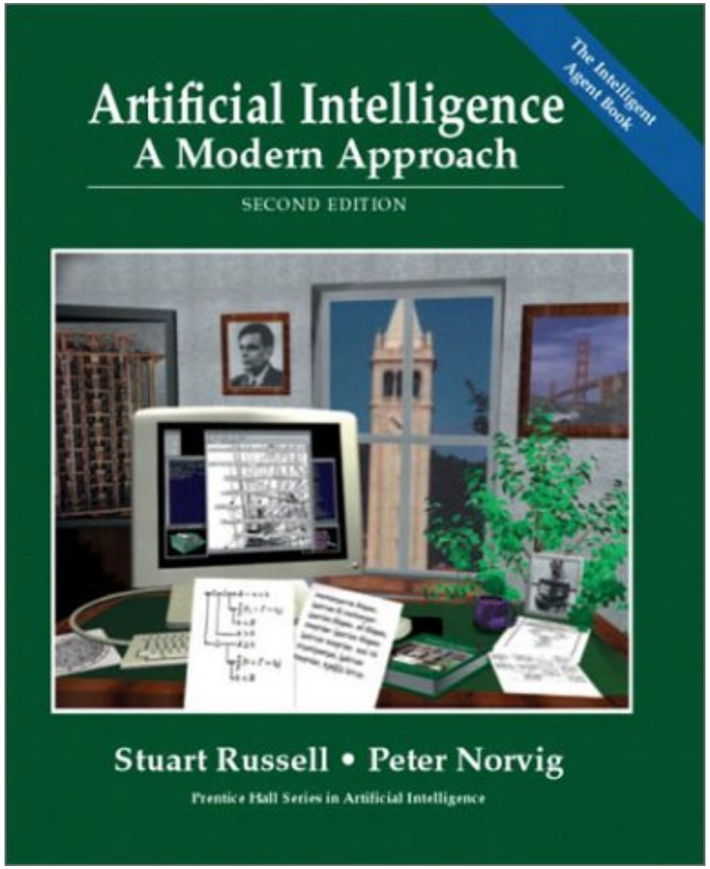


Bayesian Networks

Chapter 14

Section 1 – 2

Issues

- If a state is described by n propositions, then a belief space contains 2^n states (possibly, some have probability 0)
- → Modeling difficulty: many numbers must be entered in the first place
- → Computational issue: memory size and time

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

- Toothache and pcatch are independent given cavity (or ¬cavity), but this relation is hidden in the numbers ! [Verify this]
- Bayesian networks explicitly represent independence among propositions to reduce the number of probabilities defining a belief state

Verification

$$P(toothache, pcatch | cavity) = P(toothache | cavity) * P(pcatch | cavity)$$

$$\frac{P(toothache, pcatch, cavity)}{P(cavity)} = \frac{P(toothache, cavity)}{P(cavity)} * \frac{P(pcatch, cavity)}{P(cavity)}$$

$$P(toothache, pcatch, cavity) = P(toothache, cavity) * \frac{P(pcatch, cavity)}{P(cavity)}$$

$$0{,}108 = ((0{,}108+0{,}012) * (0{,}108+0{,}072)) / (0{,}108+0{,}012+0{,}072+0{,}008)$$

$$0{,}108 = 0{,}12 * 0{,}18 / 0{,}2$$

$$0{,}108 = 0{,}0216/0{,}2 = 216/2000 = 0{,}108$$

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

Bayes rule

- Applying the bayes rule (chain rule) does not help so much.

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\,\mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \mathbf{P}(X_1,\ldots,X_{n-2})\,\mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\,\mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \ldots$$

P(toothache, cavity, pcatch) = P(toothache)* P(cavity| toothache)
*P(pcatch | toothache, cavity)

Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

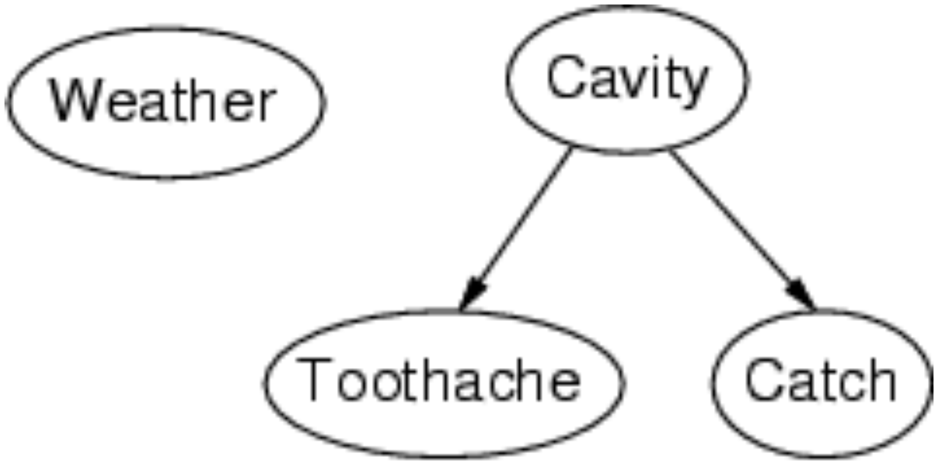
$$\mathbf{P}(X_i \mid Parents(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Following slides by Hwee Tou Ng (Singapore)

Example (1)

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

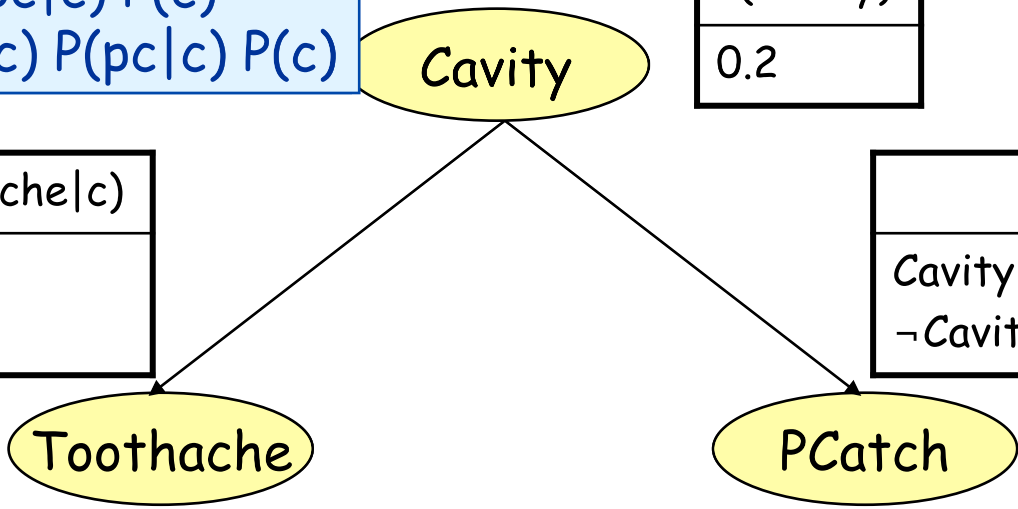
Bayesian Network

- Notice that Cavity is the "cause" of both Toothache and PCatch, and represent the causality links explicitly
- Give the prior probability distribution of Cavity
- Give the conditional probability tables of Toothache and PCatch

