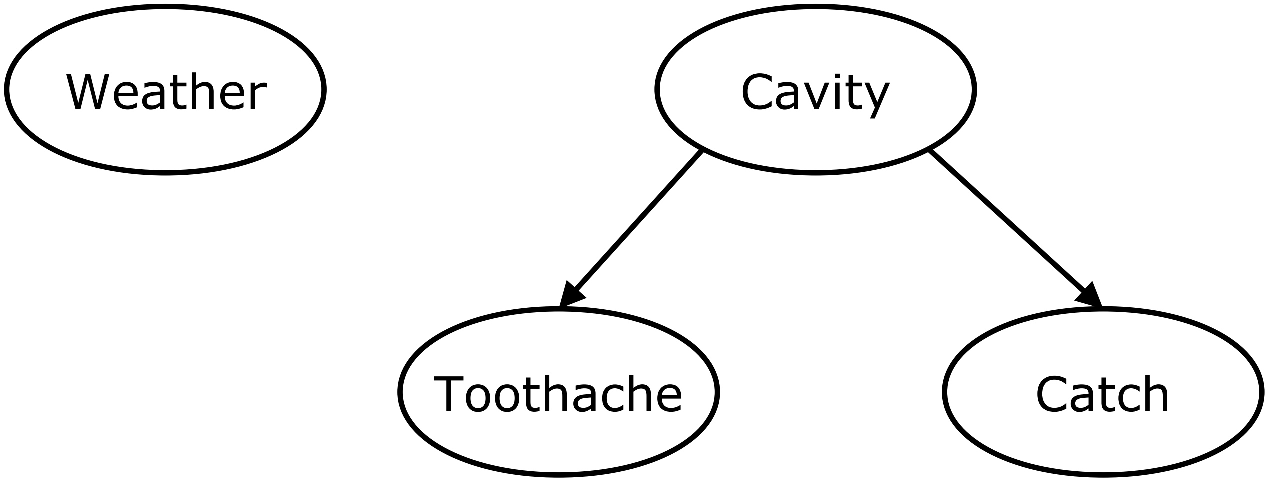
- We need an efficient method for doing this which is more widely applicable than the naïve Bayes model

Bayesian networks

A **Bayesian network** is a directed graph in which each node is annotated with quantitative probability information:

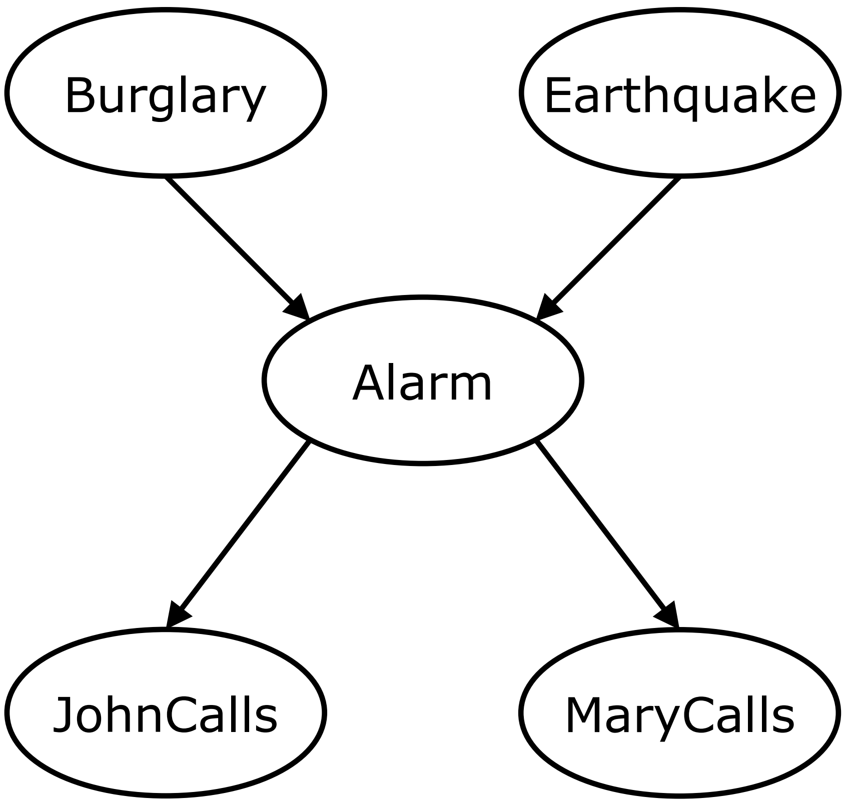
1. The set of nodes of the network corresponds to a set of random variables $\{X_1, X_2, X_3, \ldots\}$
2. pairs of nodes can be connected by directed links defining a parent $\rightarrow$ child relation
3. Each node X_i contains a conditional probability distribution $\mathbf{P}(X_i \mid Parents(X_i))$
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



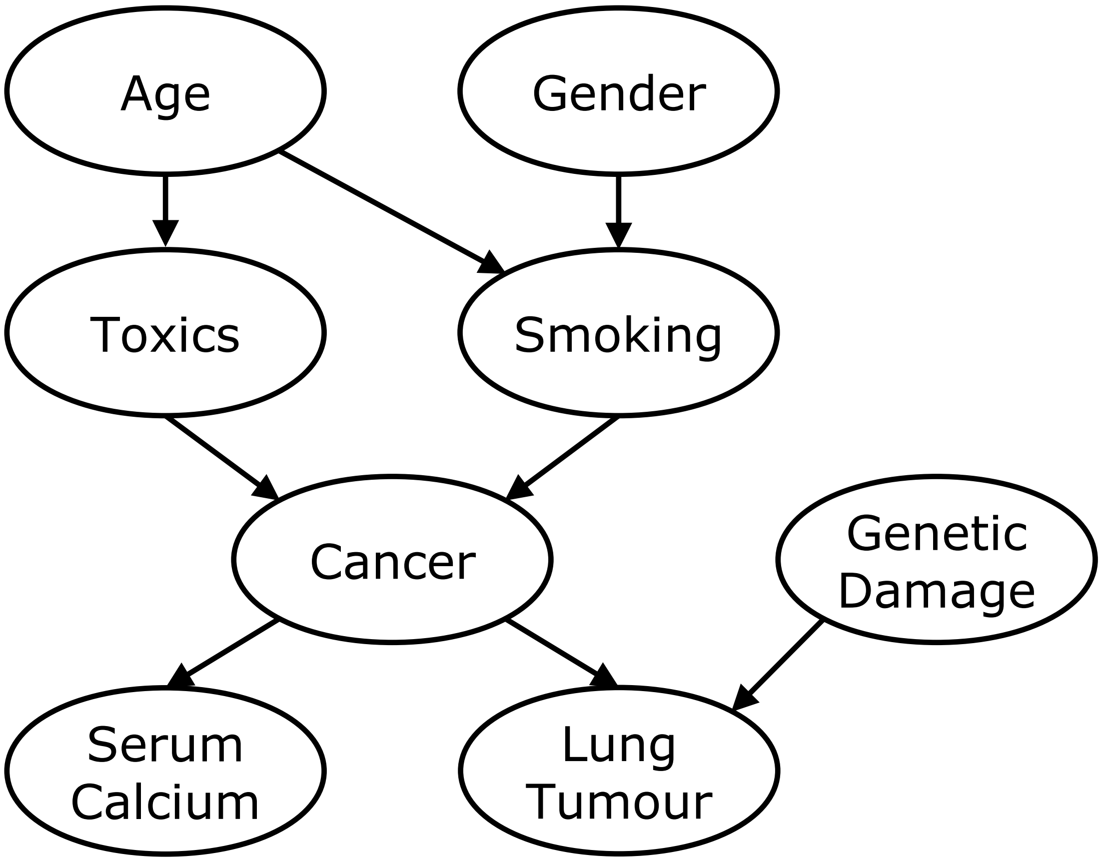
Burglar alarm responds to both earthquakes and burglars.

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).

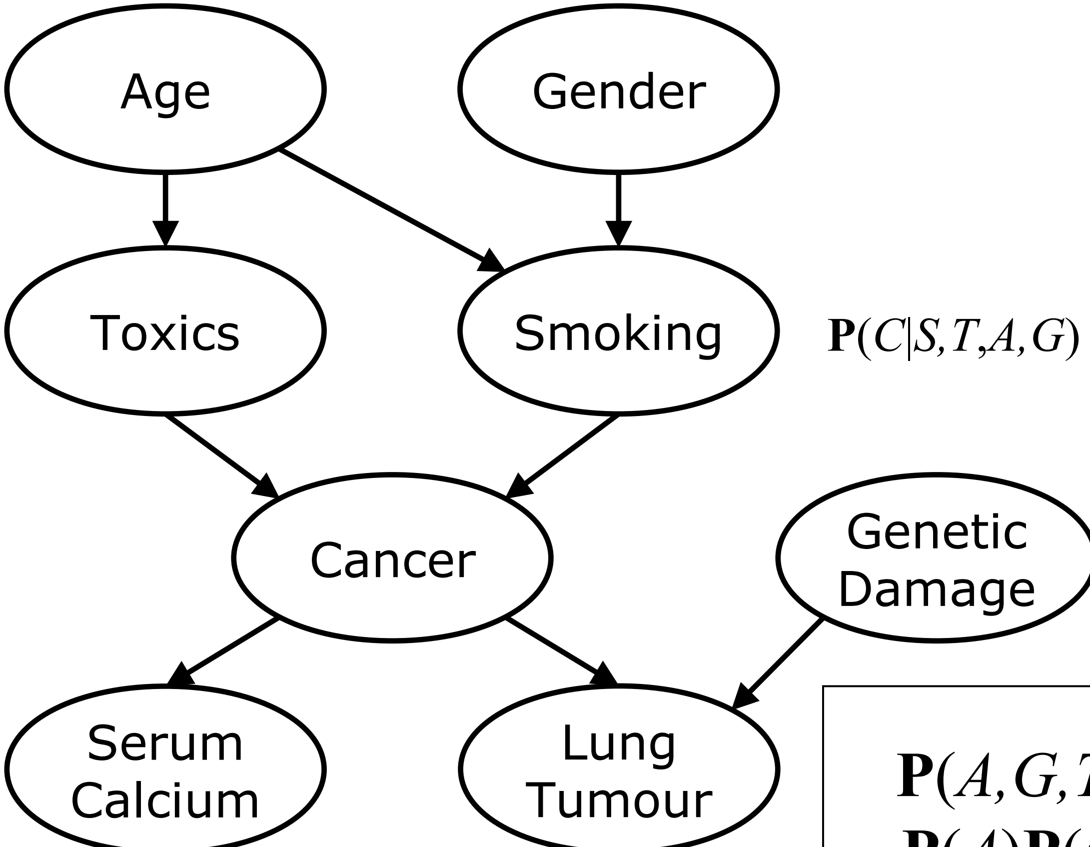
The cancer network



From Breese and Coller 1997

The cancer network

$\mathbf{P}(A,G) = \mathbf{P}(A)\mathbf{P}(G)$



$\mathbf{P}(C|S,T,A,G) = \mathbf{P}(C|S,T)$

$\mathbf{P}(SC,C,LT,GD) = \mathbf{P}(SC|C)\mathbf{P}(LT|C,GD)\mathbf{P}(C)\ \mathbf{P}(GD)$

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

From Breese and Coller 1997

Meaning of Bayesian network

The general chain rule (always true):

$$P(x_1, x_2, \ldots, x_n) = P(x_1 \mid x_2, x_3, \ldots, x_n) P(x_2, x_3, \ldots, x_n) =$$

$$P(x_1 \mid x_2, x_3, \ldots, x_n) P(x_2 \mid x_3, x_4, \ldots, x_n) P(x_3, x_4, \ldots, x_n) = \cdots$$