$P(c \wedge t \wedge pc) = P(t \wedge pc|c)\,P(c)$

Cond. Ind $= P(t|c)\,P(pc|c)\,P(c)$

P(Cavity)
0.2



	P(Toothache|c)
Cavity	0.6
¬Cavity	0.1

	P(PCatch|c)
Cavity	0.9
¬Cavity	0.02

5 probabilities, instead of 7

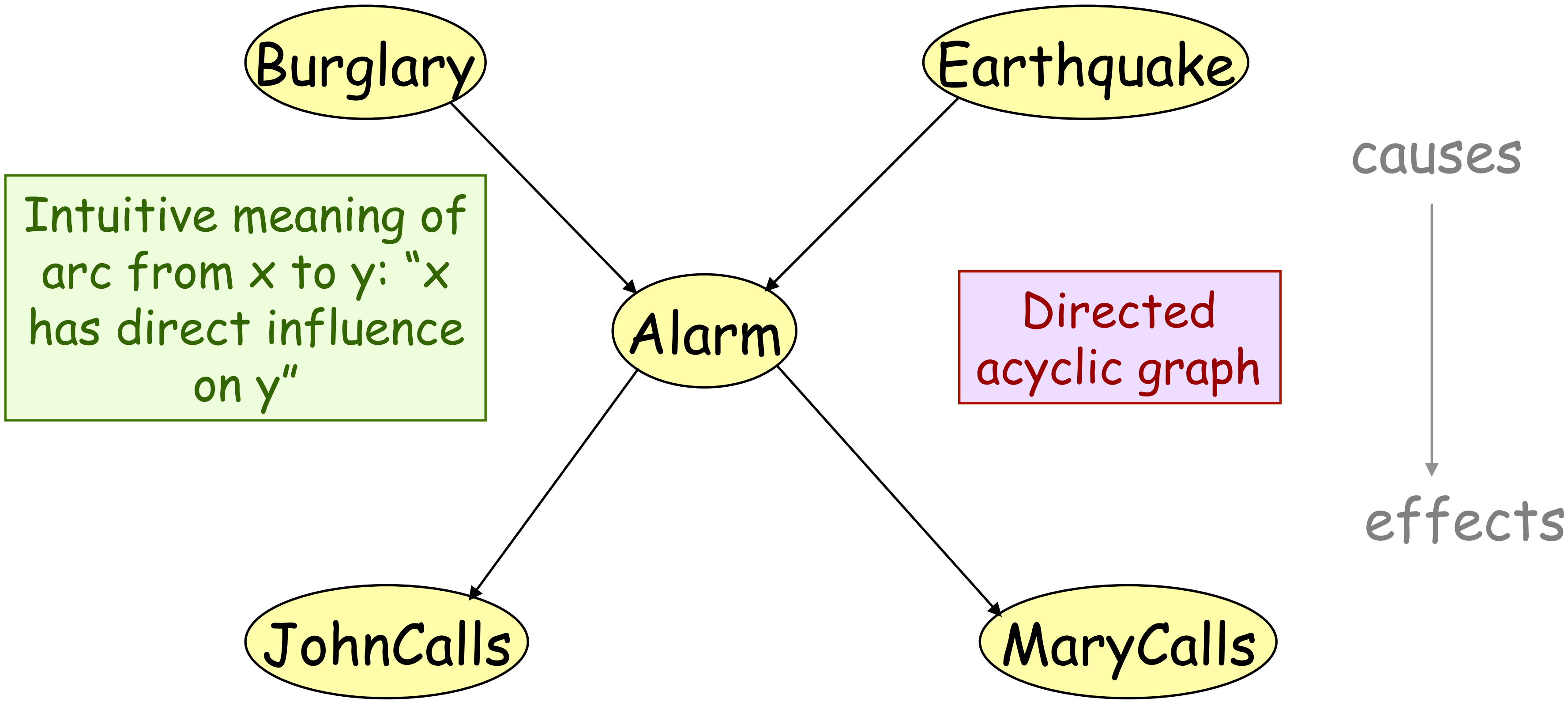
Example (2)

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

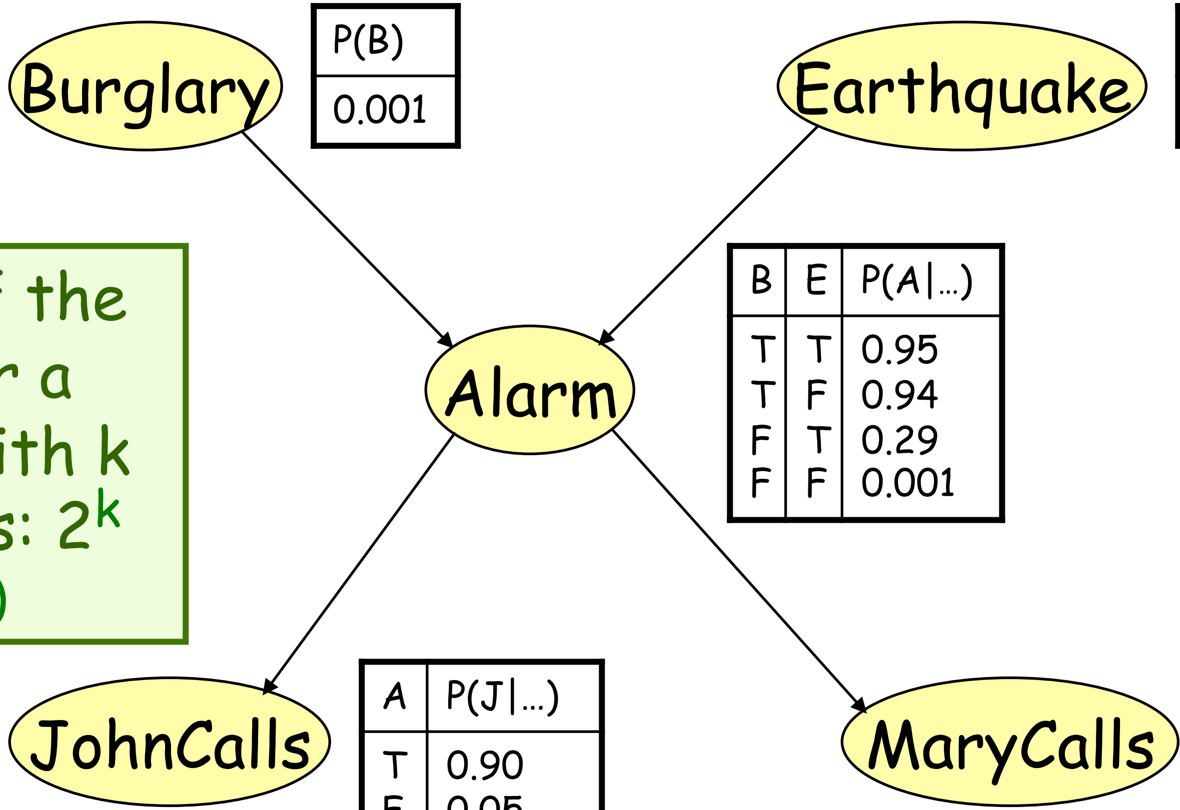
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