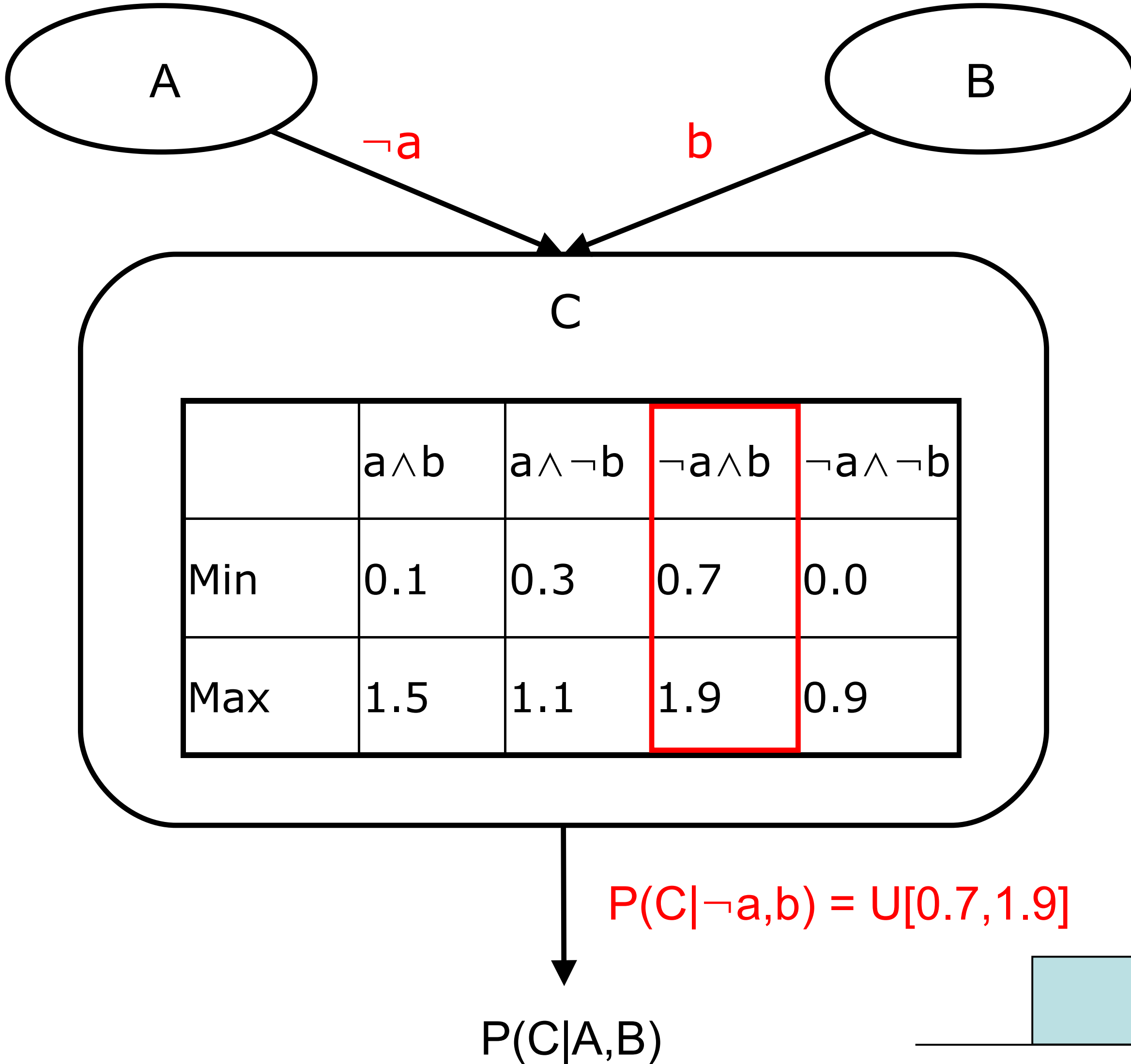
$$= \prod_{i=1}^{n} P(x_i \mid x_{i+1}, \ldots, x_n)$$

The Bayesian network chain rule:

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

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

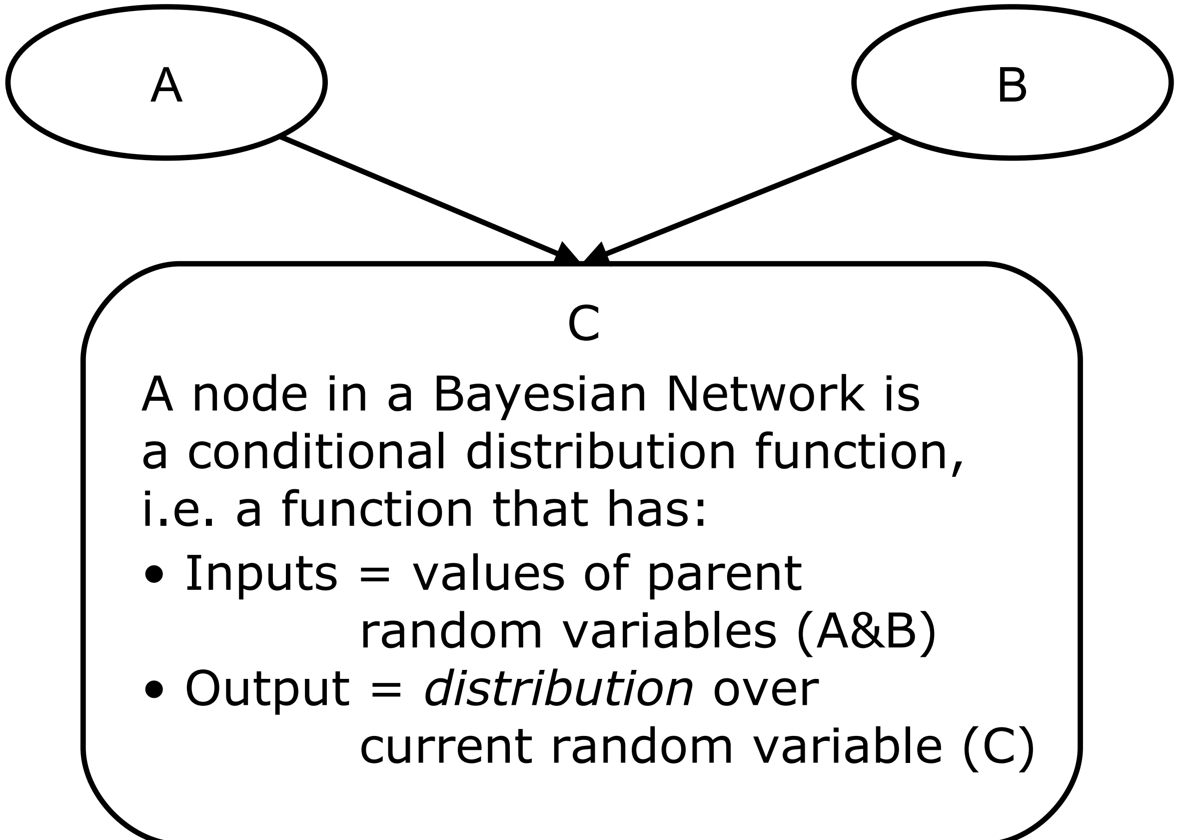
Bayes network node is a function

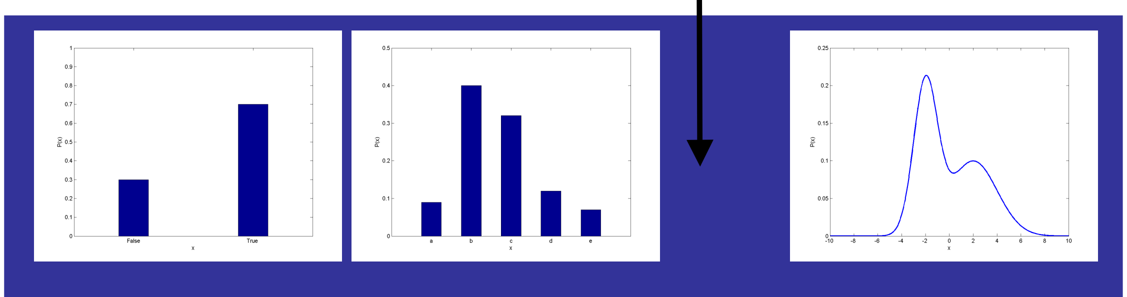


	a∧b	a∧¬b	¬a∧b	¬a∧¬b
Min	0.1	0.3	0.7	0.0
Max	1.5	1.1	1.9	0.9

P(C|¬a,b) = U[0.7,1.9]

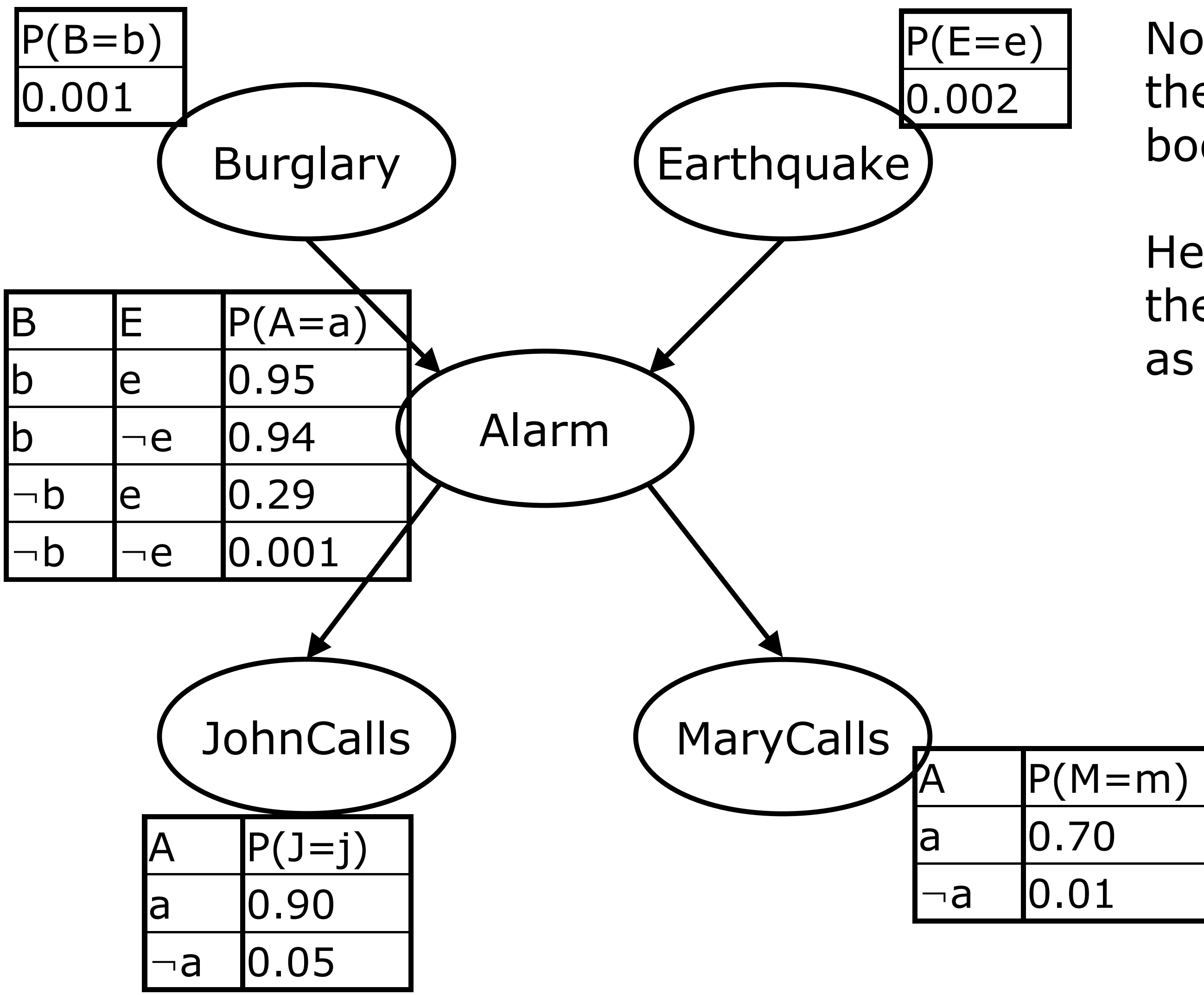
Bayes network node is a function





Can be any type of function from values to distributions.

Example: The alarm network



P(B=b)
0.001

P(E=e)
0.002

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

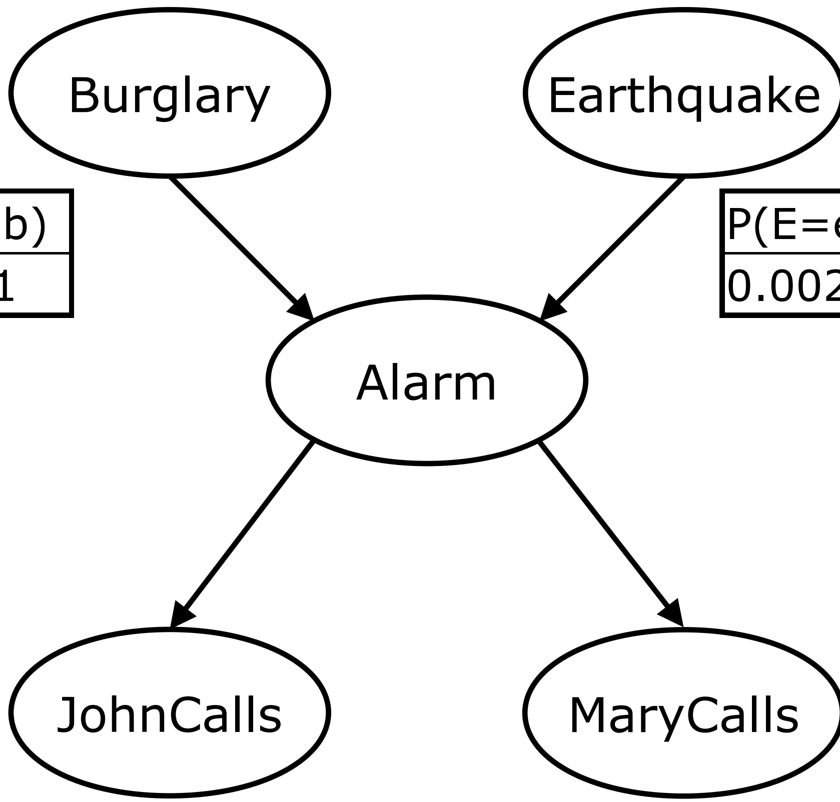
Note: Each number in the tables represents a boolean distribution.

Hence, for every input, there is a distribution as an output.

Example: The alarm network

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$$
$$= P(\neg b)P(\neg e)P(a \mid \neg b, \neg e)P(m \mid a)P(j \mid a)$$
$$= 0.999 \cdot 0.998 \cdot 0.001 \cdot 0.70 \cdot 0.90 = 0.00063$$

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



P(B=b)
0.001

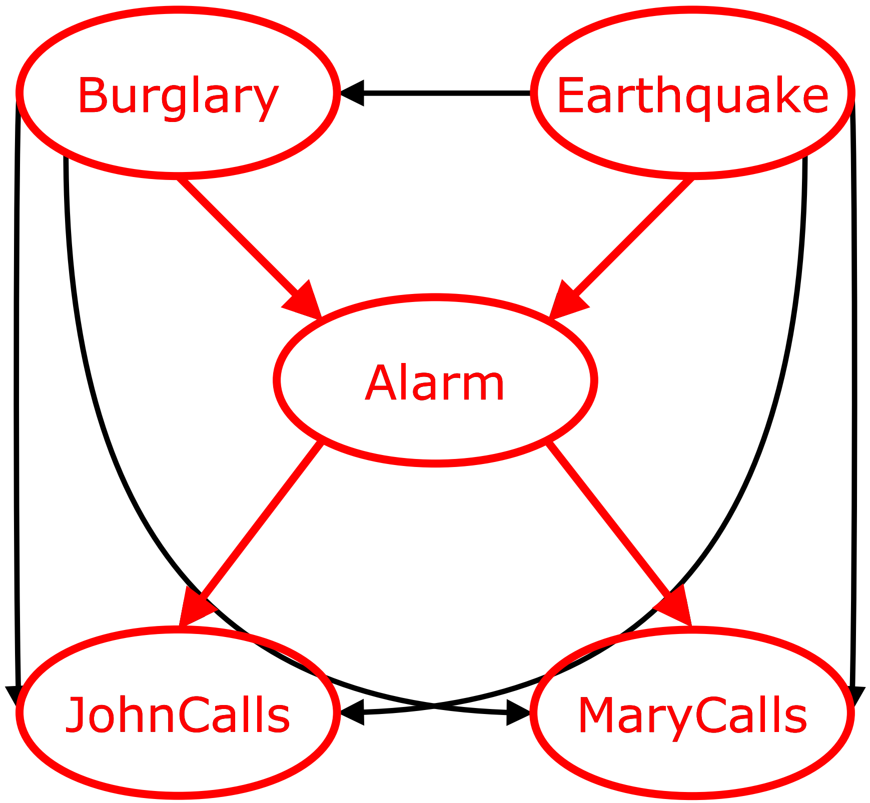
P(E=e)
0.002

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬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.

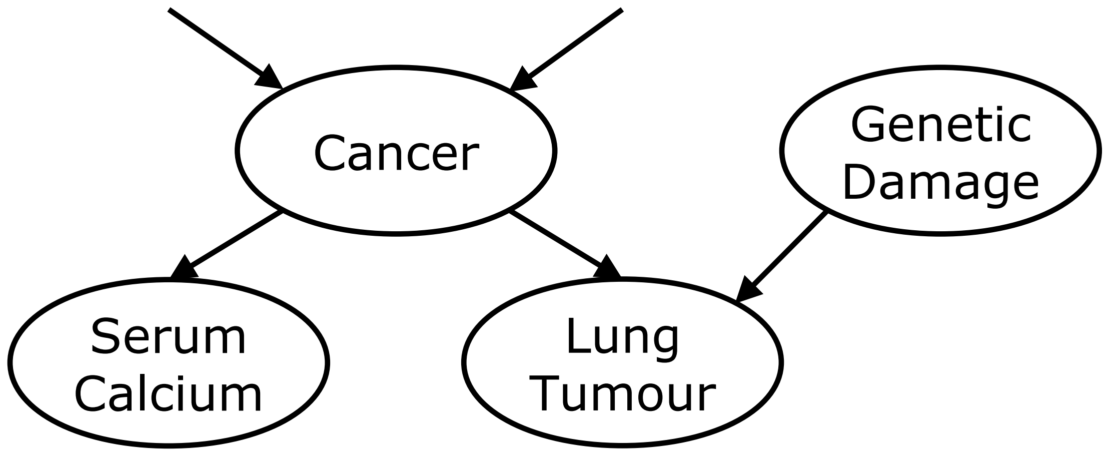
The correctness of the Bayesian network, of course, depends on the validity of these assumptions!

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= *true* or *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*.

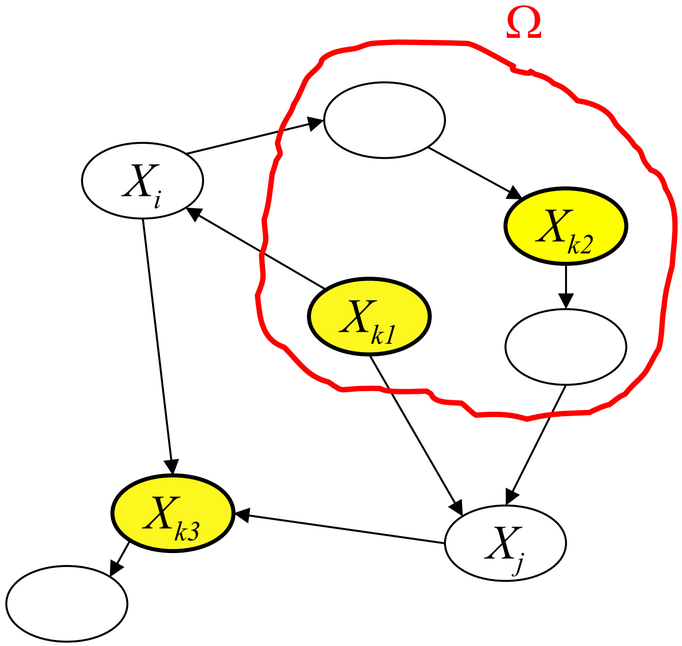
Path blocking & d-separation

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

Two nodes X_i and X_j are conditionally independent given a set $\Omega = \{X_1, X_2, X_3, \ldots\}$ 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 Ω, and both arcs on the path lead into X_k.

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



$$P(X_i, X_j \mid \Omega) = P(X_i \mid \Omega) P(X_j \mid \Omega)$$

Some definitions of BN
(from Wikipedia)

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

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

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

G = (V,E) is a directed acyclic graph (DAG) of vertices (V) and edges (E)

Some definitions of BN
(from Wikipedia)

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

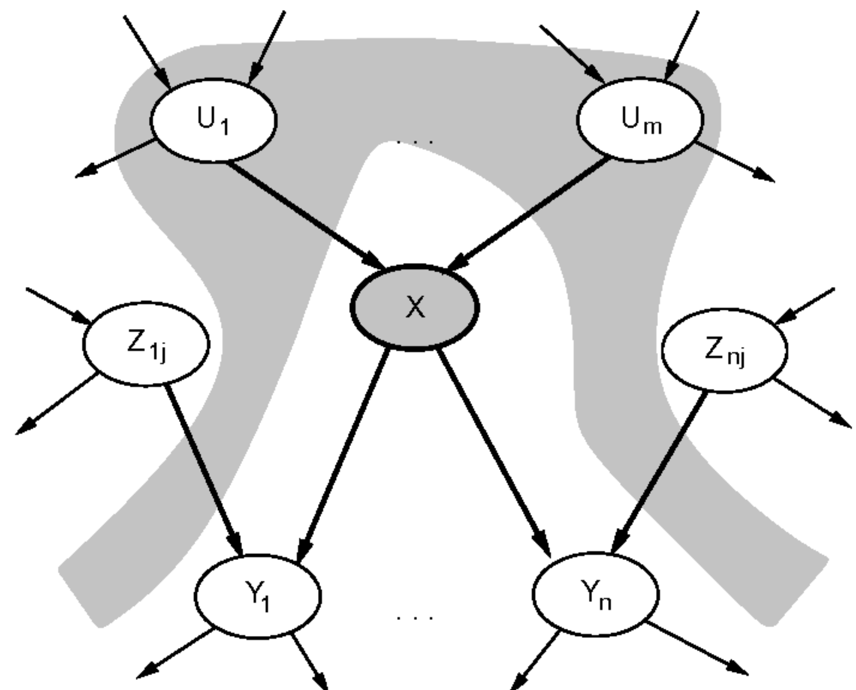
$$P(x_1 \mid non\text{-}descentants(X_i)) = P(x_i \mid parents(X_i))$$

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

X = {X_1, X_2, ..., X_N} is a set of random variables
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.



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

$$P(X \mid U_1, \ldots, U_m) P(Z_{1j}, \ldots, Z_{nj} \mid U_1, \ldots, U_m)$$

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

$$P(X \mid U_1, \ldots, U_m)$$

Some definitions of BN
(from Wikipedia)

3. X is a Bayesian network with respect to **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.

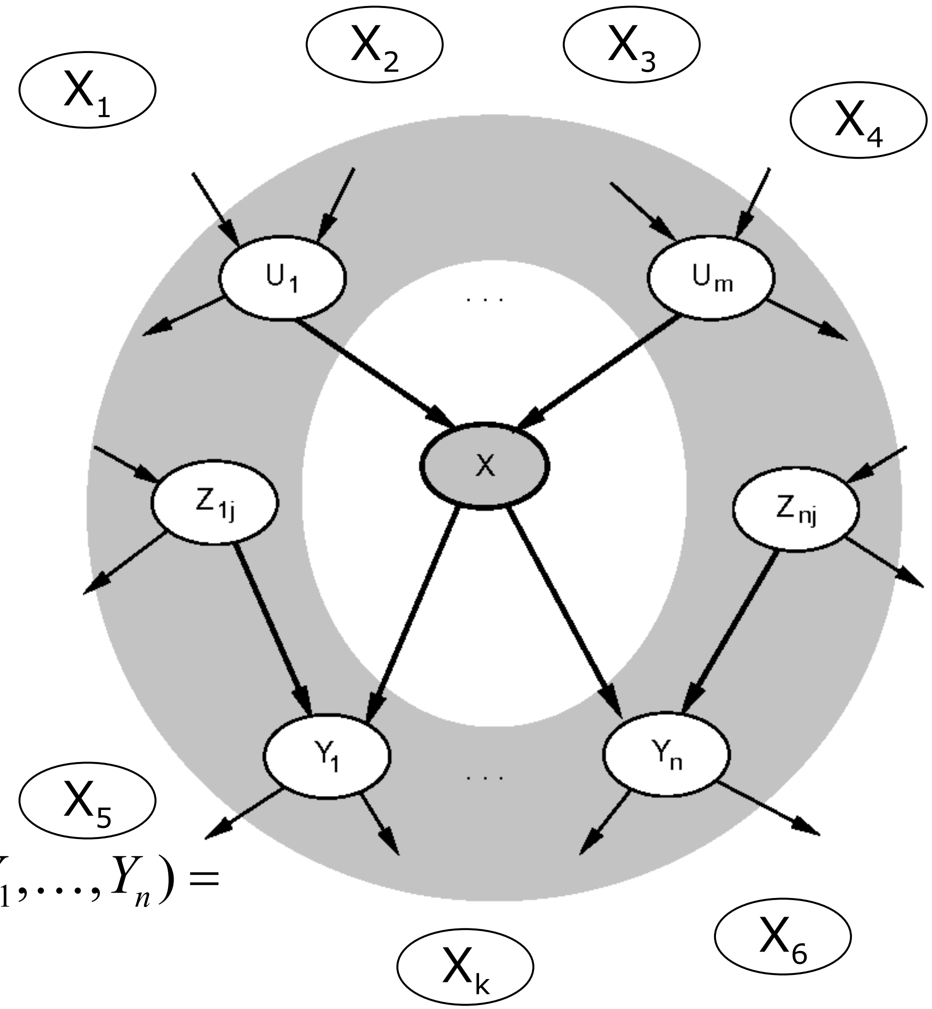
$$P(x_i \mid \mathrm{all\,nodes}) = P(x_i \mid Markov\,blanket(X_i))$$

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



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

Some definitions of BN
(from Wikipedia)

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

$$P(x_i, x_j \mid X_1, X_2, \ldots, X_N) =$$
$$P(x_i \mid d\text{-}separating\ set(i, j))P(x_j \mid d\text{-}separating\ set(i, j))$$

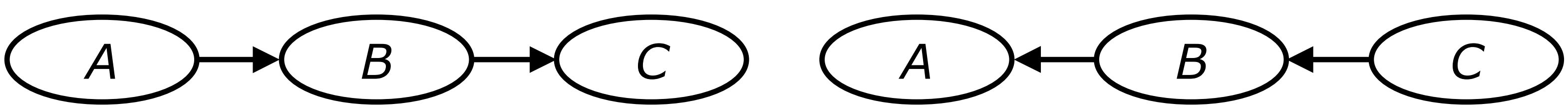
The d-separating set(i,j) is the set of nodes that d-separate node i and j.

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

X = $\{X_1, X_2, \ldots, X_N\}$ is a set of random variables
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 <u>necessary</u>: 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:



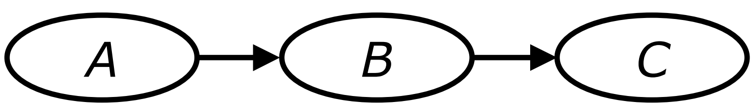
are equivalent. They impose the same conditional independence requirements.

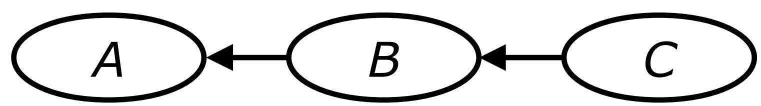
A <u>causal network</u> is a Bayesian network with an explicit requirement that the relationships be causal.

Causal networks

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

$$= P(A\,|\,B) \,*.....? \quad *...$$





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

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 preceeding information?

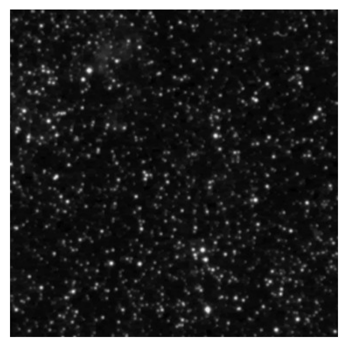
*In the 2:nd edition is this exercise 14.3 and the figure is 14.19.

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 preceeding information?