A More Complex BN



A More Complex BN



P(B)
0.001

P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

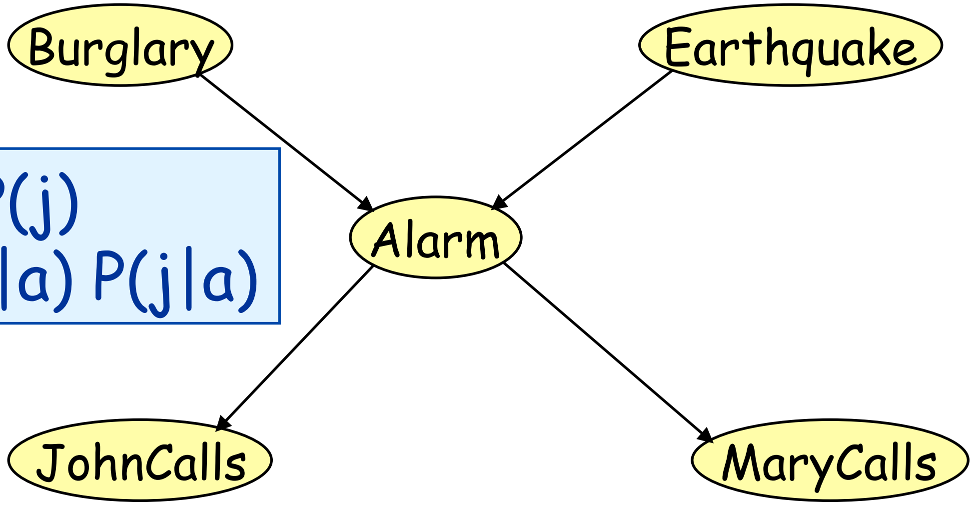
A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

Size of the CPT for a node with k parents: 2^k
($\sum$row=1)

10 probabilities, instead of 31

What does the BN encode?

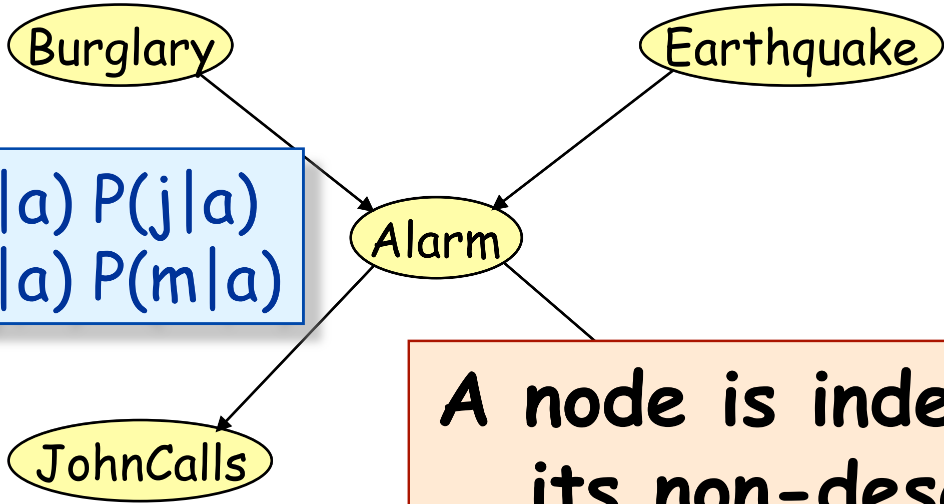


$P(b \wedge j) \neq P(b)\,P(j)$
$P(b \wedge j \mid a) = P(b \mid a)\,P(j \mid a)$

Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or ¬Alarm

For example, John does not observe any burglaries directly

What does the BN encode?



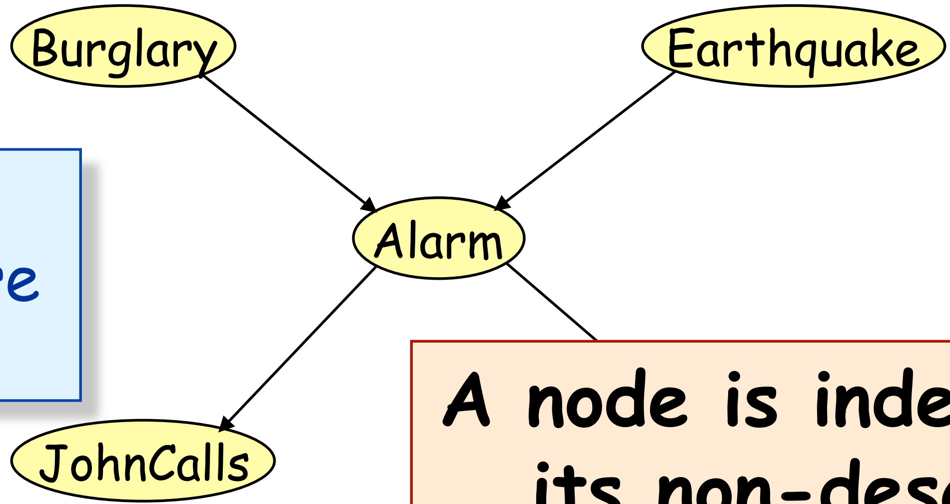
$P(b \wedge j | a) = P(b|a)\ P(j|a)$
$P(j \wedge m | a) = P(j|a)\ P(m|a)$

A node is independent of its non-descendants given its parents

The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

What does the BN encode?