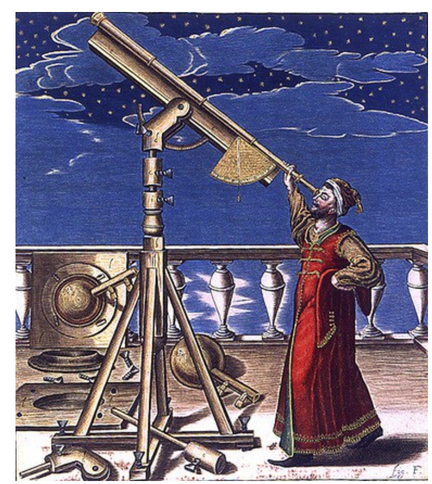
F_1

M_1



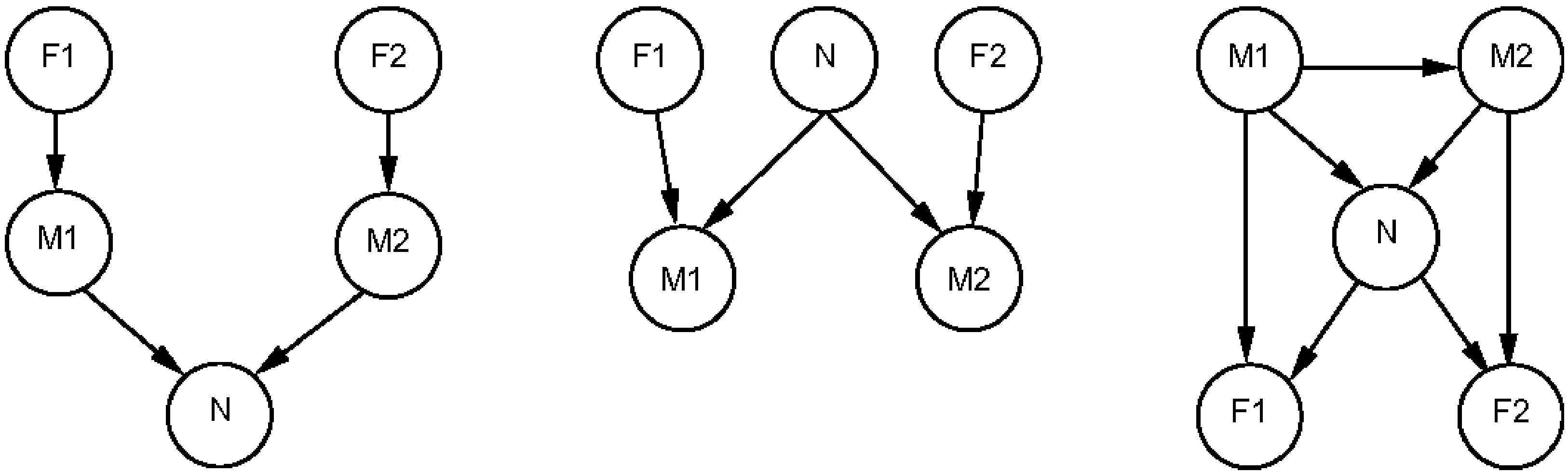
N

F_2



M_2

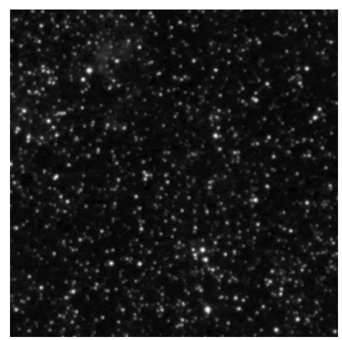
Exercise 14.12 (a) in AIMA



(i) (ii) (iii)

F_1

M_1



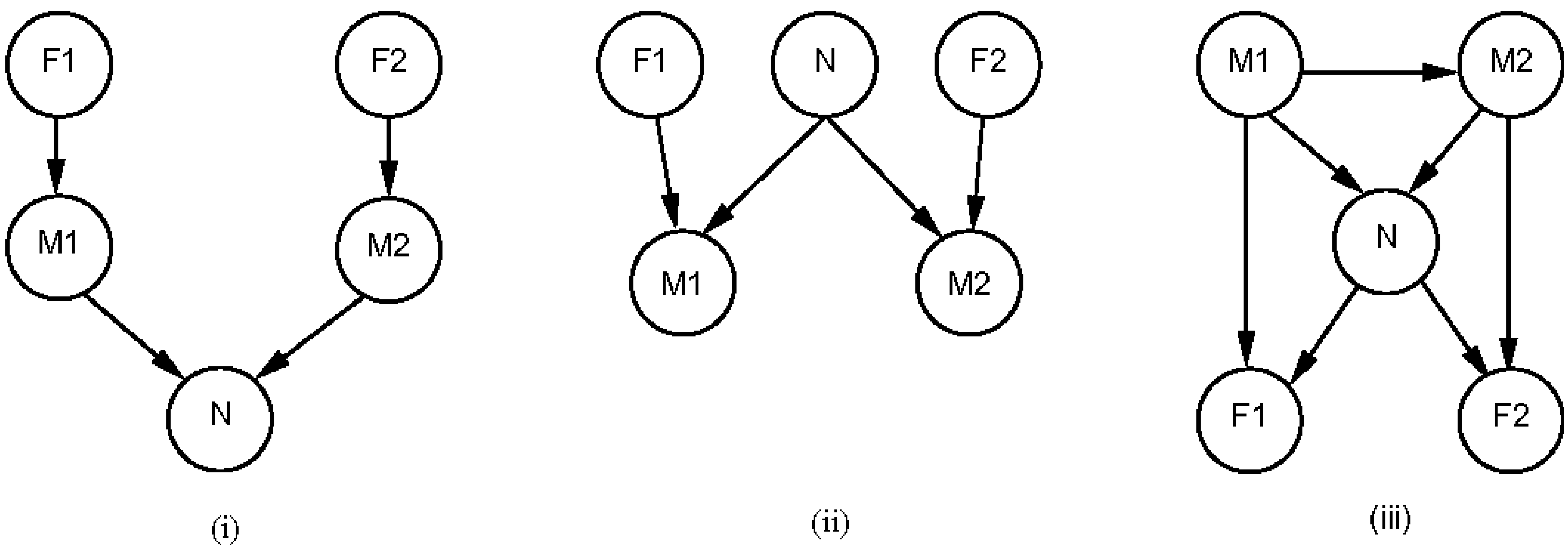
N

F_2

M_2

Exercise 14.12 (a) in AIMA

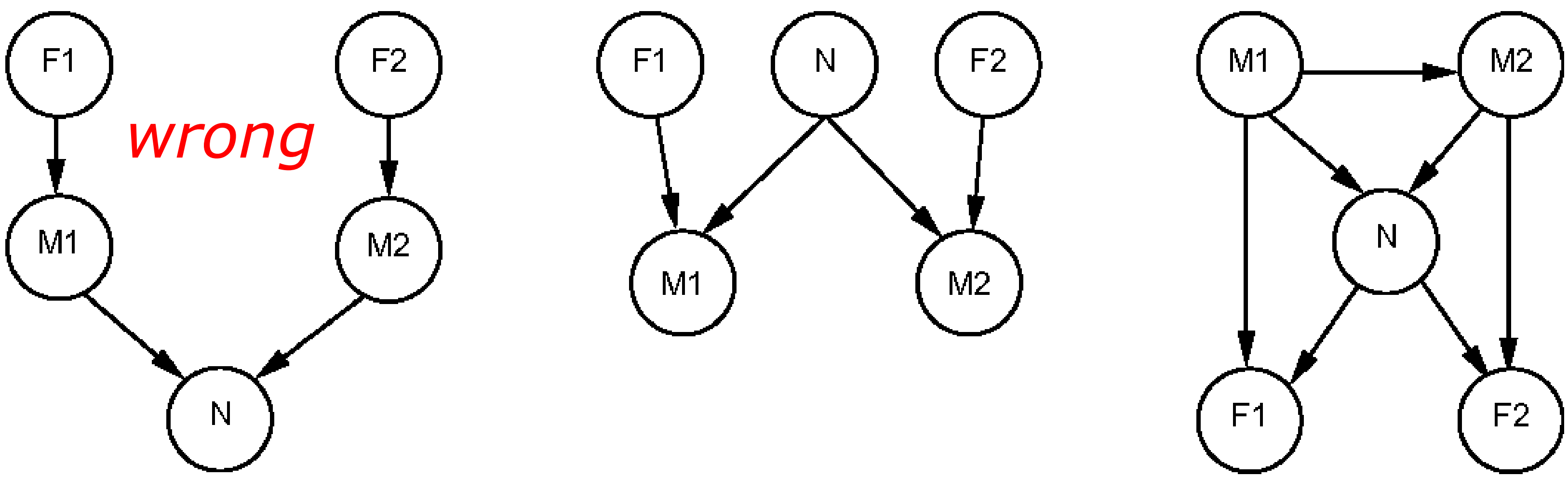
- Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky, using their telescopes. Normally there is a small possibility e of error up to one star in each direction. Each telescope can also (with a much smaller probability 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 preceeding information?



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:*

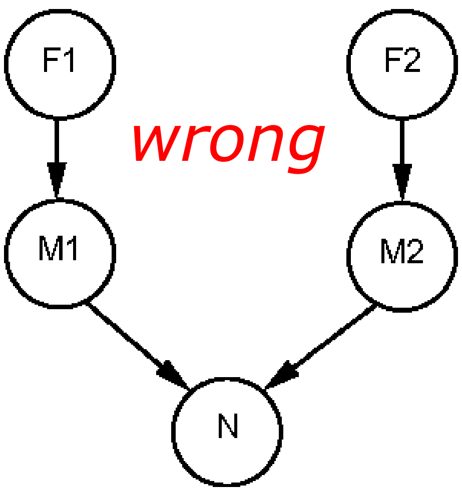
This cannot be correct!



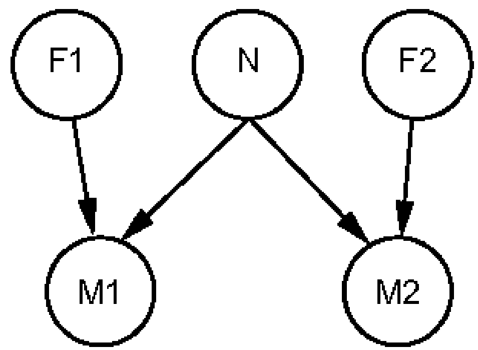
(i) (ii) (iii)

Exercise 14.12 (a) in AIMA

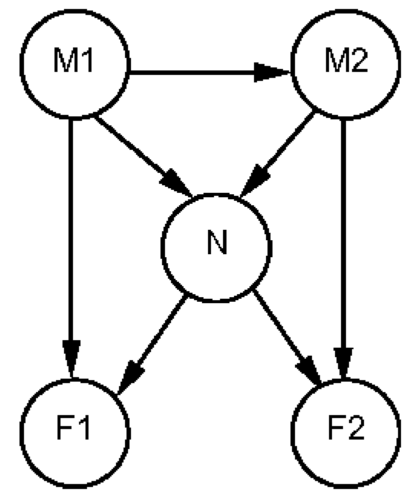
- Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky, using their telescopes. Normally there is a small possibility e of error up to one star in each direction. Each telescope can also (with a much smaller probability 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 preceeding information?



(i)



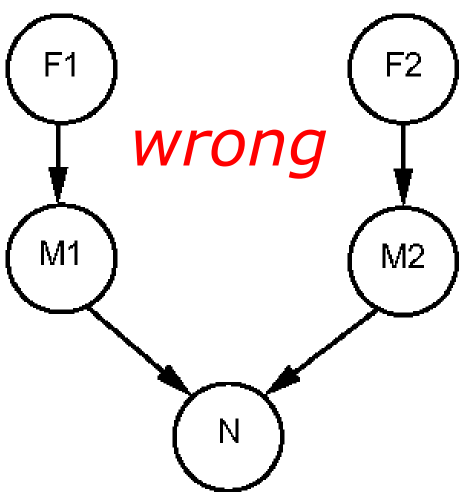
(ii)



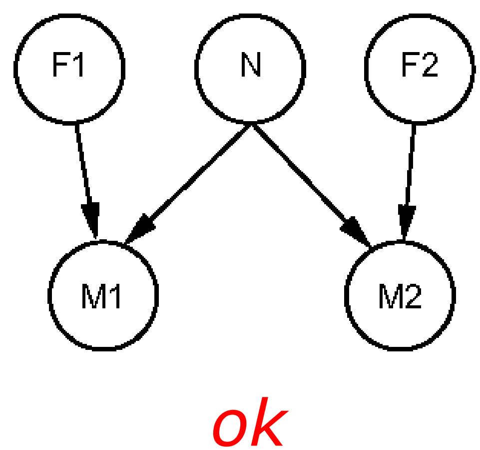
(iii)

Exercise 14.12 (a) in AIMA

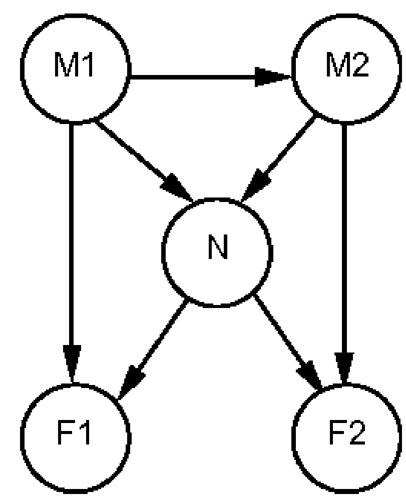
- (ii) is correct – it describes the causal relationships. It is a causal network.



(i)



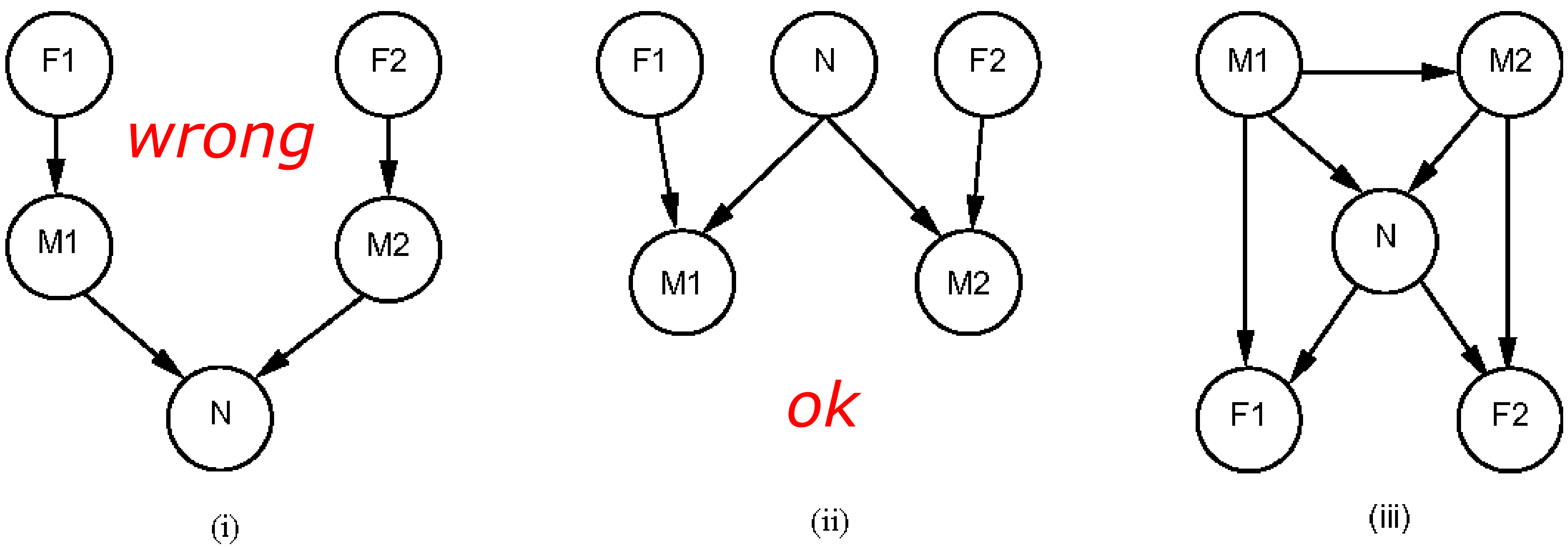
(ii)



(iii)

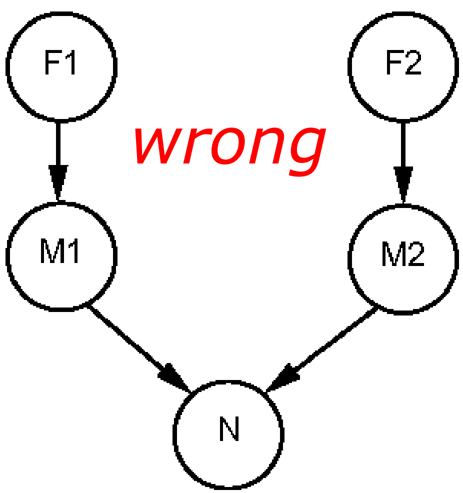
Exercise 14.12 (a) in AIMA

- 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 preceeding information?

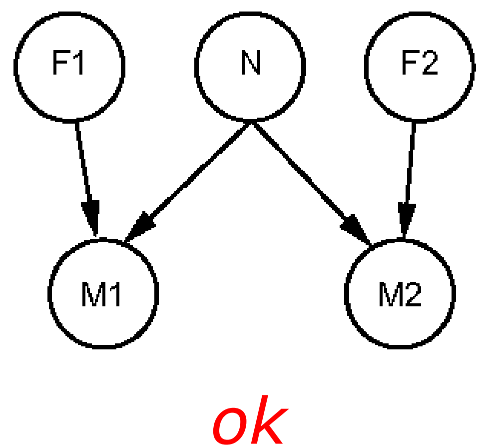


Exercise 14.12 (a) in AIMA

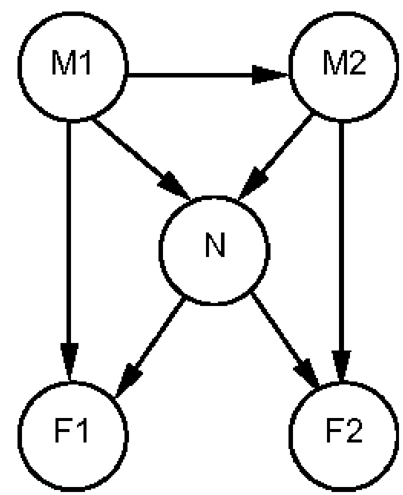
- (iii) is also ok – a fully connected graph would be correct (but not efficient). (iii) has all connections except M_i–F_j and F_i–F_j. (iii) is not causal and not efficient.



(i)



(ii)



(iii)

ok – but not good

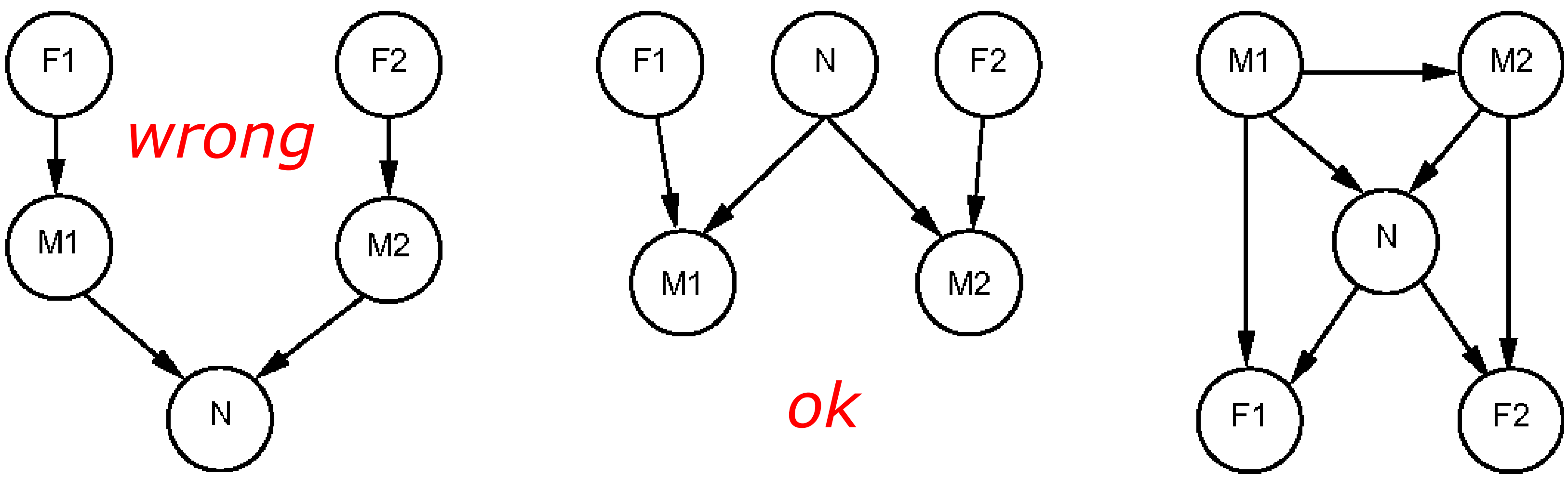
Exercise 14.12 (a) in AIMA

(ii) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(N)P(M1 \mid F1, N)P(M2 \mid F2, N)$$

(iii) says:

$$P(F1, F2, N, M1, M2) = P(M1)P(F1 \mid N, M1)P(M2 \mid M1)P(N \mid M1, M2)P(F2 \mid N, M2)$$



(i) (ii) (iii)

ok – but not good

Exercise 14.12 (a) in AIMA

(ii) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(N)P(M1 \mid F1, N)P(M2 \mid F2, N)$$

The full correct expression (one version) is:

$$P(F1, F2, N, M1, M2) = P(M1 \mid F1, F2, N, M2)P(F1, F2, N, M2) =$$

$$P(M1 \mid F1, F2, N, M2)P(M2 \mid F1, F2, N)P(F1, F2, N) =$$

$$P(M1 \mid F1, F2, N, M2)P(M2 \mid F1, F2, N)P(F1 \mid F2, N)P(F2, N) =$$

$$\underbrace{P(M1 \mid F1, F2, N, M2)}_{\approx P(M1 \mid F1, N)}\underbrace{P(M2 \mid F1, F2, N)}_{\approx P(M2 \mid F2, N)}\underbrace{P(F1 \mid F2, N)}_{\approx P(F1)}\underbrace{P(F2 \mid N)}_{\approx P(F2)}P(N)$$

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:

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

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

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

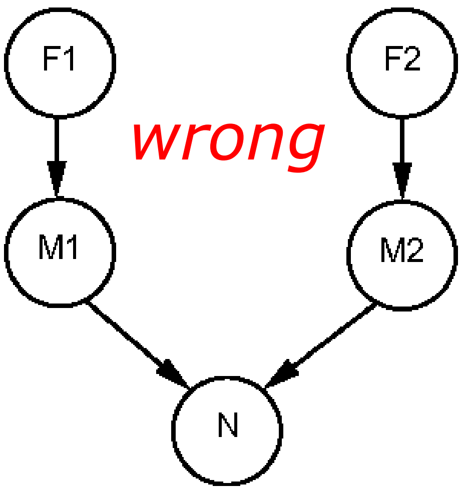
$$\underbrace{P(F1\,|\,F2,N,M1,M2)}_{\approx P(F1\,|\,N,M1)}\underbrace{P(F2\,|\,N,M1,M2)}_{\approx P(F2\,|\,N,M2)}P(N\,|\,M1,M2)P(M2\,|\,M1)P(M1)$$

ok

Exercise 14.12 (a) in AIMA

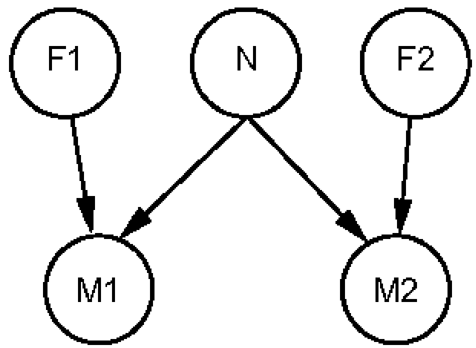
(i) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(M1 \mid F1)P(M2 \mid F2)P(N \mid M1, M2)$$



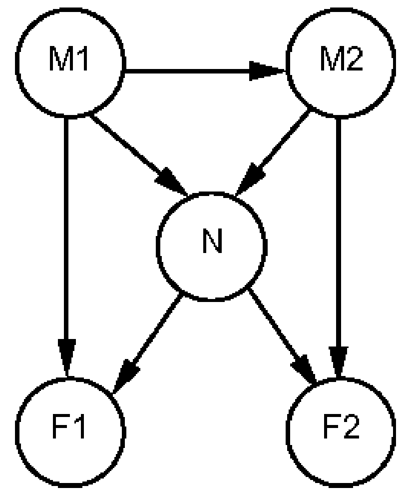
wrong

(i)



ok

(ii)



(iii)

ok – but not good

Exercise 14.12 (a) in AIMA

(i) says:

$$P(F1, F2, N, M1, M2) = P(F1)P(F2)P(M1 \mid F1)P(M2 \mid F2)P(N \mid M1, M2)$$

The full correct expression (a third version) is:

$$P(F1, F2, N, M1, M2) = P(N \mid M1, M2, F1, F2)P(M1, M2, F1, F2) =$$

$$P(N \mid M1, M2, F1, F2)P(M1 \mid M2, F1, F2)P(M2, F1, F2) =$$

$$P(N \mid M1, M2, F1, F2)P(M1 \mid M2, F1, F2)P(M2 \mid F1, F2)P(F1, F2) =$$

$$\underbrace{P(N \mid M1, M2, F1, F2)}_{\approx P(N \mid M1, M2)}\underbrace{P(M1 \mid M2, F1, F2)}_{\approx P(M1 \mid F1)}\underbrace{P(M2 \mid F1, F2)}_{\approx P(M2 \mid F2)}\underbrace{P(F1 \mid F2)}_{\approx P(F1)}P(F2)$$

Exercise 14.12 (a) in AIMA

(i) says:

$$P(F1,F2,N,M1,M2) = P(F1)P(F2)P(M1 \mid F1)P(M2 \mid F2)P(N \mid M1,M2)$$

This is an unreasonable approximation. The rest is ok.

The full correct expression (a third version) is:

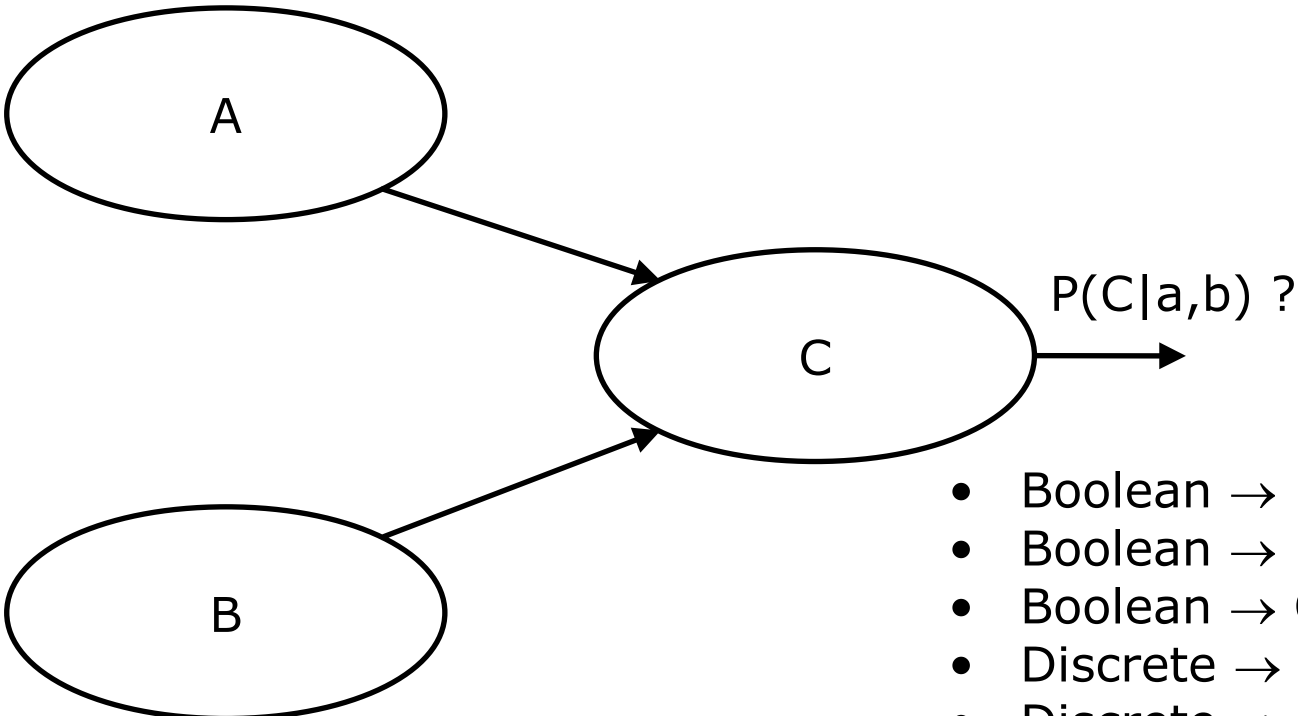
$$P(F1,F2,N,M1,M2) = P(N \mid M1,M2,F1,F2)P(M1,M2,F1,F2) =$$

$$P(N \mid M1,M2,F1,F2)P(M1 \mid M2,F1,F2)P(M2,F1,F2) =$$

$$P(N \mid M1,M2,F1,F2)P(M1 \mid M2,F1,F2)P(M2 \mid F1,F2)P(F1,F2) =$$

$$\underbrace{P(N \mid M1,M2,F1,F2)}_{\approx P(N \mid M1,M2)}\underbrace{P(M1 \mid M2,F1,F2)}_{\approx P(M1 \mid F1)}\underbrace{P(M2 \mid F1,F2)}_{\approx P(M2 \mid F2)}\underbrace{P(F1 \mid F2)}_{\approx P(F1)}P(F2)$$

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 $\to$ Boolean:
  Noisy-OR, Noisy-AND

Boolean/Discrete $\to$ Discrete:
  Noisy-MAX

Bool./Discr./Cont. $\to$ Continuous:
  Parametric distribution (e.g. Gaussian)

Continuous $\to$ Boolean:
  Logit/Probit

Noisy-OR example

Boolean → Boolean

$P(E\|C_1,C_2,C_3)$								
C_1	0	1	0	0	1	1	0	1
C_2	0	0	1	0	1	0	1	1
C_3	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.

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

Example from L.E. Sucar

Noisy-OR general case

Boolean → Boolean

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

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

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

Needs only n parameters, not 2^n parameters.

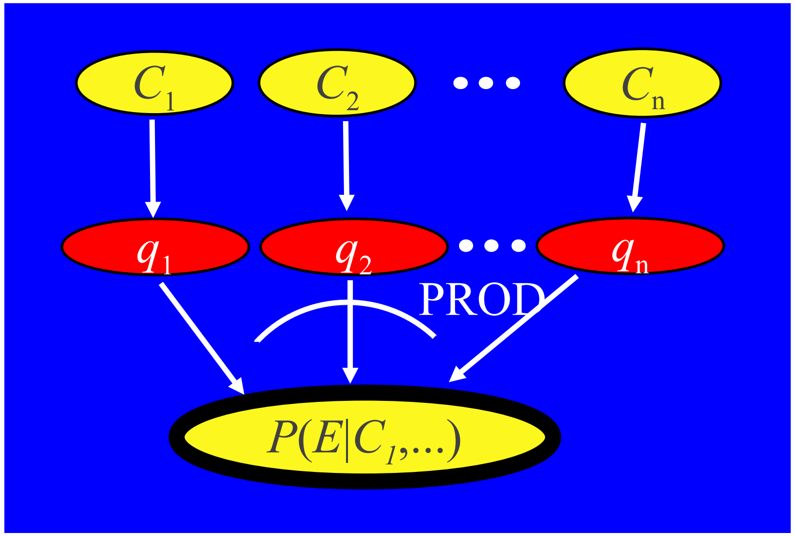
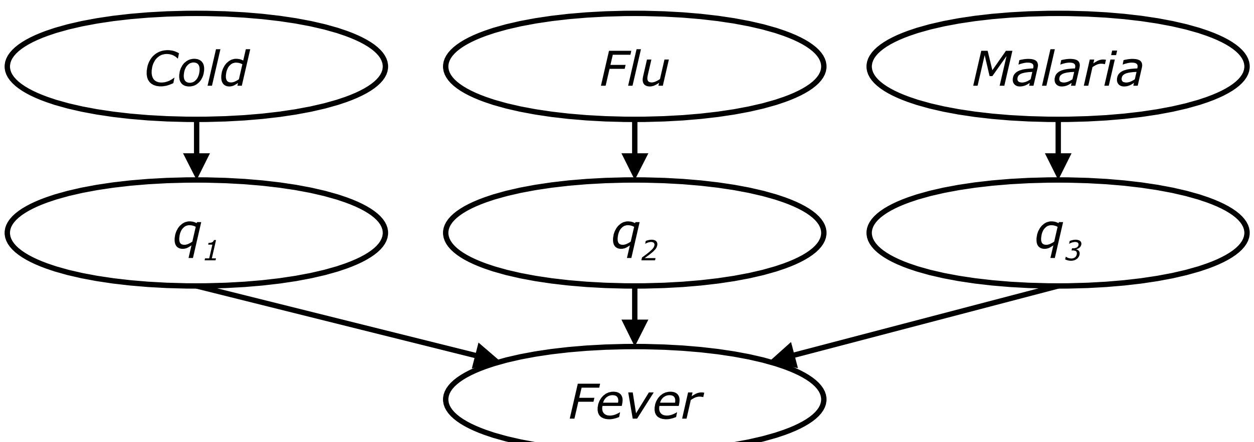


Image adapted from Laskey & Mahoney 1999

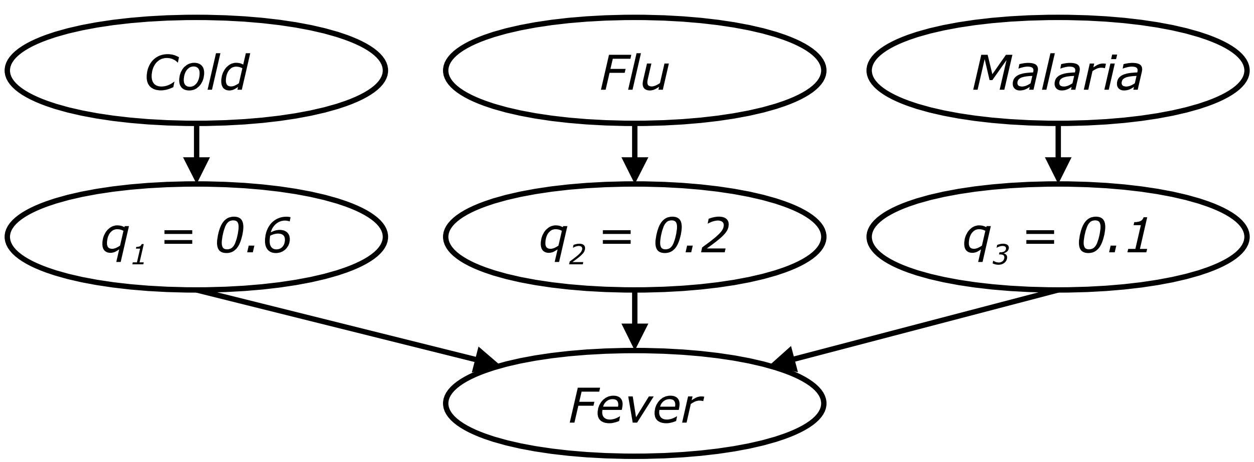
Noisy-OR example (II)

- *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)

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

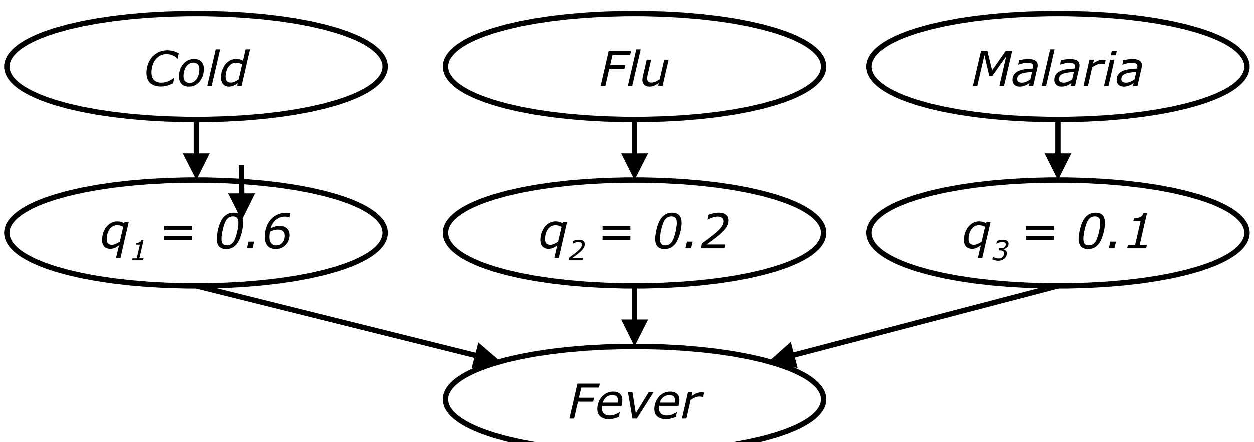


Noisy-OR example (II)

- P(*Fever* | *Cold*) = 0.4 ⇒ q_1 = 0.6
- P(*Fever* | *Flu*) = 0.8 ⇒ q_2 = 0.2
- P(*Fever* | *Malaria*) = 0.9 ⇒ q_3 = 0.1

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

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

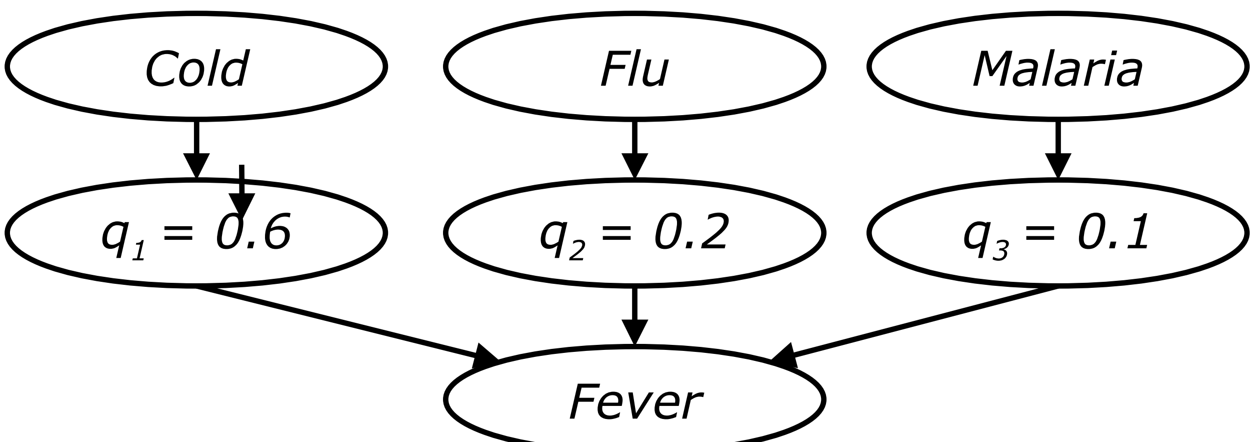


Noisy-OR example (II)

- P(*Fever* | *Cold*) = 0.4 ⇒ q_1 = 0.6
- P(*Fever* | *Flu*) = 0.8 ⇒ q_2 = 0.2
- P(*Fever* | *Malaria*) = 0.9 ⇒ q_3 = 0.1

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

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

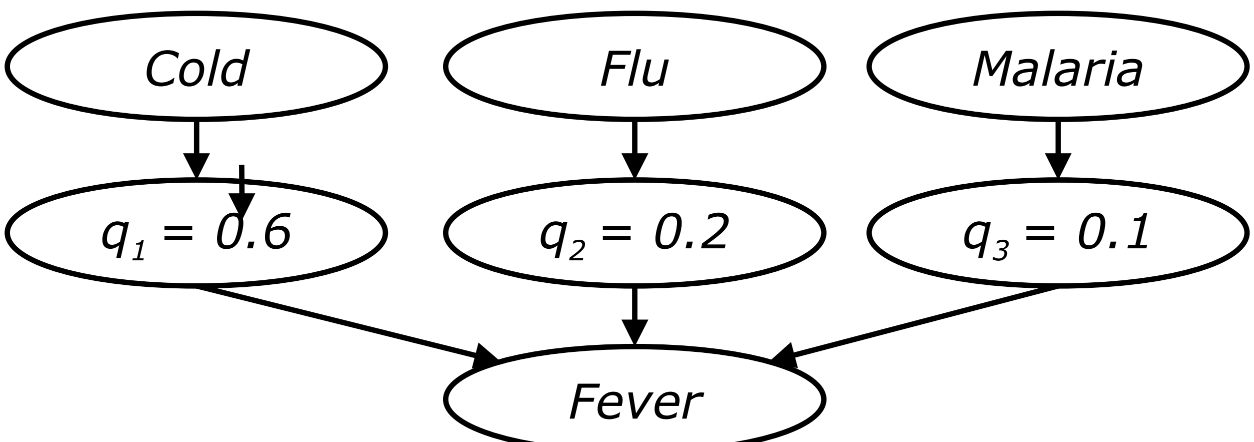


Noisy-OR example (II)

- P(*Fever* | *Cold*) = 0.4 ⇒ q_1 = 0.6
- P(*Fever* | *Flu*) = 0.8 ⇒ q_2 = 0.2
- P(*Fever* | *Malaria*) = 0.9 ⇒ q_3 = 0.1

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

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



Parametric probability densities

Boolean/Discr./Continuous → Continuous

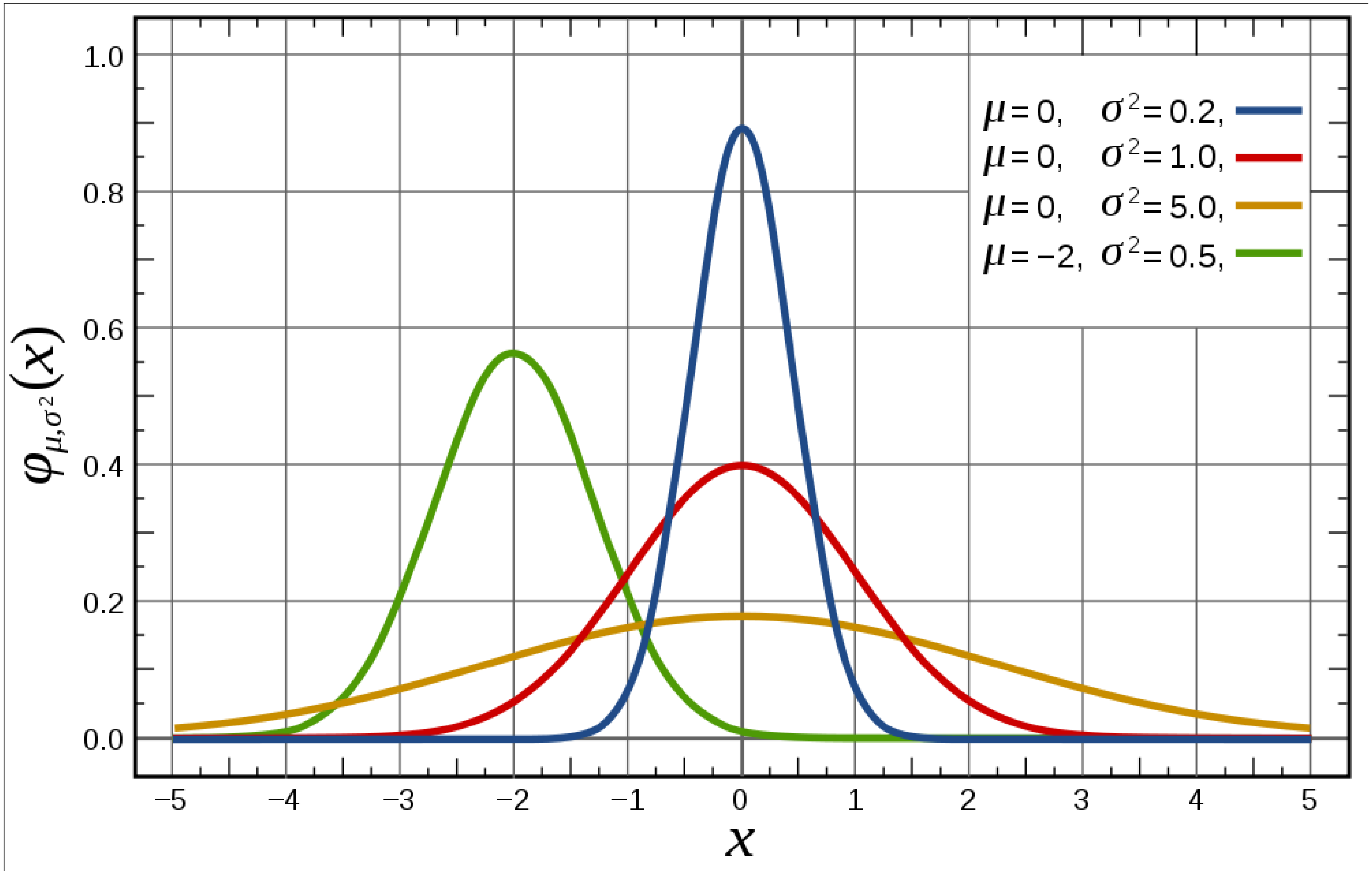
Use parametric probability densities, e.g., the normal distribution

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

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



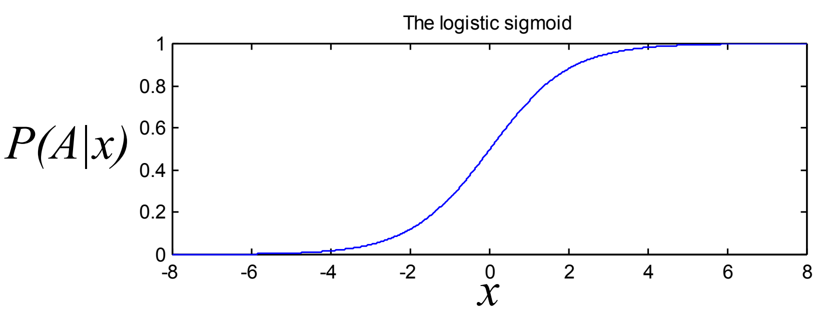
Probit & Logit

Discrete → Boolean

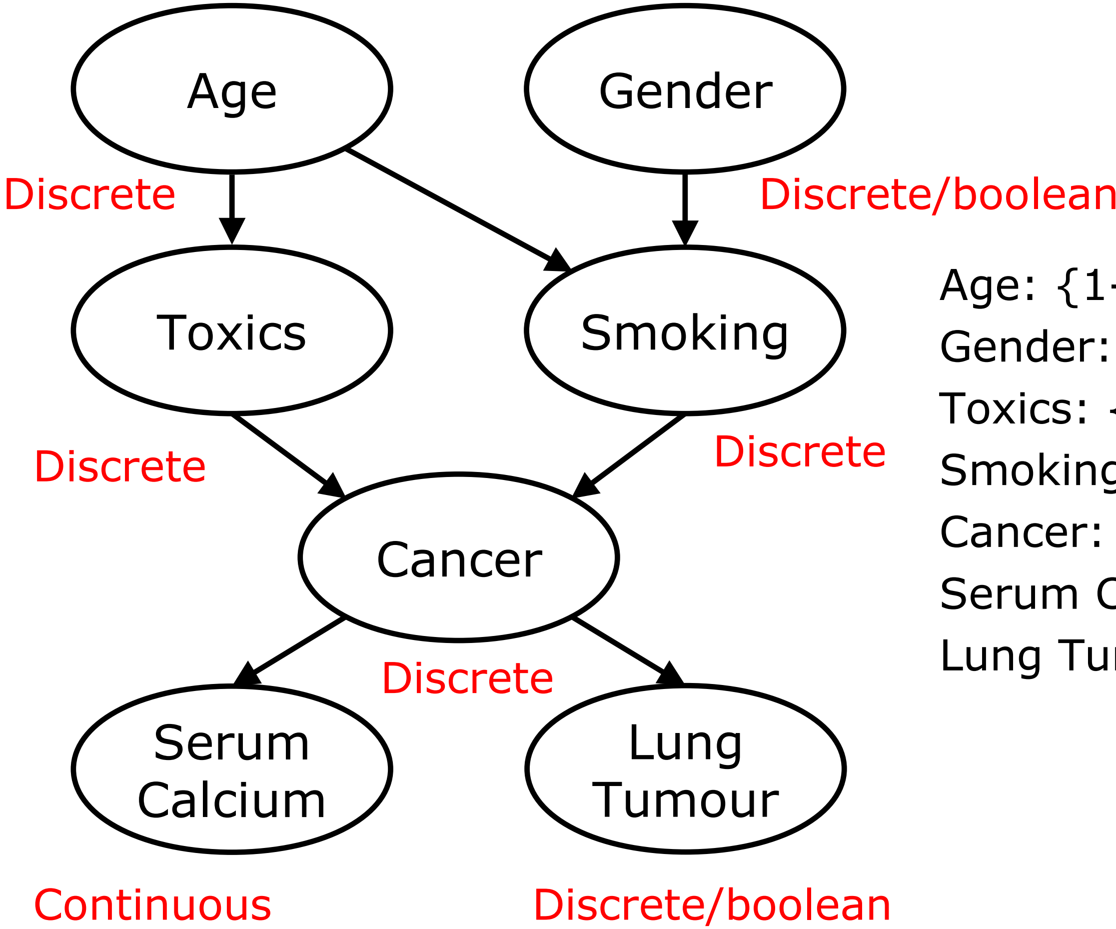
If the input is continuous but output is boolean, use probit or logit

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

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



The cancer network



Age: {1-10, 11-20,...}
Gender: {M, F}
Toxics: {Low, Medium, High}
Smoking: {No, Light, Heavy}
Cancer: {No, Benign, Malignant}
Serum Calcium: Level
Lung Tumour: {Yes, No}

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

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

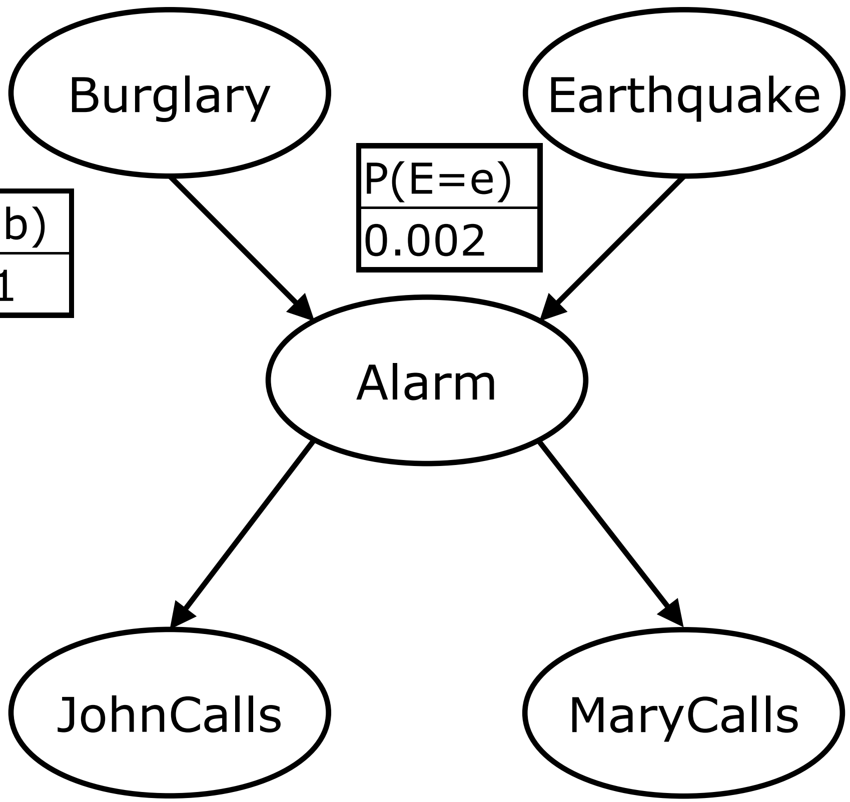
”Doable” for boolean variables: Look up entries in conditional probability tables (CPTs).

Example: The alarm network

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

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

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



P(B=b)
0.001

P(E=e)
0.002

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

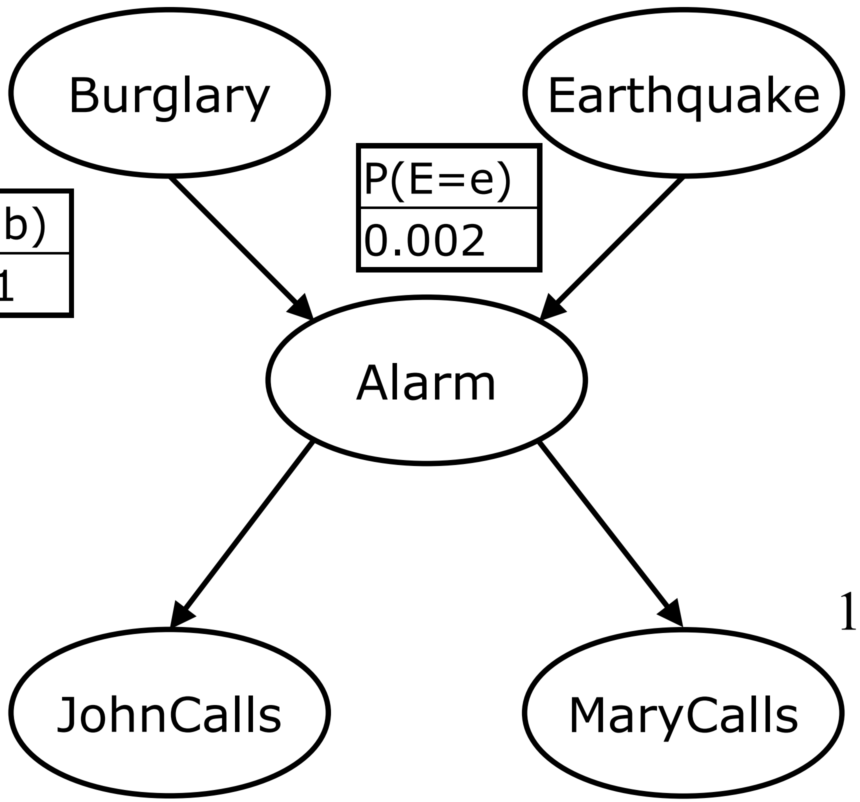
Example: The alarm network

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

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

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

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



P(B=b)
0.001

P(E=e)
0.002

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

Example: The alarm network

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

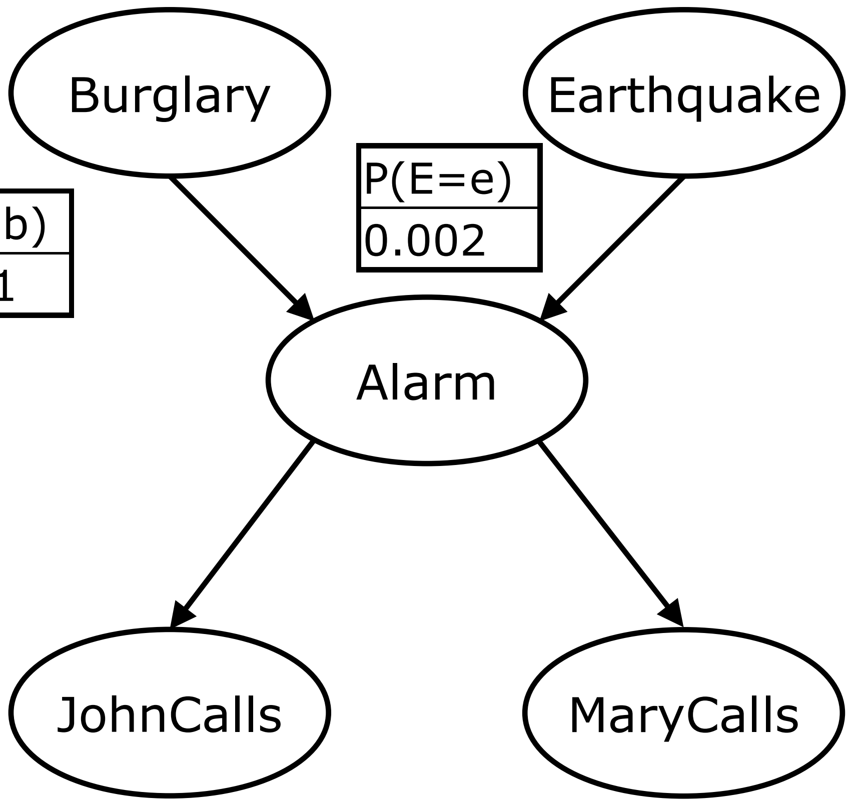
$$\mathbf{P}(B \mid j,m) = \alpha \sum_{E=\{e,\neg e\}} \sum_{A=\{a,\neg a\}} \mathbf{P}(B,E,A,j,m)$$

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

$$10^{-3} \sum_{\substack{A=\{a,\neg a\}\\ E=\{e,\neg e\}}} P(j \mid A)P(m \mid A)P(A \mid b,E)P(E) =$$

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

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



P(B=b)
0.001

P(E=e)
0.002

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

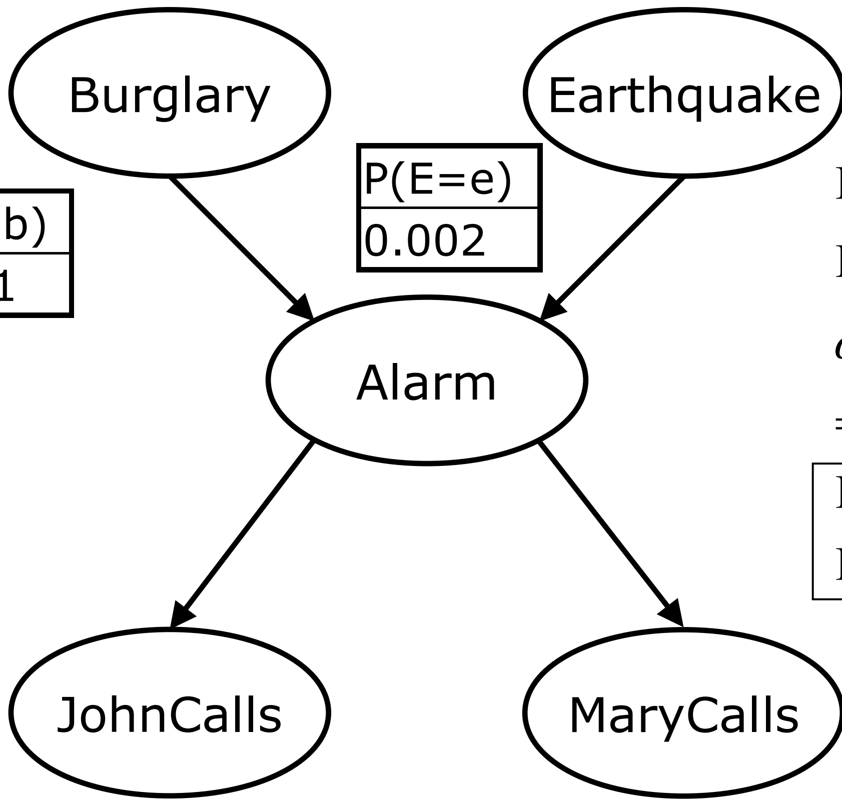
A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

Example: The alarm network

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

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



P(B=b)
0.001

P(E=e)
0.002

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

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

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

$$\alpha = \mathbf{P}(j,m)^{-1} = \left[\mathbf{P}(b,j,m) + \mathbf{P}(\neg b,j,m)\right]^{-1} =$$

$$= [2.083 \times 10^{-3}]^{-1}$$

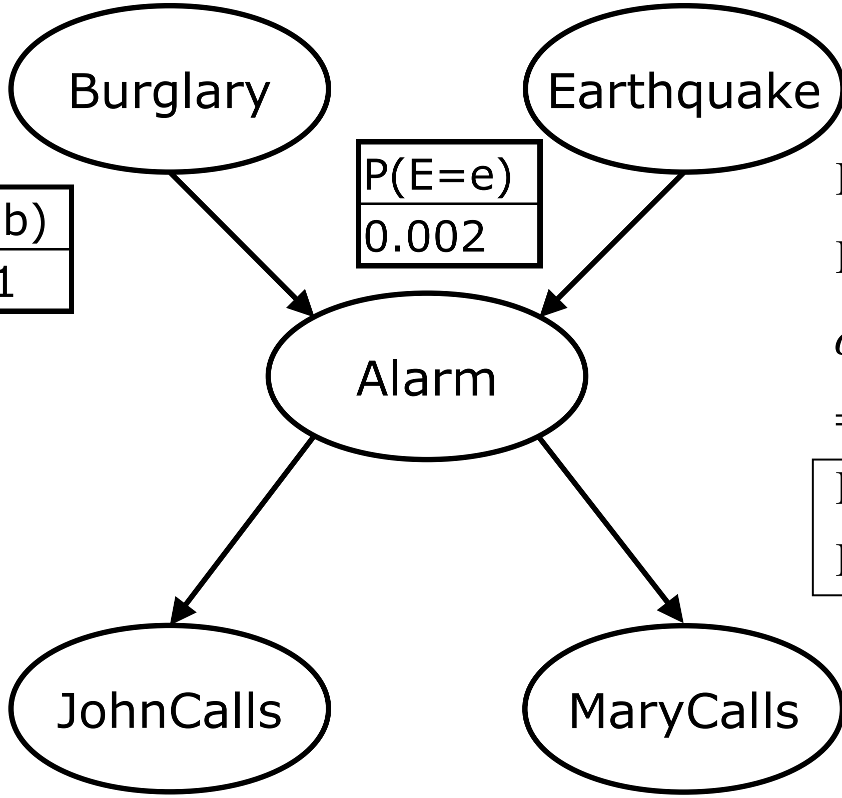
$$\boxed{\begin{aligned}\mathbf{P}(b \mid j,m) &= \alpha \mathbf{P}(b,j,m) = 0.284 \\ \mathbf{P}(\neg b \mid j,m) &= \alpha \mathbf{P}(\neg b,j,m) = 0.716\end{aligned}}$$

Example: The alarm network

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

Answer: 28%

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



P(B=b)
0.001

P(E=e)
0.002

A	P(J=j)
a	0.90
¬a	0.05

A	P(M=m)
a	0.70
¬a	0.01

B	E	P(A=a)
b	e	0.95
b	¬e	0.94
¬b	e	0.29
¬b	¬e	0.001

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

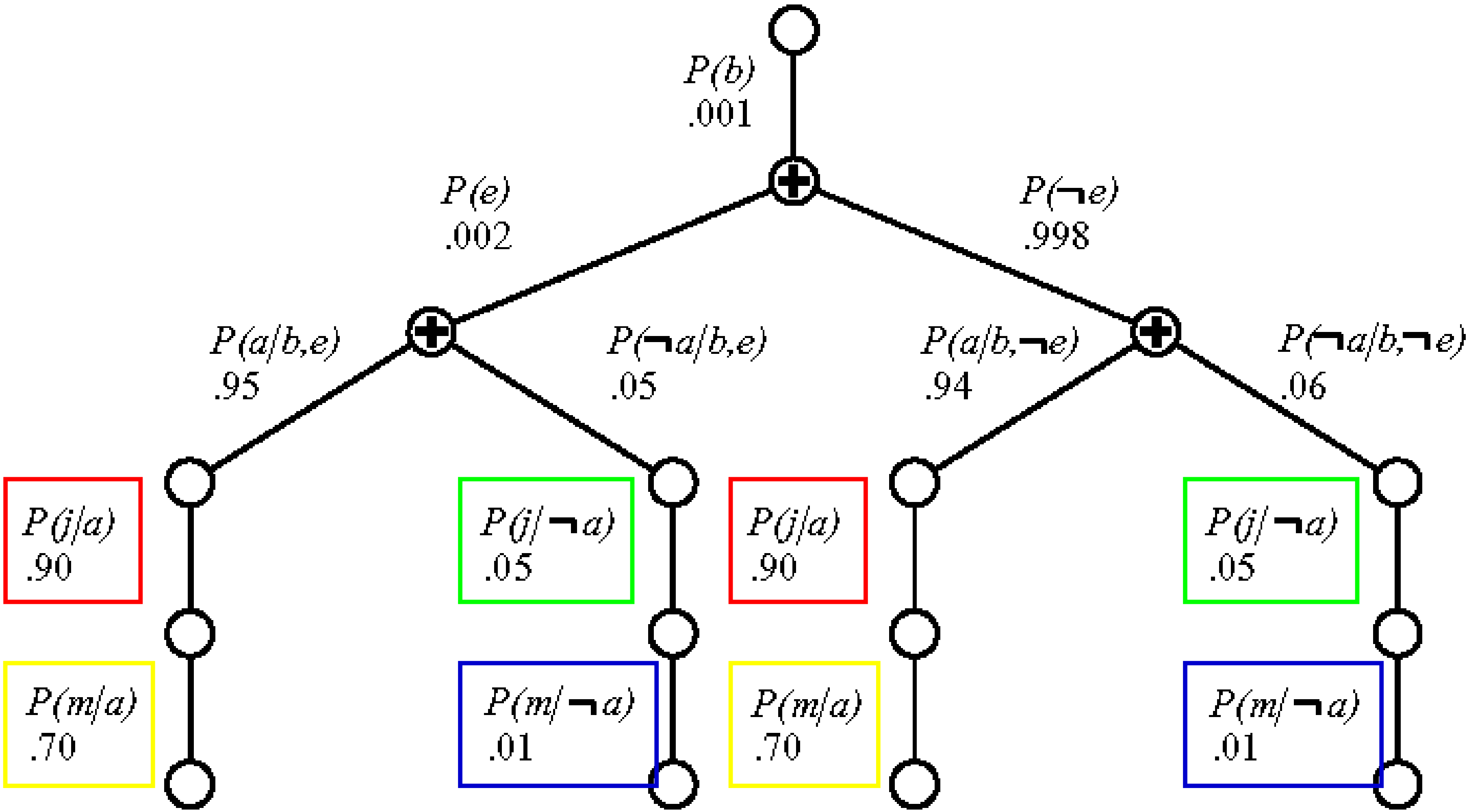
$$\mathbf{P}(\neg b,j,m) = 1.491 \times 10^{-3}$$

$$\alpha = \mathbf{P}(j,m)^{-1} = \left[\mathbf{P}(b,j,m) + \mathbf{P}(\neg b,j,m)\right]^{-1} =$$

$$= [2.083 \times 10^{-3}]^{-1}$$

$$\boxed{\begin{aligned} \mathbf{P}(b \mid j,m) &= \alpha \mathbf{P}(b,j,m) = 0.284 \\ \mathbf{P}(\neg b \mid j,m) &= \alpha \mathbf{P}(\neg b,j,m) = 0.716 \end{aligned}}$$

Use depth-first search



A lot of unneccesary 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, ...) 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.