Burglary and Earthquake are independent

A node is independent of its non-descendants given its parents

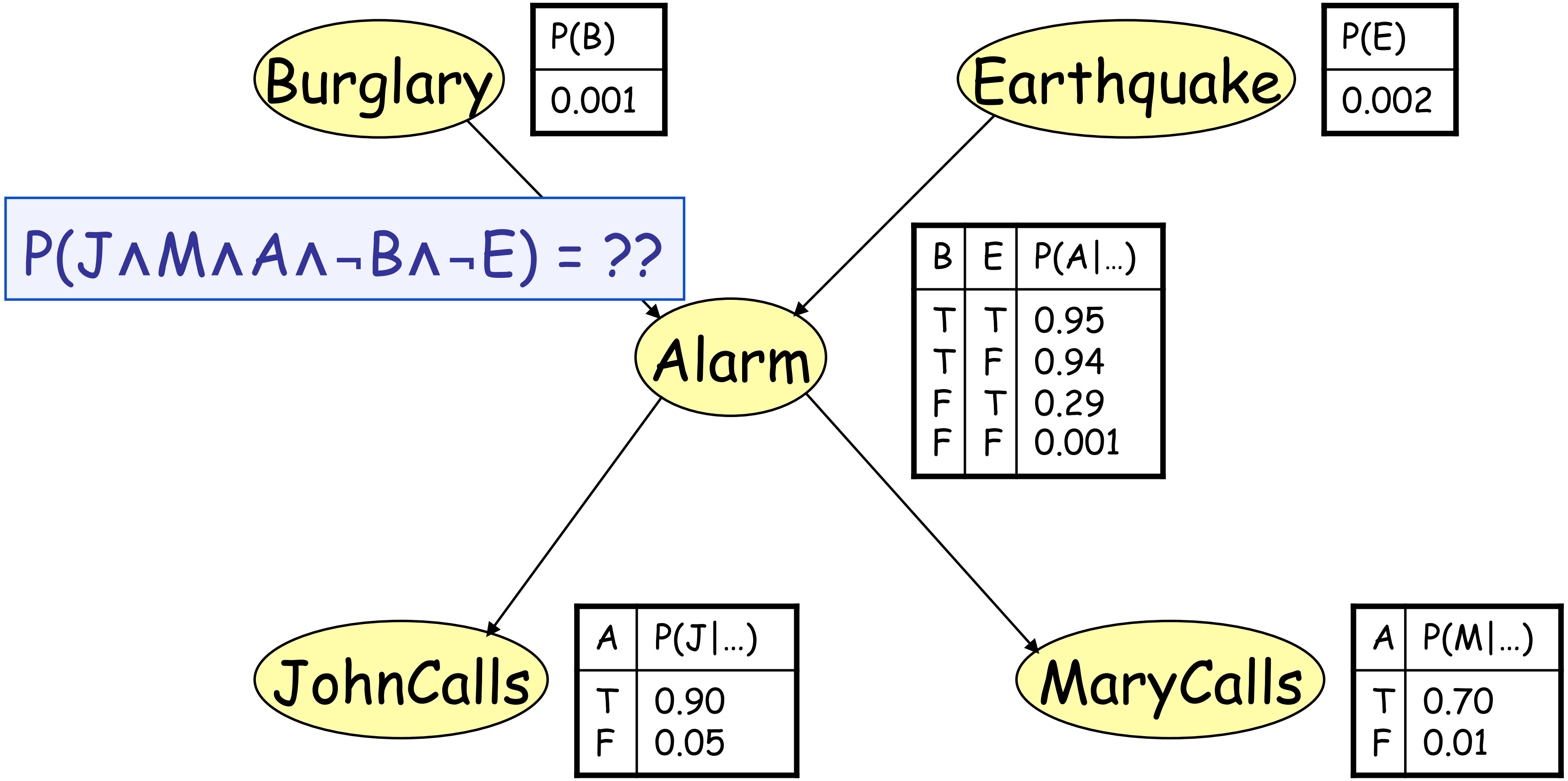
The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

Locally Structured World

- A world is locally structured (or sparse) if each of its components interacts directly with relatively few other components
- In a sparse world, the CPTs are small and the BN contains much fewer probabilities than the full joint distribution
- If the # of entries in each CPT is bounded by a constant, i.e., O(1), then the # of probabilities in a BN is linear in n – the # of propositions – instead of 2^n for the joint distribution

Calculation of Joint Probability



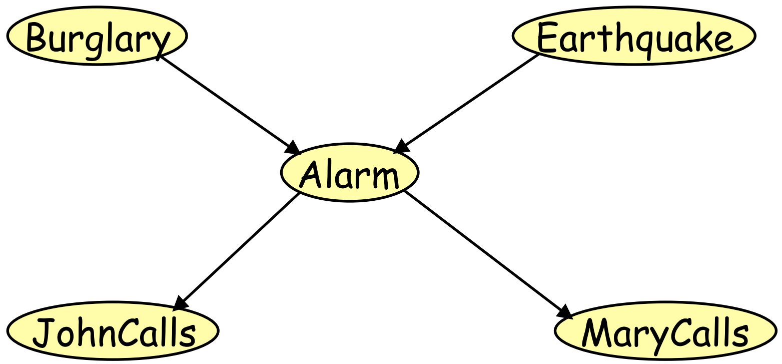
P(B)
0.001

P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

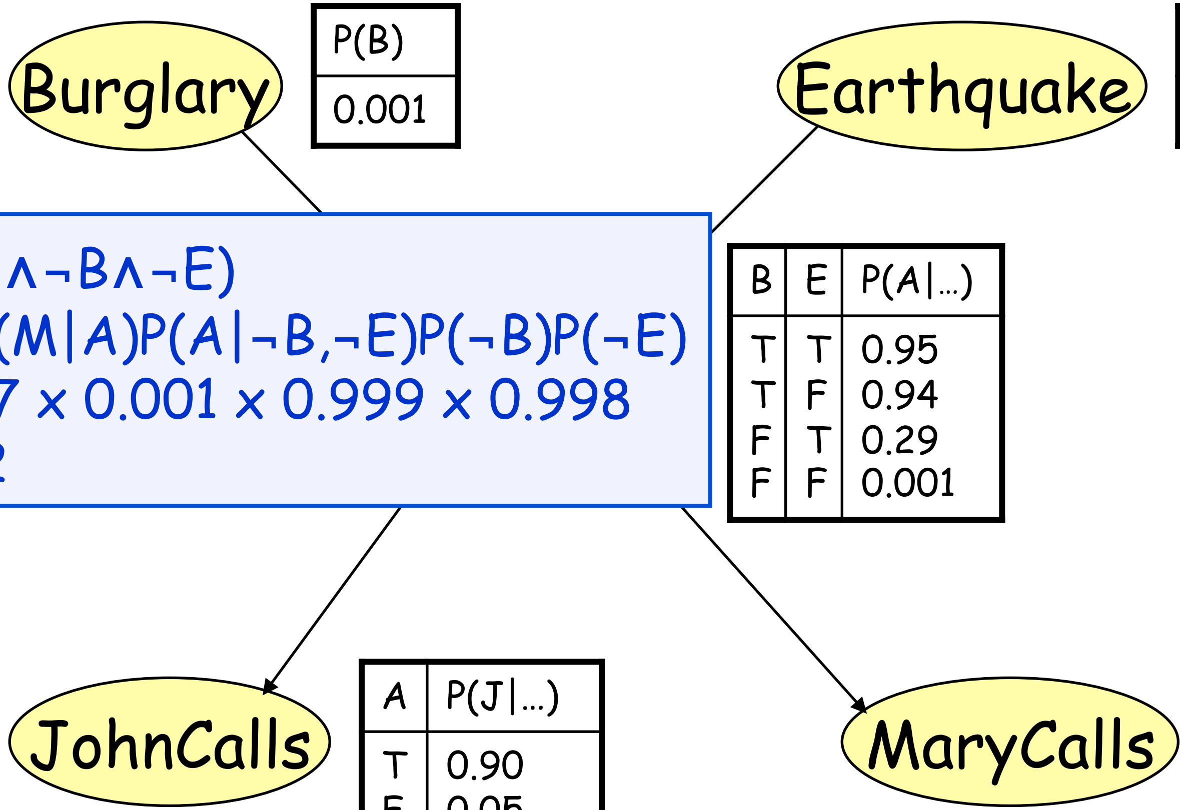
A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01



- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
  $= P(J \wedge M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  $= P(J | A, \neg B, \neg E) \times P(M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  (J and M are independent given A)
- $P(J | A, \neg B, \neg E) = P(J | A)$
  (J and $\neg B \wedge \neg E$ are independent given A)
- $P(M | A, \neg B, \neg E) = P(M | A)$
- $P(A \wedge \neg B \wedge \neg E) = P(A | \neg B, \neg E) \times P(\neg B | \neg E) \times P(\neg E)$
  $= P(A | \neg B, \neg E) \times P(\neg B) \times P(\neg E)$
  ($\neg B$ and $\neg E$ are independent)
- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J | A) P(M | A) P(A | \neg B, \neg E) P(\neg B) P(\neg E)$

Calculation of Joint Probability



$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

P(B)
0.001

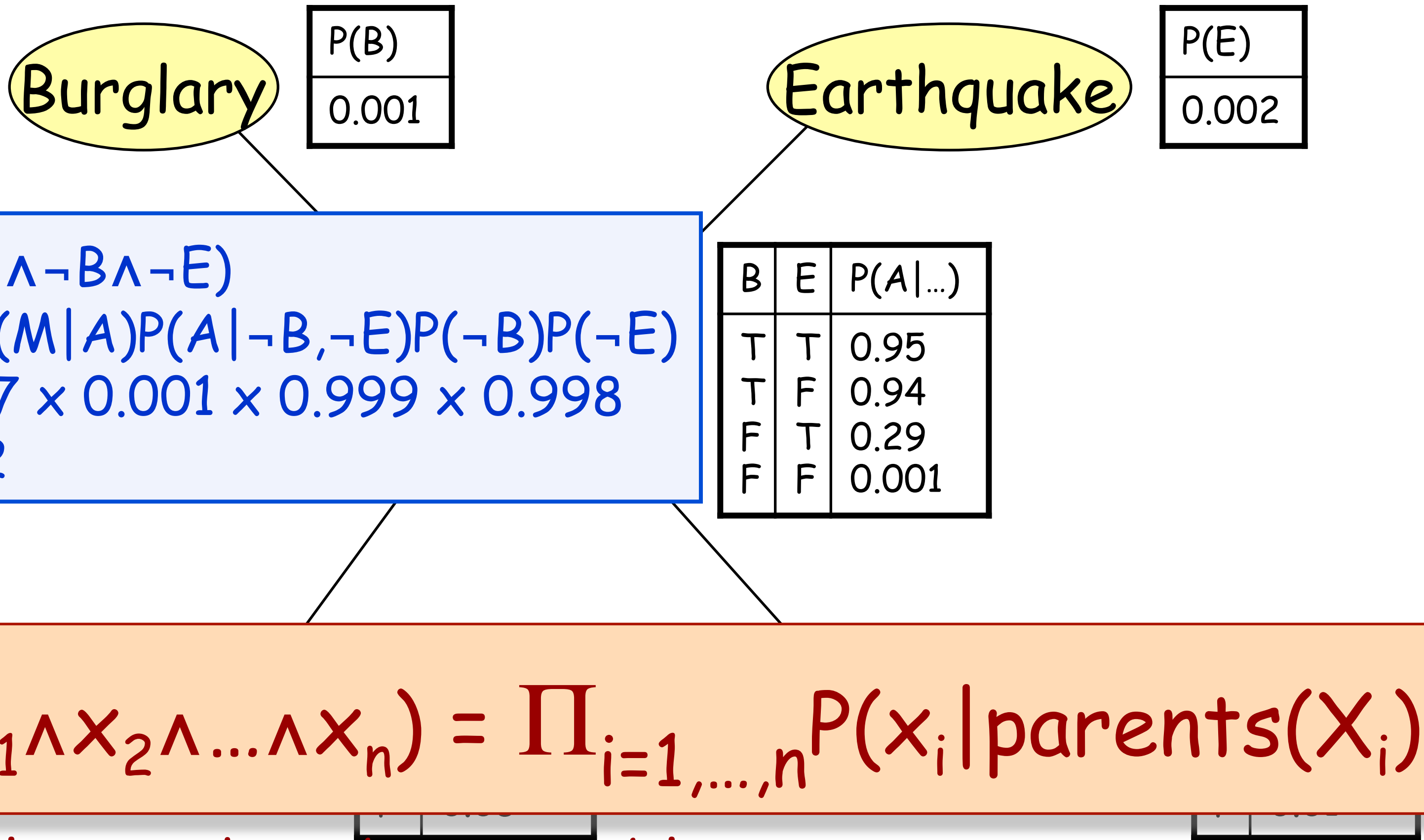
P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

Calculation of Joint Probability



P(B)
0.001

P(E)
0.002

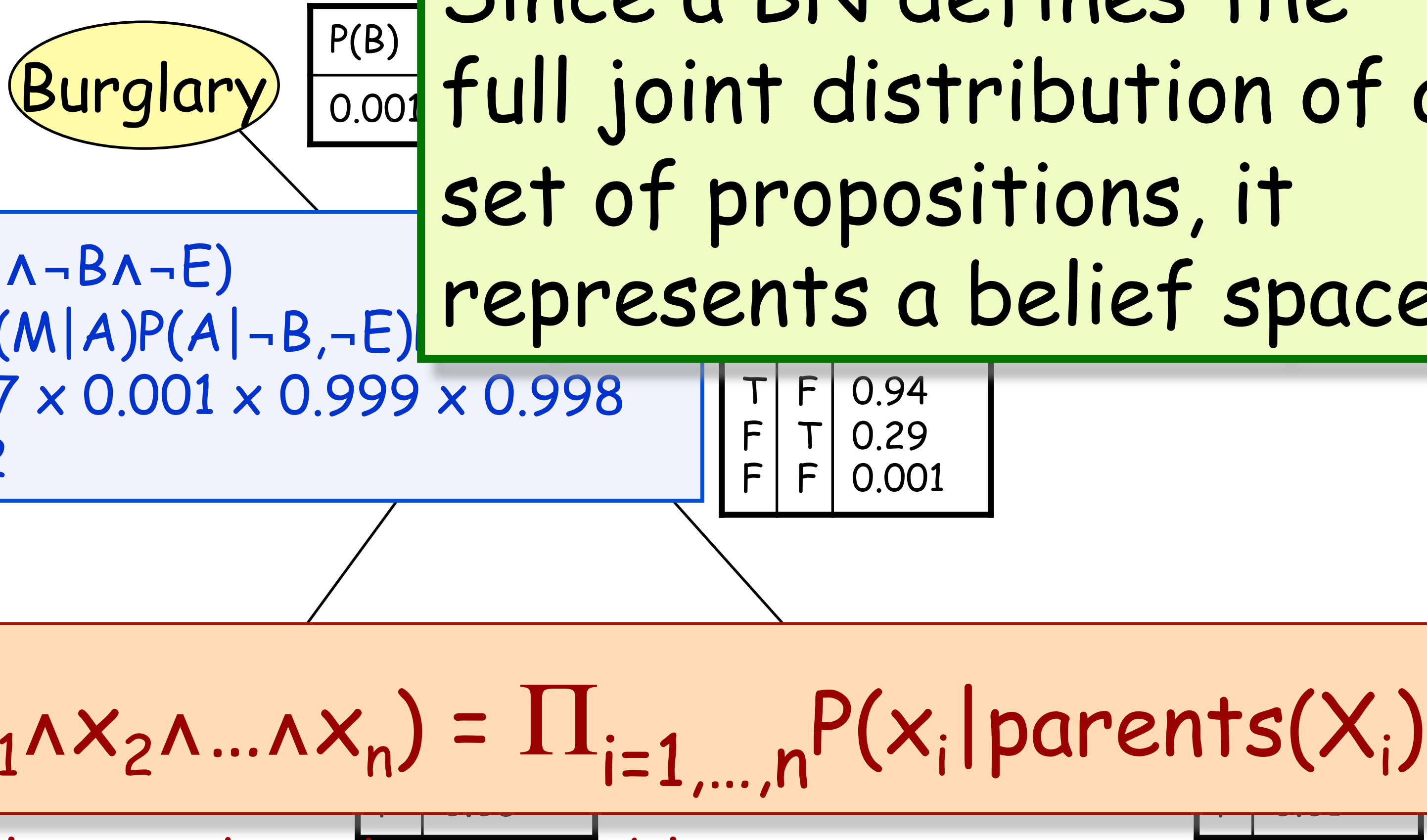
B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

$P(J \land M \land A \land \neg B \land \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

$$P(x_1 \land x_2 \land \ldots \land x_n) = \Pi_{i=1,\ldots,n} P(x_i|\mathrm{parents}(X_i))$$

→ full joint distribution table

Calculation of Joint Probability



$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)$...
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

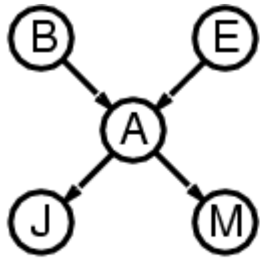
Since a BN defines the full joint distribution of a set of propositions, it represents a belief space

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_n) = \Pi_{i=1,\ldots,n} P(x_i|\text{parents}(X_i))$$

→ full joint distribution table

Semantics of a BN: Full Joint Distribution

The full joint distribution is defined as the product of the local conditional distributions:



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

e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

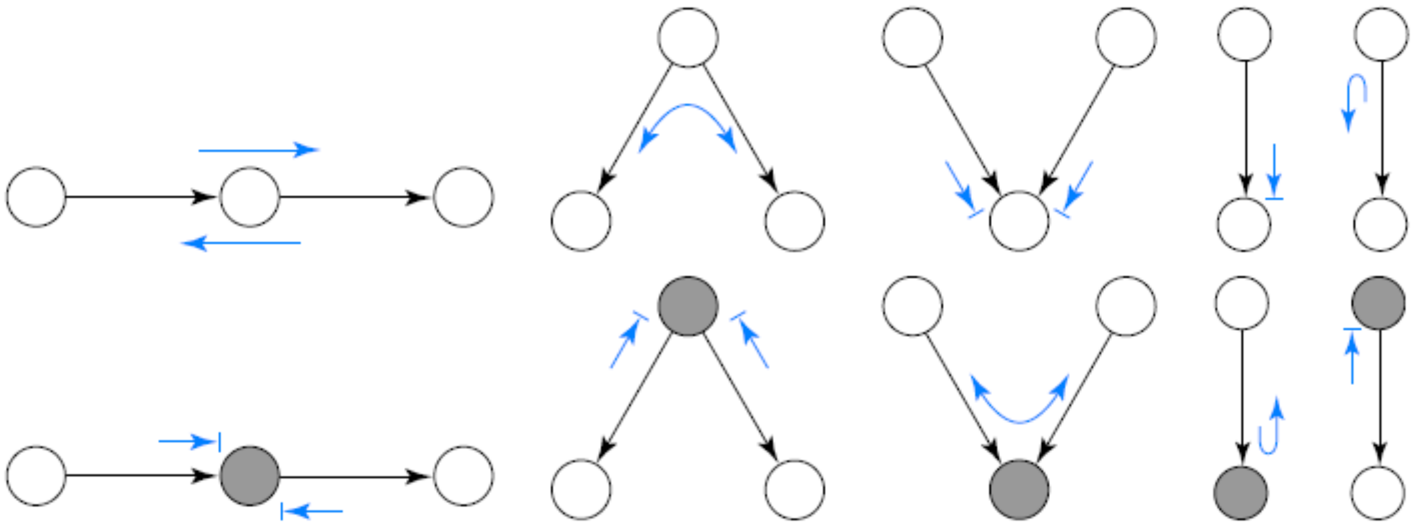
$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$

Encoding conditional independence via d-separation

- Bayesian networks encode the independence properties of a density.
- We can determine if a conditional independence $X \perp\!\!\!\perp Y \mid \{Z_1, \ldots, Z_k\}$ holds by appealing to a graph separation criterion called *d-separation* (which stands for *direction-dependent separation*).
- X and Y are d-separated if there is no *active path* between them.
- The formal definition of active paths is somewhat involved. The *Bayes Ball Algorithm* gives a nice graphical definition.

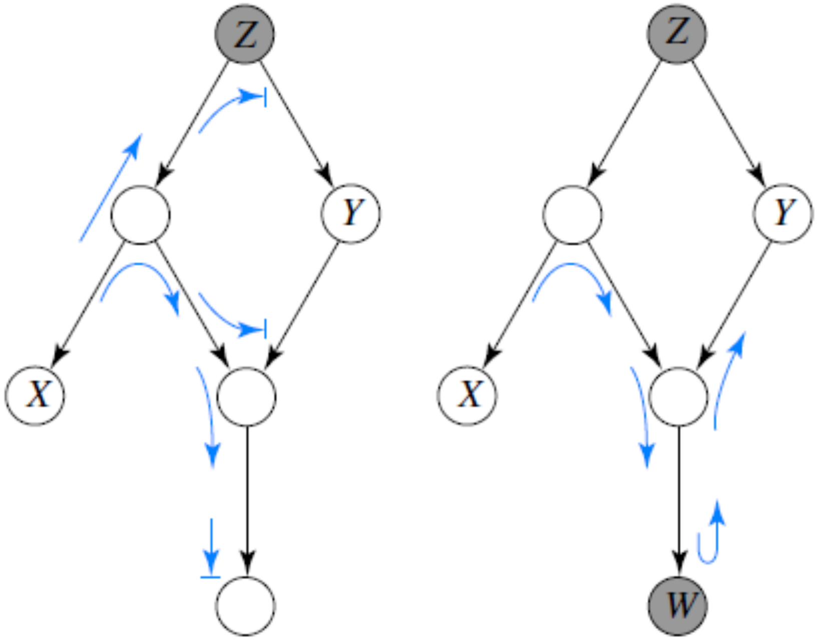
The ten rules of Bayes Ball

An undirected path is active if a Bayes ball travelling along it never encounters the "stop" symbol: $\longrightarrow\!\mid$



If there are no active paths from X to Y when $\{Z_1, \ldots, Z_k\}$ are shaded, then $X \perp\!\!\!\perp Y \mid \{Z_1, \ldots, Z_k\}$.

A double-header: two games of Bayes Ball



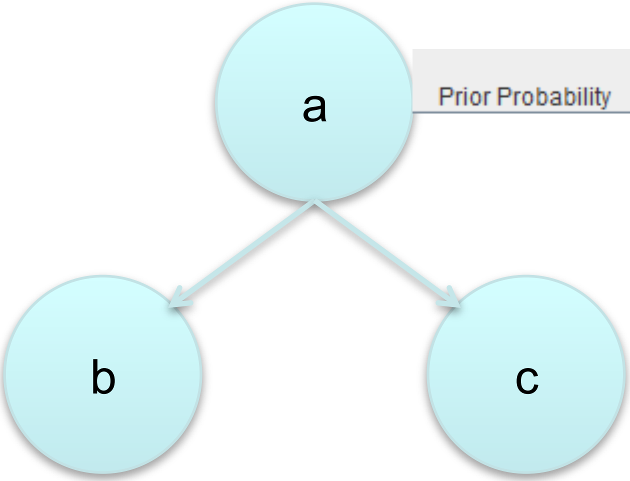
no active paths

$X \perp\!\!\!\perp Y \mid Z$

one active path

$X \not\perp\!\!\!\perp Y \mid \{W, Z\}$

Demonstration of CI



	P(a=T)	P(a=F)
Prior Probability	0.5	0.5

a	P(b=T)	P(b=F)
T	0.7	0.3
F	0.3	0.7

a	P(c=T)	P(c=F)
T	0.4	0.6
F	0.3	0.7

P(b=true) ?

P(c=true) ?

Suppose a is given. What are the effects on b,c

What if also b is given. How does this affect c?

What if only c is given. Effects on a on b?

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \dots, X_n$
- 2. For i = 1 to n
  - add X_i to the network
  - select parents from $X_1, \dots, X_{i-1}$ such that
    $$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \dots X_{i-1})$$

This choice of parents guarantees:

$\mathbf{P}(X_1, \dots, X_n) \quad = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$

(chain rule)

$\quad = \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$

(by construction)

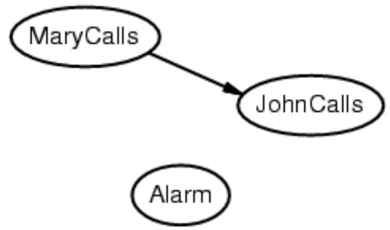
Example

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

P(*J* | *M*) = ***P***(*J*)?

Example

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



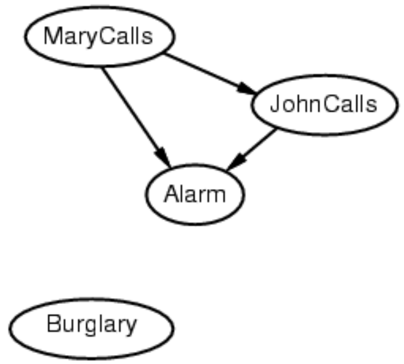
$\mathbf{P}(J \mid M) = \mathbf{P}(J)?$

No

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

Example

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



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

No

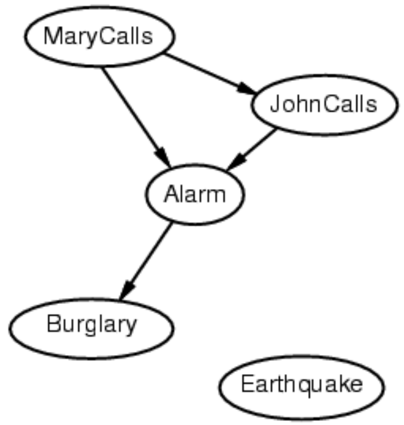
$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **No**

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

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

Example

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



P(*J* | *M*) = ***P***(*J*)?

No

P(*A* | *J, M*) = ***P***(*A* | *J*)? ***P***(*A* | *J, M*) = ***P***(*A*)? **No**

P(*B* | *A, J, M*) = ***P***(*B* | *A*)? **Yes**

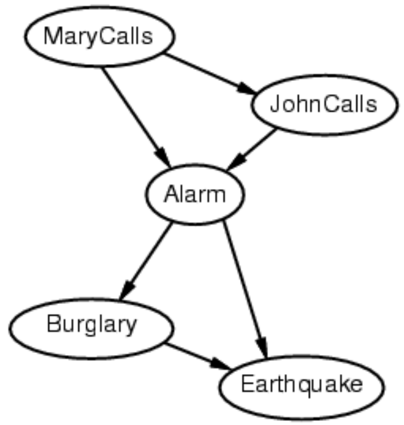
P(*B* | *A, J, M*) = ***P***(*B*)? **No**

P(*E* | *B, A ,J, M*) = ***P***(*E* | *A*)?

P(*E* | *B, A, J, M*) = ***P***(*E* | *A, B*)?

Example

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



***P**(J | M) = **P**(J)?*

No

***P**(A | J, M) = **P**(A | J)*? ***P**(A | J, M) = **P**(A)*? **No**

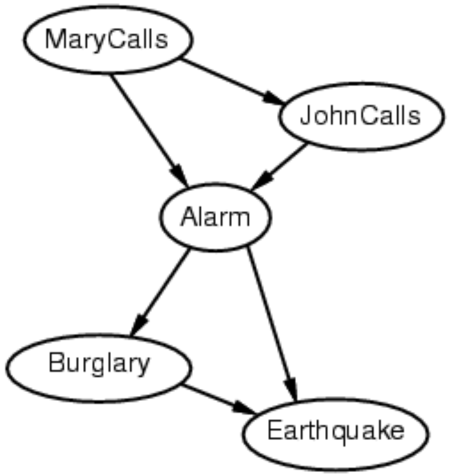
***P**(B | A, J, M) = **P**(B | A)*? **Yes**

***P**(B | A, J, M) = **P**(B)*? **No**

***P**(E | B, A ,J, M) = **P**(E | A)*? **No**

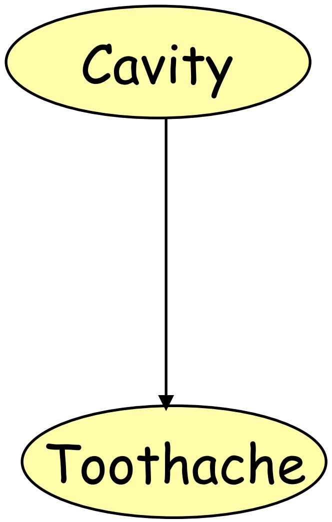
***P**(E | B, A, J, M) = **P**(E | A, B)*? **Yes**

Example contd.



- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Querying the BN



P(C)
0.1

C	P(T|c)
T	0.4
F	0.01111

- The BN gives P(t|c)
- What about P(c|t)?
- P(Cavity|t)
  = P(Cavity ∧ t)/P(t)
  = P(t|Cavity) P(Cavity) / P(t)
  [Bayes' rule]
- P(c|t) = α P(t|c) P(c)
- Querying a BN is just applying the trivial Bayes' rule on a larger scale

Slides:
Jean Claude Latombe

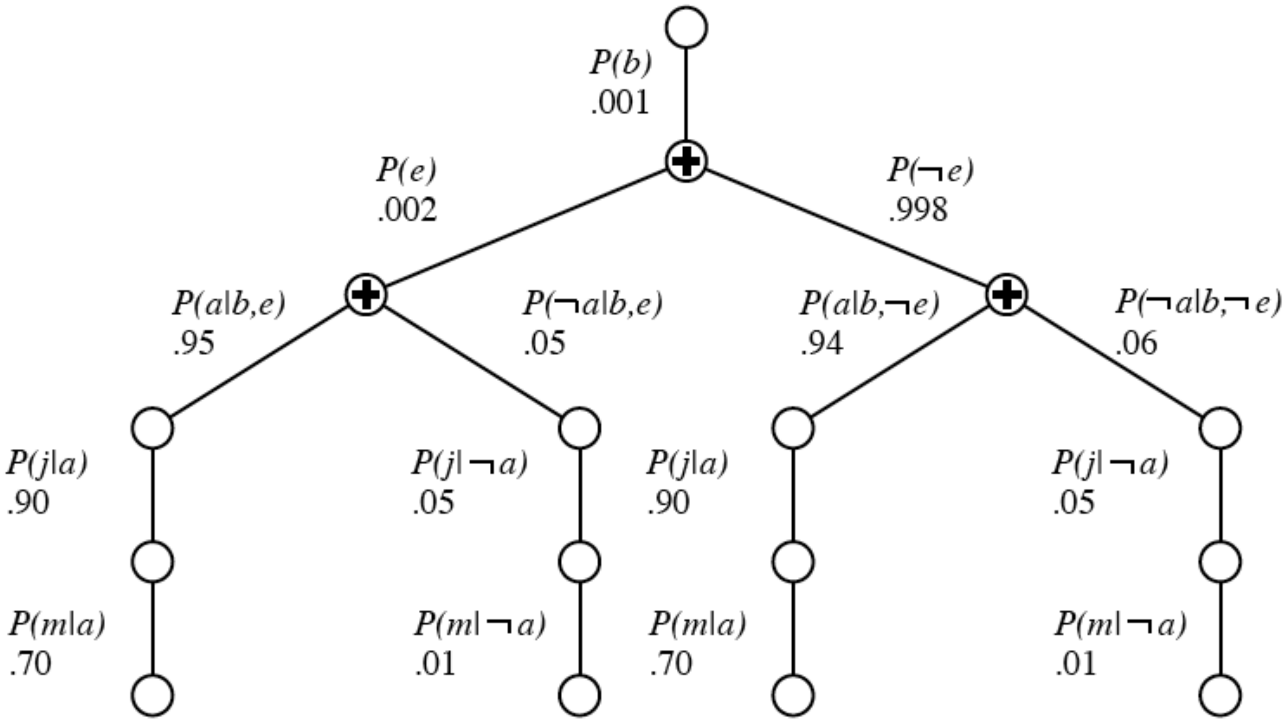
Querying the BN

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

  $= \alpha \sum_a \sum_e P(b \wedge j \wedge m \wedge a \wedge e)$ [marginalization]

  $= \alpha \sum_a \sum_e P(b)P(e)P(a|b,e)P(j|a)P(m|a)$ [BN]

  $= \alpha\, P(b) \sum_e P(e) \sum_a P(a|b,e)P(j|a)P(m|a)$ [re-ordering]

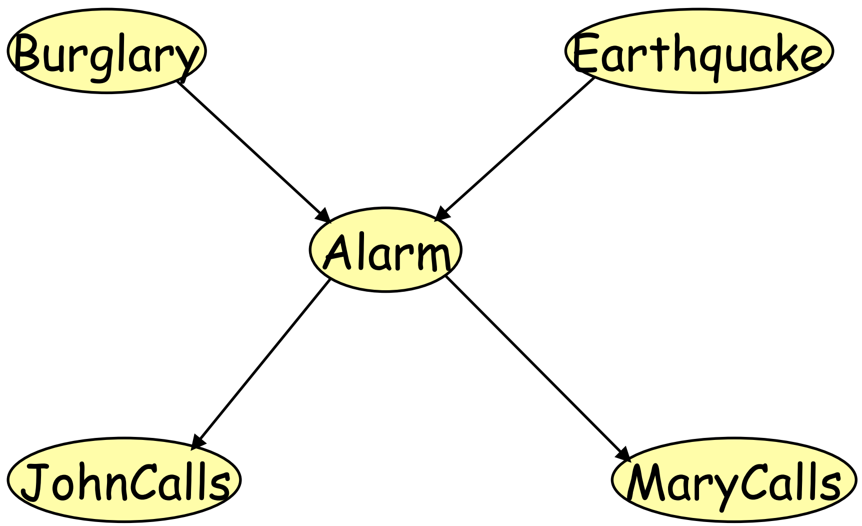


Querying the BN

- Depth-first evaluation of P(b|j) leads to computing each of the 2 following products twice:
  P(j|a) P(m|a), P(j|¬a) P(m|¬a)
- Bottom-up (right-to-left) computation + caching – e.g., variable elimination algorithm (see R&N) – avoids such repetition
- For singly connected BN, the computation takes time linear in the total number of CPT entries (→ time linear in the # propositions if CPT's size is bounded)

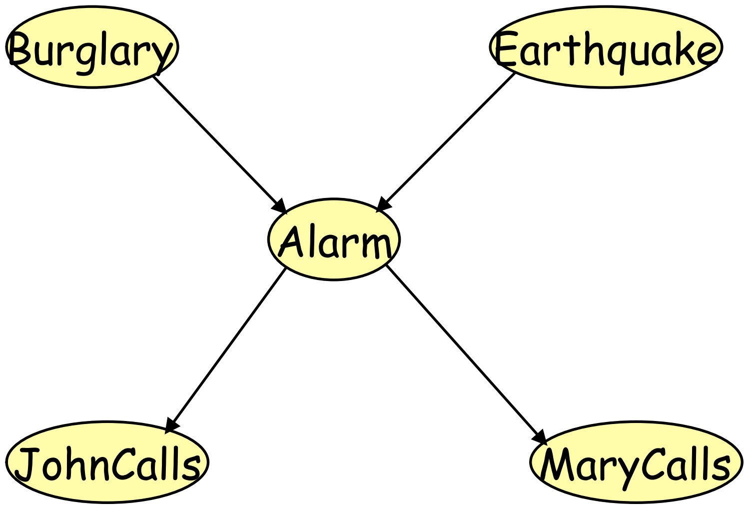
Singly Connected BN

A BN is singly connected if there is at most one undirected path between any two nodes



is singly connected

Comparison to Classical Logic



Burglary → Alarm
Earthquake → Alarm
Alarm → JohnCalls
Alarm → MaryCalls

φ	ψ	$\varphi \Rightarrow \psi$
T	T	T
T	F	F
F	T	T
F	F	T

If the agent observes
¬JohnCalls,
it infers ¬Alarm,
¬Burglary, and ¬Earthquake

If it observes JohnCalls, then it infers nothing

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct