

CPSC 322: Introduction to Artificial Intelligence

Uncertainty: Belief Networks and Introduction to VE

Textbook reference: [8.3,8.4]


Instructor: Varada Kolhatkar
University of British Columbia

Credit: These slides are adapted from the slides of the previous offerings of the course. Thanks to all instructors for creating and improving the teaching material and making it available!

Announcements

- Teaching evaluations are open. You should have received an email.
 - I am teaching undergrad for the first time and I will very much appreciate constructive feedback.
- Final exam
 - **Time:** Dec 9 at 7:00pm and **Location:** SRC A
 - Difficulty level: Given that you did so well on midterm we would like to challenge you a bit in the final. So please start studying now and make use of all the help available to you.
- Assignment 4 has been released.
 - Due date: **Nov 29th, 11:59 PM**

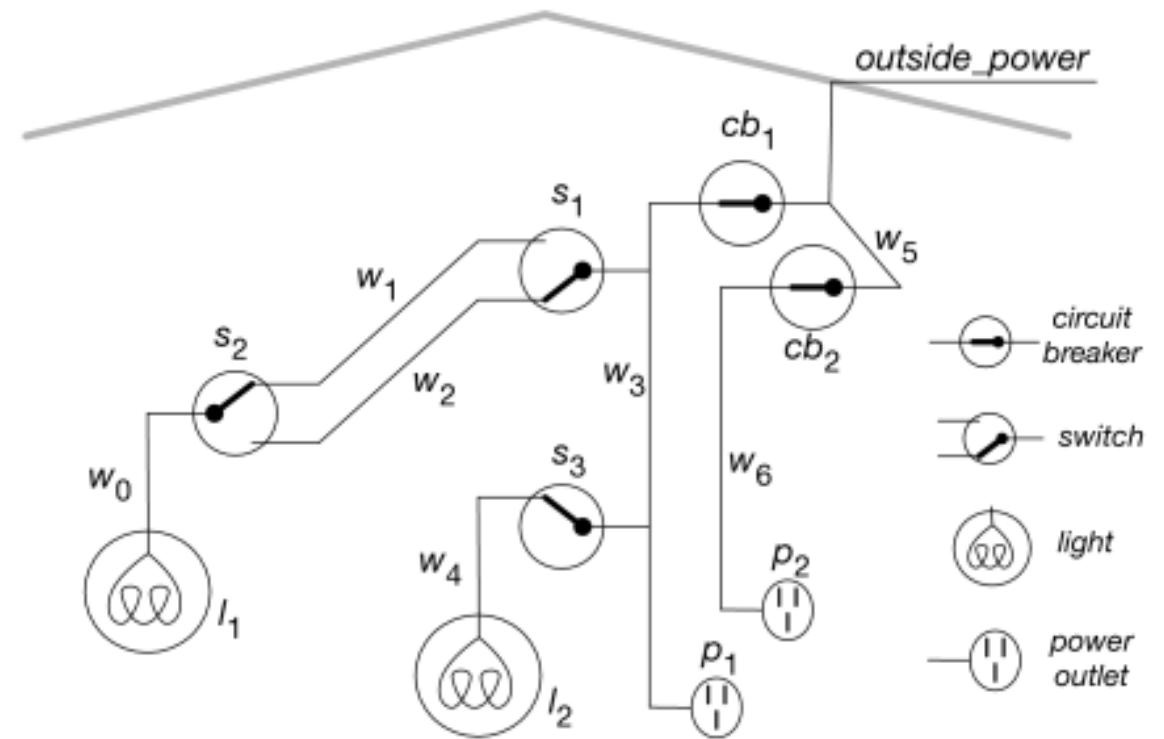
Lecture outline

- Recap 
- Bayesian networks definition and examples
- Bayesian networks types of inference
- Factors and factor operations

Conditional independence: example

Whether light l_1 is lit or not is conditionally independent from the position of the switch s_2 given whether there is power in w_0 .

Once we know $Power(w_0)$, learning values from any other variable will not change our beliefs about $Lit(l_1)$.



$Lit(l_1)$ is independent of any other variable given $Power(w_0)$

Marginally but not conditionally independent: example

Two variables can be marginally but not conditionally independent

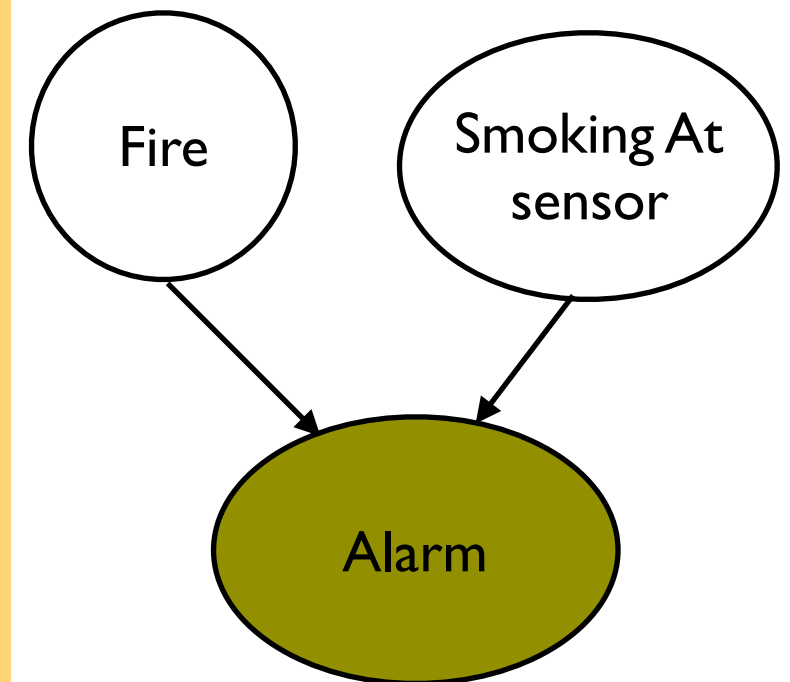
“Smoking At Sensor” (S): resident smokes cigarette next to fire sensor

“Fire” (F): there is a fire somewhere in the building

“Alarm” (A): the fire alarm rings

S and F are marginally independent: Learning $S=\text{true}$ or $S=\text{false}$ does not change your belief in F

But they are not conditionally independent given alarm: If the alarm rings and you learn $S=\text{true}$ your belief in F decreases



Why conditional independence?

- Conditional independence is more common.
- Marginal independence between variables is relatively rare in a given domain.
- Also, in any given domain, we typically know something, and independence queries should be conditional on that something.

Bayesian networks (BNs): Motivation

- With JPDs we can do inference by enumeration BUT that's **extremely inefficient**; it does not scale.
- We want a representation and reasoning system that is based on **conditional (and marginal) independence**.
- **Compact** yet **expressive** representation.
- **Efficient reasoning** procedures

BNs: Motivation

- Bayes[ian] (Belief) Net[work]s are such a representation.
- Named after Thomas Bayes (1702 –1761)
- Term coined in 1985 by Judea Pearl
- Their invention changed the primary focus of AI from logic to probability!

Thomas Bayes



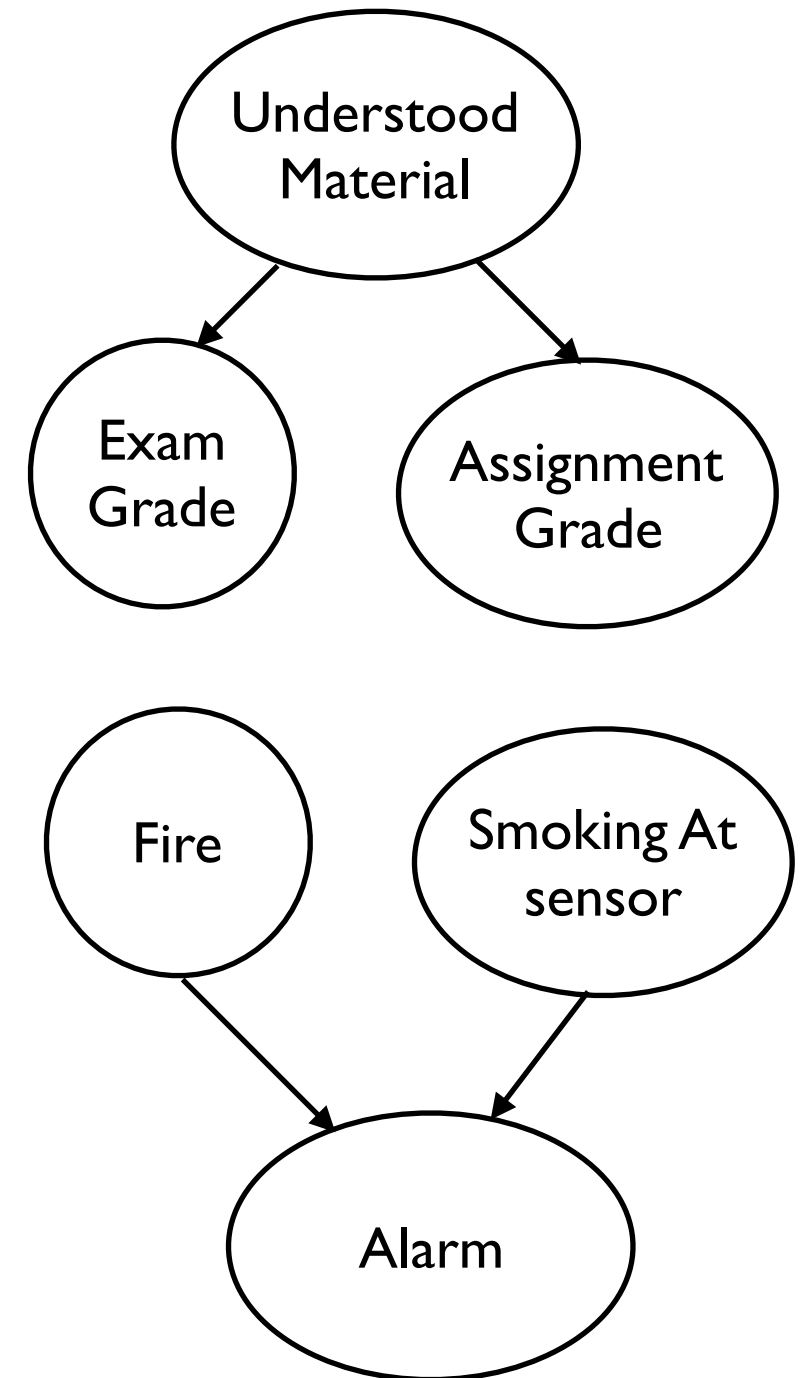
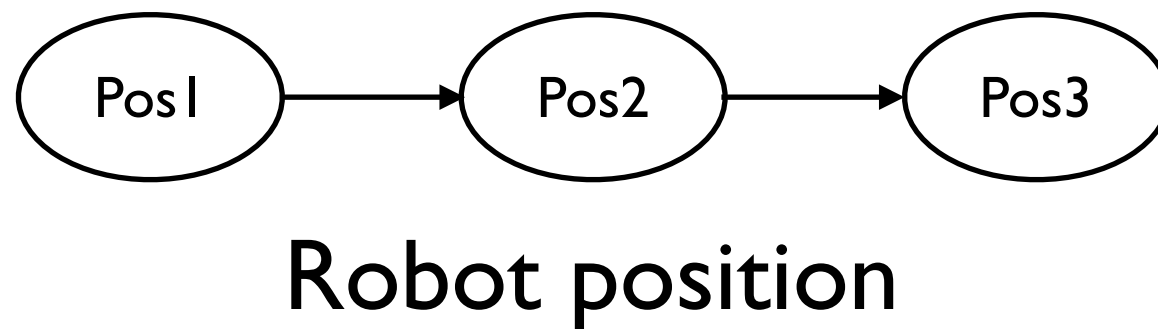
Judea Pearl



BNs intuition

A graphical representation
for a joint probability distribution

- Nodes are random variables
- Directed edges between nodes reflect dependence



Today: Learning outcomes

From this lecture, students are expected to be able to:

- Build a belief network for a simple domain
- Compute the representational savings in terms of number of probabilities required
- Classify the types of inference: diagnostic, predictive, inter-causal, mixed
- Define factors and apply operations to factors, including assigning, summing out and multiplying factors

Bayesian networks (BNs) definition

A **Bayesian network** consists of

- A **directed acyclic graph** (V, E) whose nodes are labeled with random variables
- A **domain** for each random variable
- A **conditional probability distribution** for each variable V
 - Specifies $P(V | Parents(V))$
 - $Parents(V)$ is the set of variables V' with $(V', V) \in E$. For nodes without predecessors, $Parents(V) = \{ \}$

The parents of V are the ones V directly depends upon.

A Bayesian network is a compact representation of the JPD:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i))$$

Discrete BNs: definition

A **Bayesian network** consists of

- A **directed acyclic graph** (V, E) whose nodes are labeled with random variables
- A **domain** for each random variable
- A **conditional probability distribution** for each variable V
 - Specifies $P(V | Parents(V))$
 - $Parents(V)$ is the set of variables V' with $(V', V) \in E$. For nodes without predecessors, $Parents(V) = \{\}$

Discrete Bayesian networks are the ones where domain of each variable is **finite**, conditional probability distribution is a **conditional probability table (CPT)**.

We will assume this discrete case. But everything we say about independence (marginal & conditional) carries over to the continuous case.

BNs: Conditional Probability Table (CPT)

- A **CPT** for boolean X_i with k boolean parents has rows for the combinations of parent values
- Each row requires one number P_i for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1 - P_i$)

BNs: Simple example

- Two Boolean variables: Disease (D) and Symptom (S)
- The **causal ordering**: D, S
- Chain rule:
$$P(D, S) = P(S) \times P(S | D)$$
- Is $D \perp\!\!\!\perp S \mid \{\}$?
- Are they marginally independent (conditioned on nothing)?

JPD

D	S	P(D,S)
t	t	0.0099
t	f	0.0001
f	t	0.0990
f	f	0.8910

Marginals

D	P(D)	S	P(S)
t	0.01	t	0.1089
f	0.99	f	0.8911

BNs: Simple example

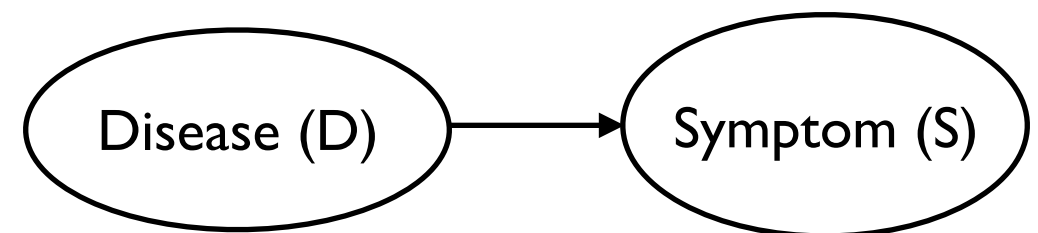
- Two Boolean variables: Disease (D) and Symptom (S)
- The causal ordering: D, S
- Chain rule: $P(D, S) = P(S) \times P(S | D)$
- Is $D \perp\!\!\!\perp S | \{\}$?
- Are they marginally independent (conditioned on nothing)?
No! That would mean $P(D, S) = P(S) \times P(D)$, which is not true.
We have to put an edge between the parent D and child S.

JPD

D	S	P(D,S)
t	t	0.0099
t	f	0.0001
f	t	0.0990
f	f	0.8910

Marginals

D	P(D)	S	P(S)
t	0.01	t	0.1089
f	0.99	f	0.8911



BNs: Simple example

Which conditional probability tables do we need?

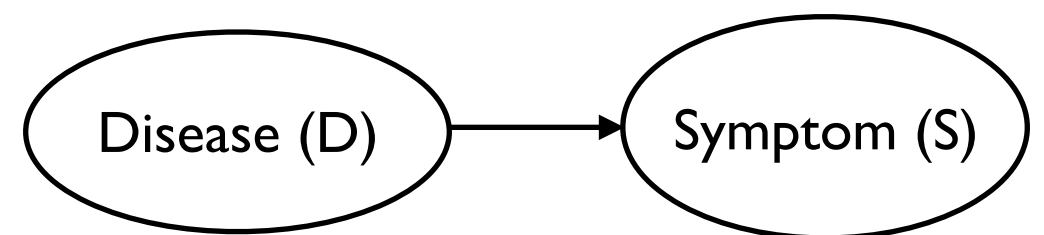
- A. $P(D)$
- B. $P(D|S)$
- C. $P(S|D)$
- D. $P(D, S)$

JPD

D	S	P(D,S)
t	t	0.0099
t	f	0.0001
f	t	0.0990
f	f	0.8910

Marginals

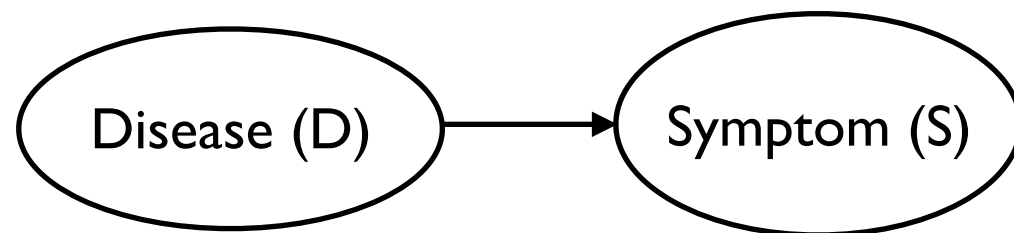
D	P(D)	S	P(S)
t	0.01	t	0.1089
f	0.99	f	0.8911



BNs: Simple example

Which conditional probability tables do we need?

$P(D)$ and $P(S|D)$



$P(D=t)$
0.01

D	$P(S=t D)$
t	$0.0099/(0.0099+0.0001) = 0.99$
f	$0.099/(0.099+0.891) = 0.1$

JPD

D	S	$P(D,S)$
t	t	0.0099
t	f	0.0001
f	t	0.0990
f	f	0.8910

Marginals

D	$P(D)$	S	$P(S)$
t	0.01	t	0.1089
f	0.99	f	0.8911

Once we have the CPTs in the network, we can compute any entry of the JPD

Constructing BNs

- Totally order the variables: e.g., X_1, \dots, X_n (e.g., temporal or causal order)
- Use chain rule with that ordering: $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{i-1} \dots X_1)$
- For every variable X_i , find the smallest set of parents $Pa(X_i) \subseteq \{X_1, \dots, X_{i-1}\}$ such that $X_i \perp\!\!\!\perp \{X_1, \dots, X_{i-1}\} / Pa(X_i)$
 X_i is conditionally independent from its other ancestors given its parents.
- Then we can rewrite $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i))$
 - This is a compact representation of JPD.

Constructing BNs

- Nodes are the random variables
- Directed arc from each variable in $Pa(X_i)$ to X_i
- For every variable X_i , construct its **conditional probability table (CPT)** $P(X_i | Pa(X_i))$. This has to specify a conditional probability distribution $P(X_i | Pa(X_i) = pa(X_i))$ for every instantiation $pa(X_i)$ of X_i 's parents

Instantiation of parents of a variable



For every variable X_i , construct its conditional probability table $P(X_i | Pa(X_i))$. This has to specify a conditional probability distribution $P(X_i | Pa(X_i) = pa(X_i))$ for every instantiation $pa(X_i)$ of X_i 's parents

If a variable has 3 parents each of which has a domain with 4 values, how many instantiations of its parents are there?

A. 4^3 

C. 3×4

B. $3^4 - 1$

D. 4

BNs: Compactness

- If each variable has no more than k parents, the complete network requires $O(n2^k)$ numbers
- For $k \ll n$, this is a substantial improvement.
- The numbers required grow linearly with n , compared to $O(2^n)$ for the full joint distribution

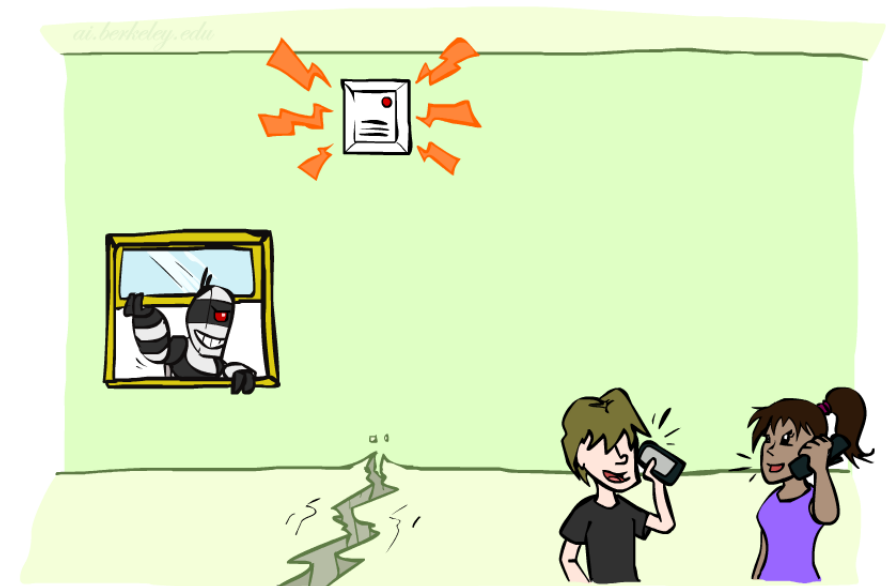
BNs: Burglary example

There might be a burglar in my house (B)

The anti-burglar alarm in my house may go off (A)

I have an agreement with two of my neighbours, John and Mary, that they call me if they hear the alarm go off when I am at work (M, J)

Minor earthquakes may occur and sometimes the set off the alarm (E)



Source: Berkeley AI material

BNs: Burglary example

Order variables to reflect causal knowledge
(i.e., causes before effects)

A burglar (B) can set the alarm (A) off

An earthquake (E) can set the alarm (A) off

The alarm can cause Mary to call (M)

The alarm can cause John to call (J)

Calculate the joint $P(B, E, A, M, J)$



BNs: Burglary example

Order variables to reflect causal knowledge
(i.e., causes before effects)

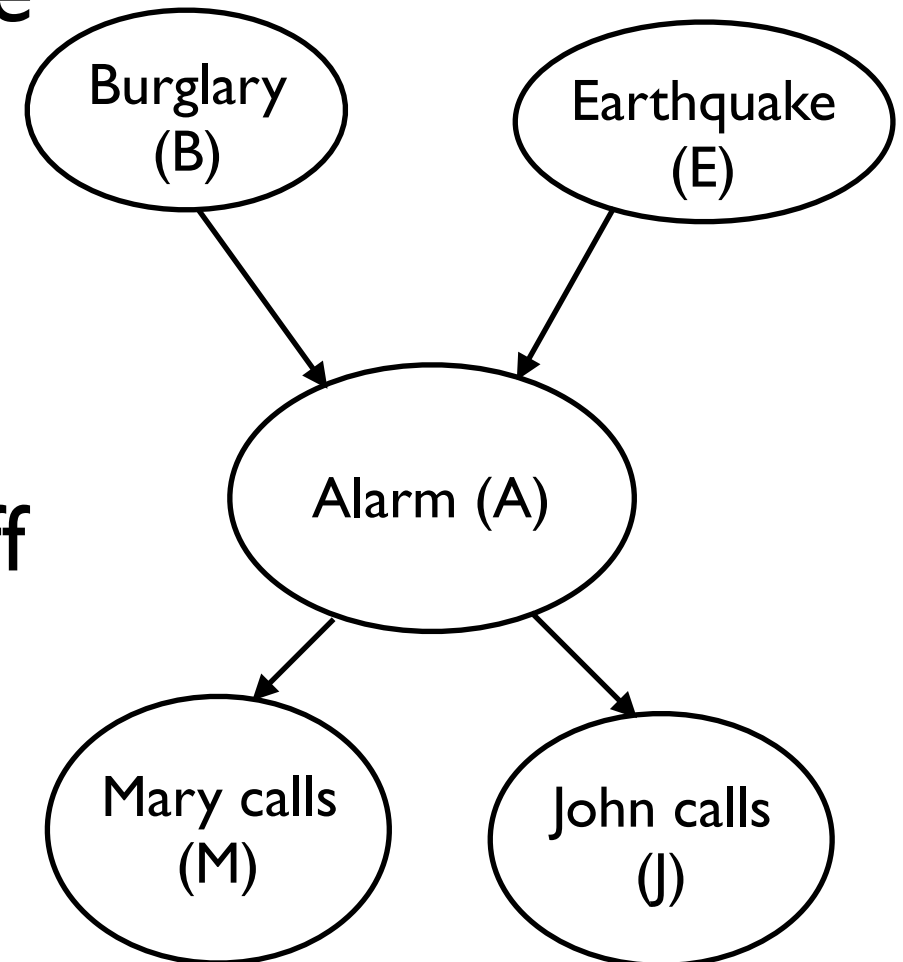
A burglar (B) can set the alarm (A) off

An earthquake (E) can set the alarm (A) off

The alarm can cause Mary to call (M)

The alarm can cause John to call (J)

Calculate the joint $P(B, E, A, M, J)$



BNs: Burglary example




Boolean variables: B, A, M, J, E ($n = 5$)

We want to calculate the joint $P(B, E, A, M, J)$

How many entries (probabilities) do we need to store with a JPD?

A. 10

C. 31 

B. 32

D. 16

BNs: Burglary example

- Goal: Calculate the joint $P(B, E, A, M, J)$

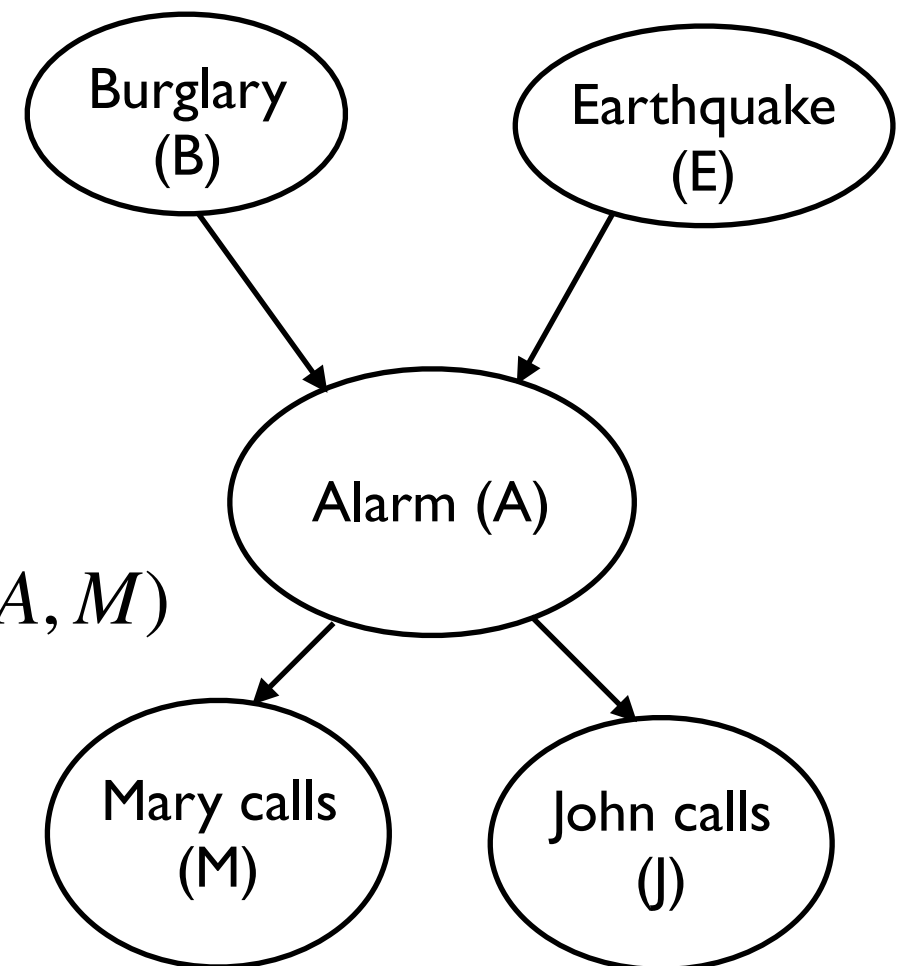
- Apply chain rule $P(B, E, A, M, J)$

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

- Simplify according to marginal and conditional independence

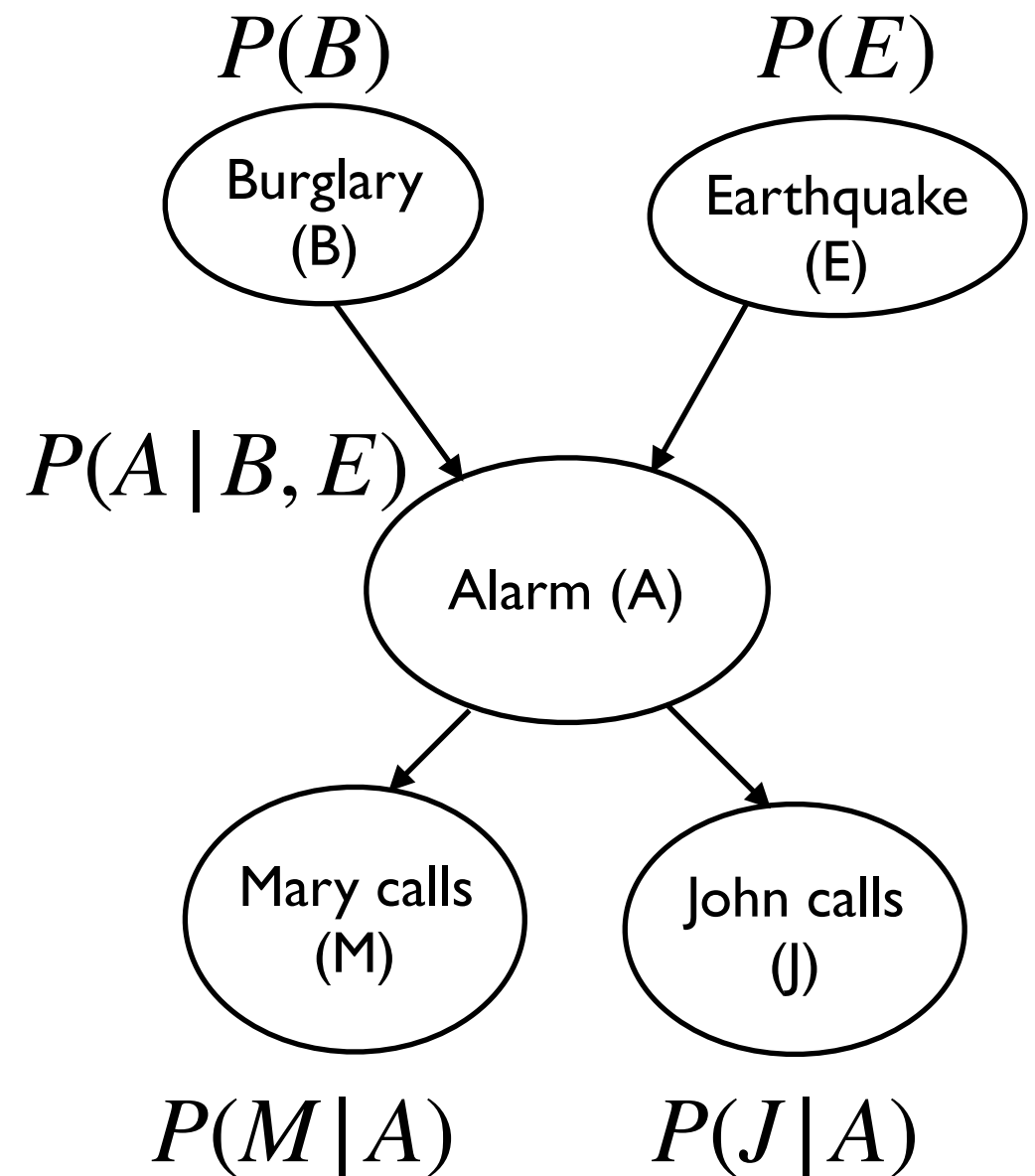
$$P(B, E, A, M, J)$$

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



BNs: Burglary example

- Simplified joint:
 $P(B, E, A, M, J) = P(B)P(E)P(A | B, E)P(M | A)P(J | A)$
- Express dependencies as a network (directed acyclic graph (DAG))
 - Each variable is a node
 - For each variable, the conditioning variables are its parents
- Associate to each node corresponding conditional probabilities



BNs: Burglary example

$$P(B)$$

$P(B=T)$	$P(B=F)$
0.001	0.999

$$P(E)$$

$P(E=T)$	$P(E=F)$
0.002	0.998

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

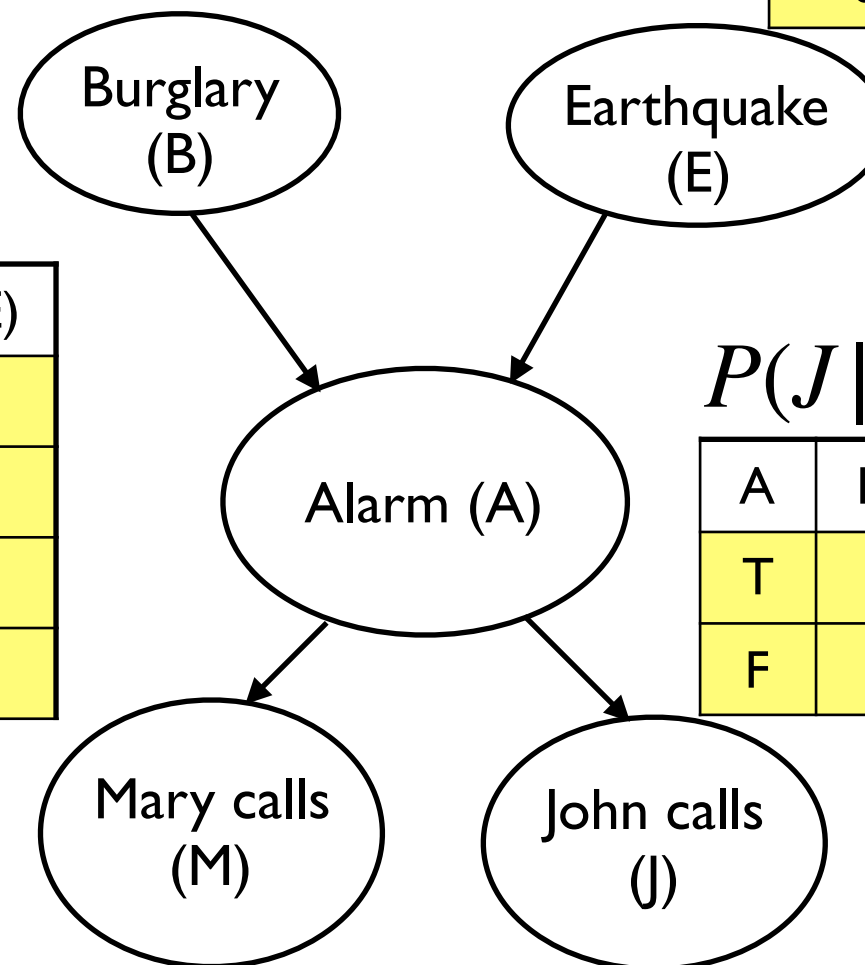
B	E	$P(A=T B,E)$	$P(A=F B,E)$
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

$$P(J | A)$$

A	$P(J=T A)$	$P(J=F A)$
T	0.90	0.10
F	0.05	0.95

$$P(M | A)$$

A	$P(M=T A)$	$P(M=F A)$
T	0.70	0.30
F	0.01	0.99



BNs: Burglary example

$P(B)$

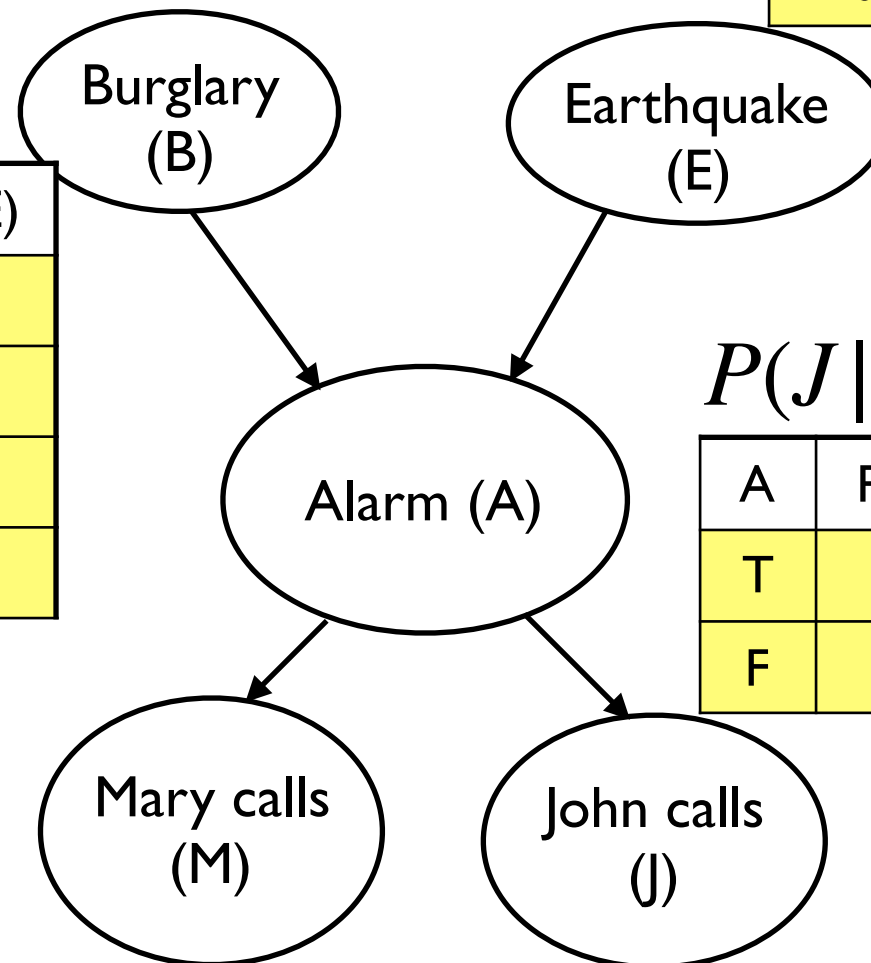
$P(B=T)$	$P(B=F)$
0.001	0.999

$P(E)$

$P(E=T)$	$P(E=F)$
0.002	0.998

$P(A | B, E)$

B	E	$P(A=T B,E)$	$P(A=F B,E)$
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999



$P(J | A)$

A	$P(J=T A)$	$P(J=F A)$
T	0.90	0.10
F	0.05	0.95

$P(M | A)$

A	$P(M=T A)$	$P(M=F A)$
T	0.70	0.30
F	0.01	0.99

$$P(B = F)P(E = F)P(A = T | B = F, E = F)P(M = T | A = T)P(J = T | A = T) \\ = 0.999 \times 0.998 \times 0.0001 \times 0.70 \times 0.90 = 0.000062$$

BN inference: Burglary example

Our BN can answer any probabilistic query that can be answered by processing the joint!

Example:

I'm at work.

Neighbour John calls.

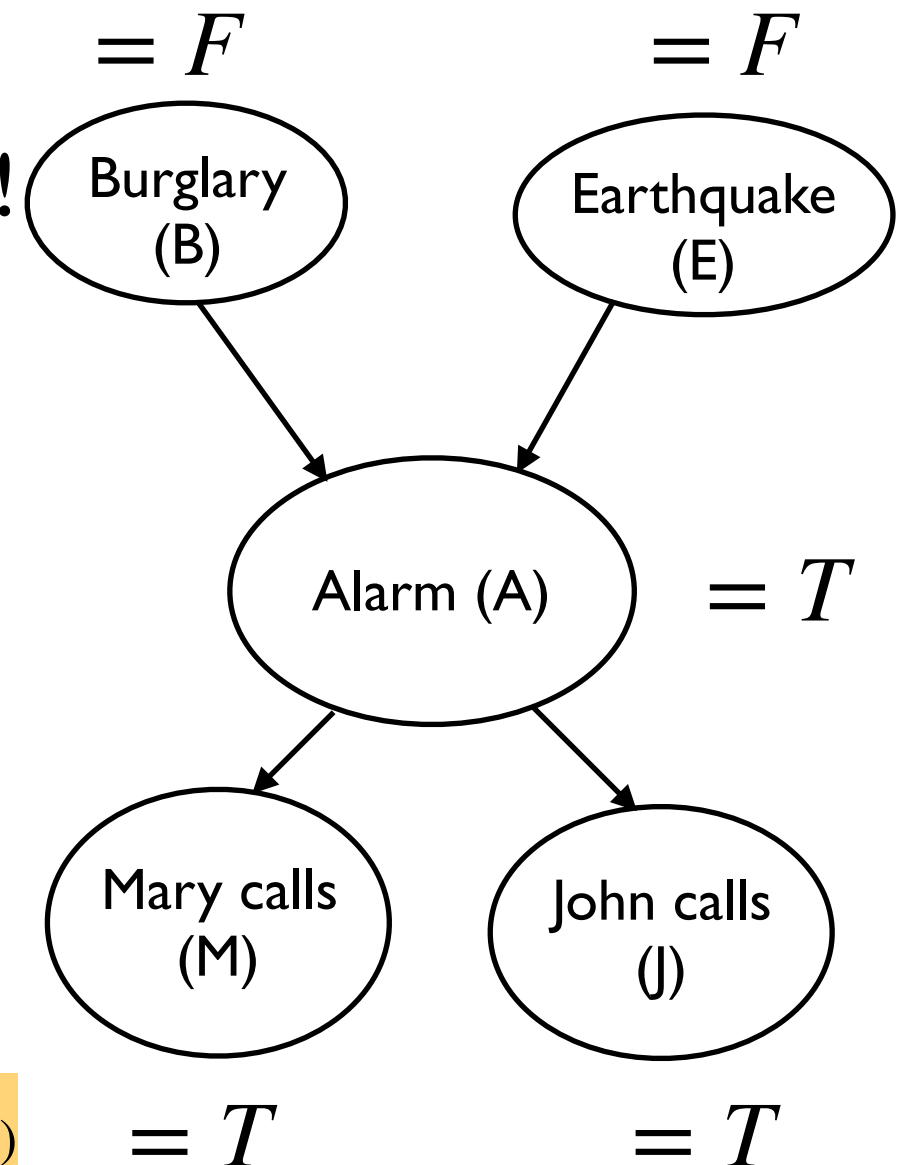
Neighbour Mary calls.

No news of any earthquakes.

Is there a burglar?

$$P(B = F, E = F, A = T, M = T, J = T)?$$

$$\begin{aligned} &P(B = F)P(E = F)P(A = T|B = F, E = F)P(M = T|A = T)P(J = T|A = T) \\ &= 0.999 \times 0.998 \times 0.0001 \times 0.70 \times 0.90 = 0.00062 \end{aligned}$$



BN: compactness

$$P(B)$$

P(B=T)	P(B=F)
0.001	0.999

$$P(E)$$

P(E=T)	P(E=F)
0.002	0.998

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

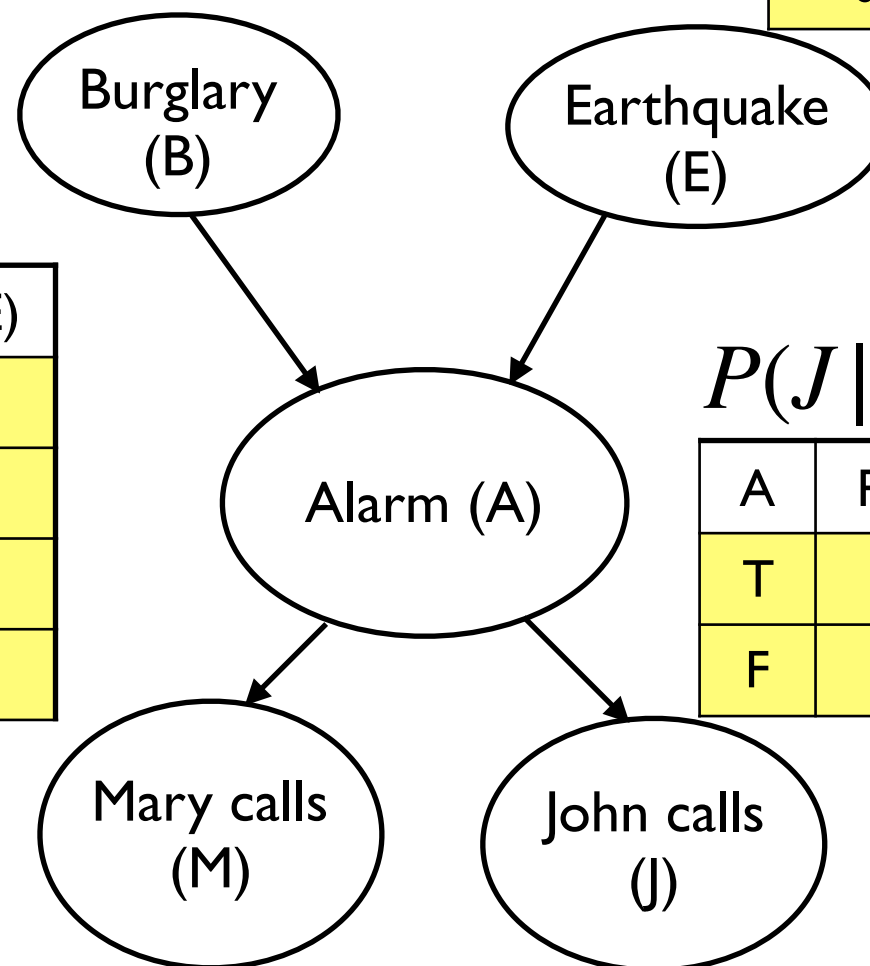
B	E	P(A=T B,E)	P(A=F B,E)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

$$P(J | A)$$

A	P(J=T A)	P(J=F A)
T	0.90	0.10
F	0.05	0.95

$$P(M | A)$$

A	P(M=T A)	P(M=F A)
T	0.70	0.30
F	0.01	0.99



BN: compactness

$P(B)$

$P(B=T)$	$P(B=F)$
0.001	0.999

$P(E)$

$P(E=T)$	$P(E=F)$
0.002	0.998

$P(A | B, E)$

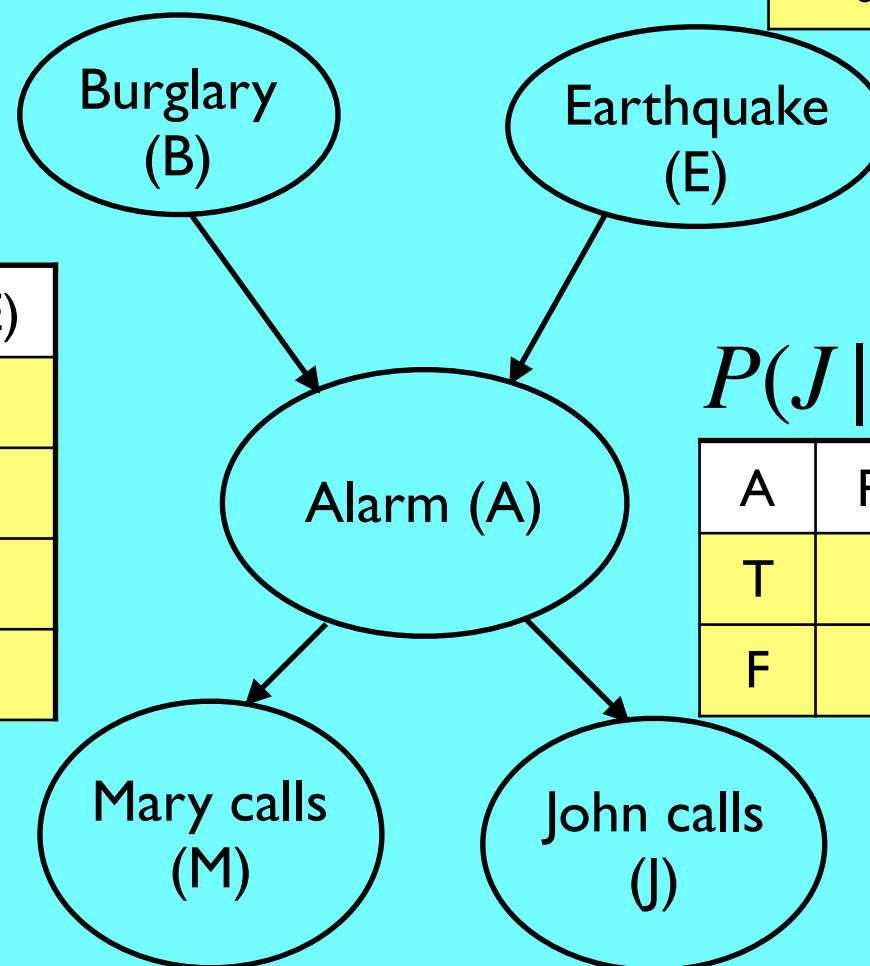
B	E	$P(A=T B,E)$	$P(A=F B,E)$
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

$P(J | A)$

A	$P(J=T A)$	$P(J=F A)$
T	0.90	0.10
F	0.05	0.95

$P(M | A)$

A	$P(M=T A)$	$P(M=F A)$
T	0.70	0.30
F	0.01	0.99



How many values do we need to store with BN?

A. 5 C. 20

B. 10 D. 16



Realistic BN: Liver diagnoses

Source: Onisko et al. 1999

Nodes: ~60

JPD: 2^{60} entries

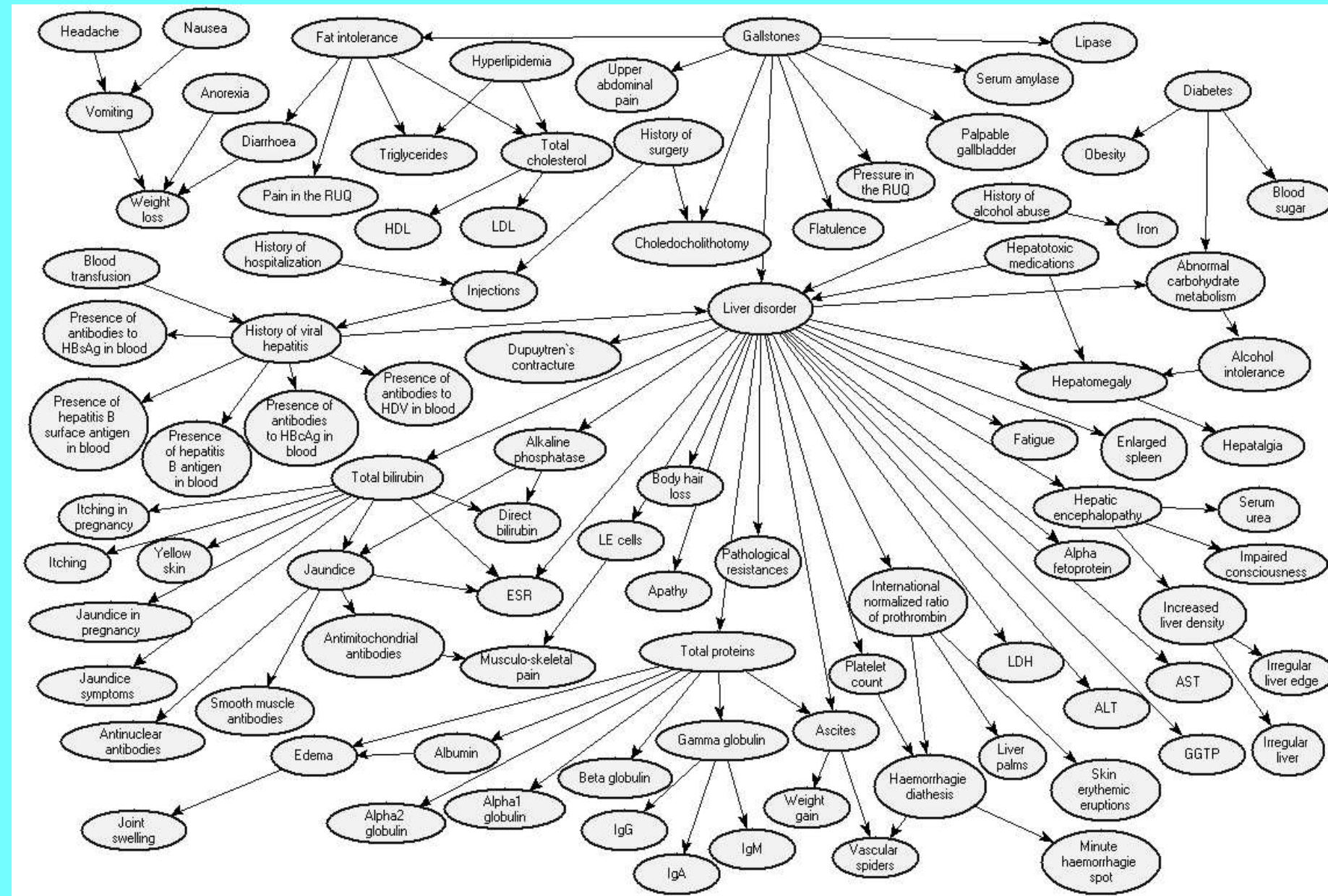
max 4 parents per node.

How many probabilities
do we need to store
with BNs?

A. 2^{60}

B. 15×2^6 ✓

C. 2^4



Interim summary

- In a Belief network, the JPD of the variables involved is defined as the product of the local conditional distributions
- $$P(X_1, \dots, X_n) = \prod_i P(X_i | X_1, \dots, X_{i-1}) = \prod_i P(X_i | \text{Parents}(X_i))$$
- Any entry (probability) in the JPD can be computed given the CPTs in the network.
- Once we know the JPD, we can answer any query about any subset of the variables (using inference by enumeration).
- A Belief network allows one to answer any query on any subset of the variables using CPTs.

Where do the conditional probabilities come from?

- The joint distribution is **not** normally the starting point
 - We would have to define exponentially many numbers
- First define the Bayesian network structure
 - Either by domain knowledge or by machine learning algorithms (see CPSC 540) (Typically based on local search)
- Then fill in the conditional probability tables
 - Either by domain knowledge or by machine learning algorithms (see CPSC 340, CPSC 422)
 - Based on statistics over the observed data

Independence assumption in BNs

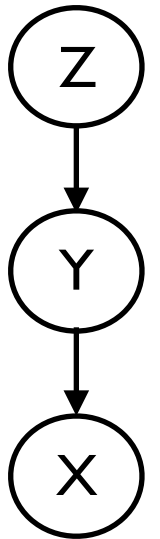
The independence of a belief network, is that each variable is independent of all of the variables that are **not descendants** of the variable (its non-descendants) **given the variable's parents**.

Independencies in BNs

Given Y , does learning the value of Z tell us nothing new about X ?
I.e., is $P(X|Y, Z)$ equal to $P(X | Y)$?

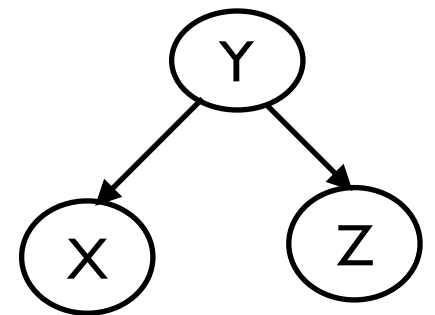
Yes. Since we know the value of all of X 's parents (namely, Y), and Z is not a descendant of X , X is conditionally independent of Z .

Also, since independence is symmetric, $P(Z|Y, X) = P(Z|Y)$.



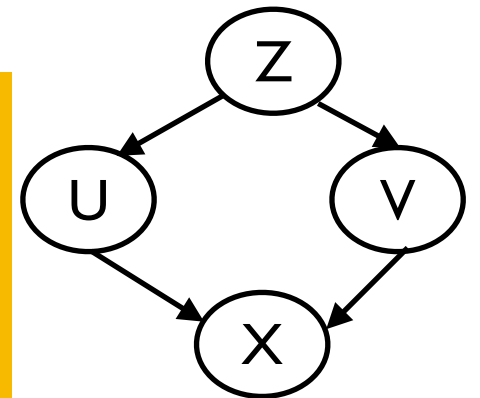
Is X conditionally independent of Z given Y ?

Yes. All X 's parents are given and Z is not a descendent of X

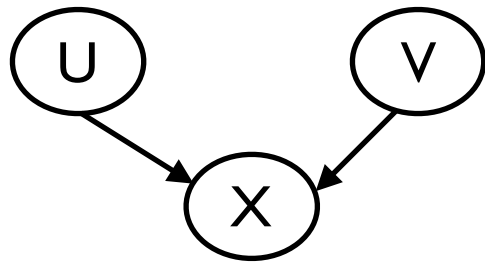


Is Z conditionally independent of X given U ? No.

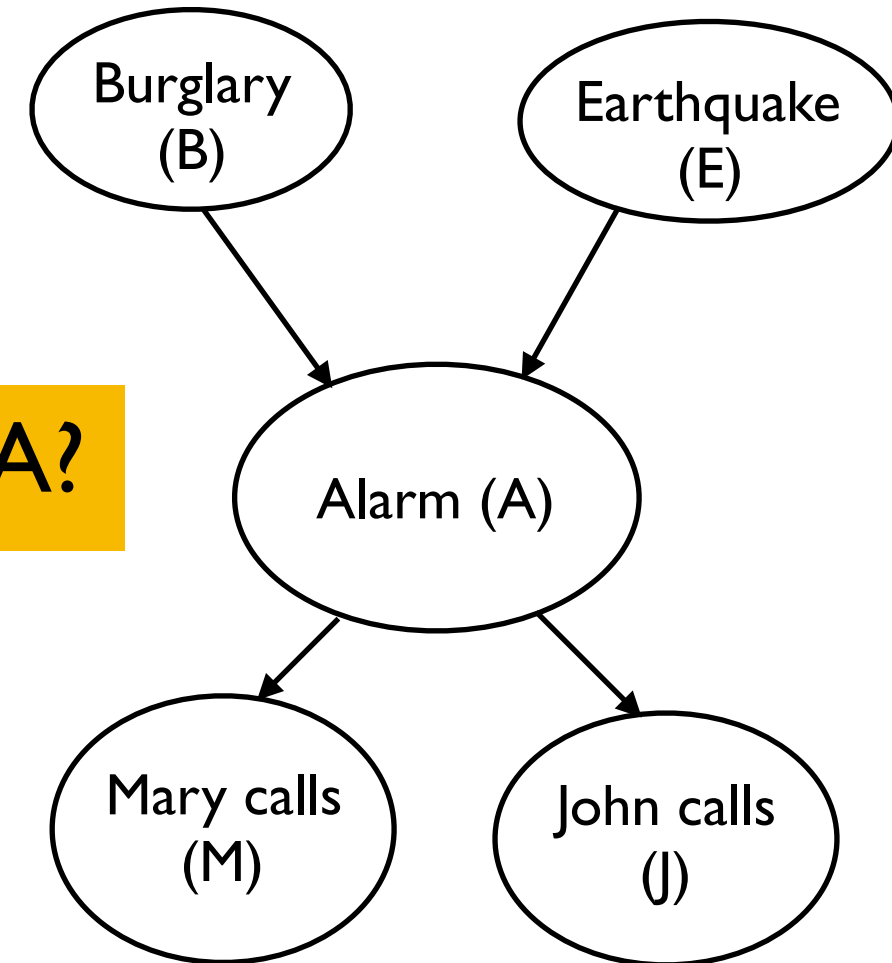
Is Z conditionally independent of X given U and V ? Yes.



Independencies in BNs



Is B conditionally independent of E given A?



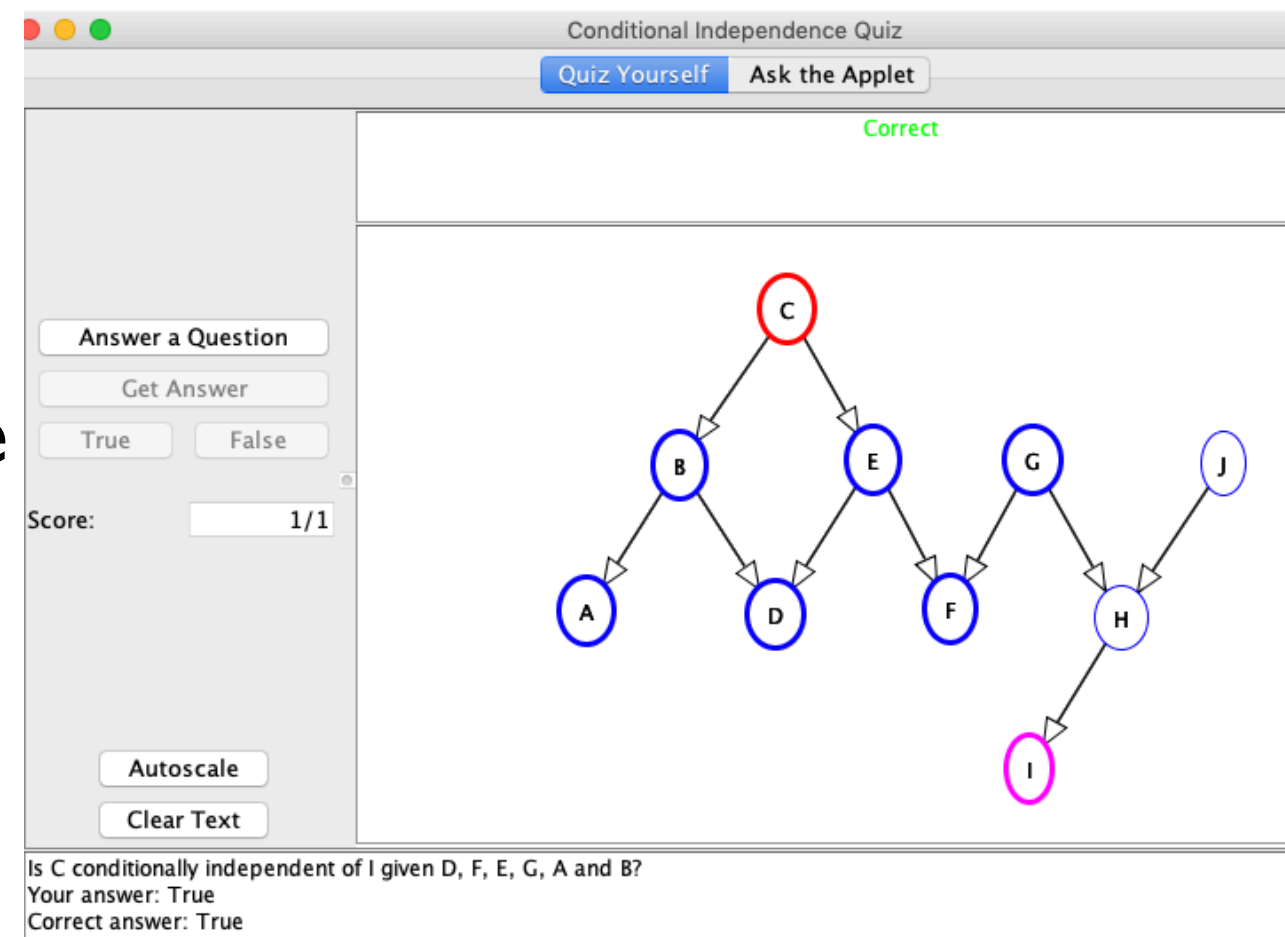
ASIDE: *d*-separation algorithm to determine whether two variables in Bayes net are conditional independent or not.

ASIDE: Independence in BNs



The independence of a belief network, is that each variable is independent of all of the variables that are not descendants of the variable (its non-descendants) given the variable's parents.

1. Download Alspace Belief and Decision networks applet
2. Load “Conditional Independence Quiz” sample problem
3. Go to “Independence Quiz” tab and quiz yourself.

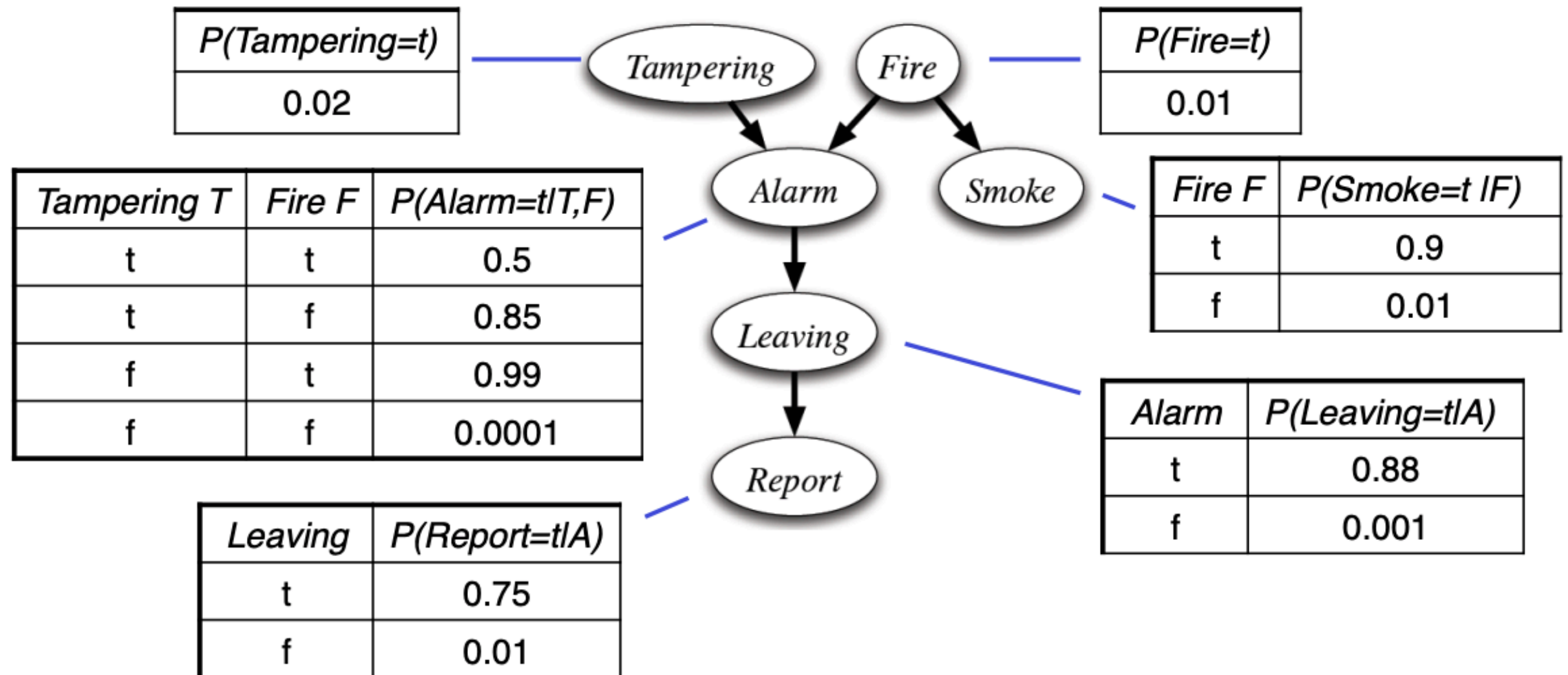


BN for report of leaving example (textbook)

- You want to diagnose whether there is a fire in a building
- You receive a noisy report about whether everyone is leaving the building
- If everyone is leaving, this may have been caused by a fire alarm
- If there is a fire alarm, it may have been caused by a fire or by tampering
- If there is a fire, there may be smoke



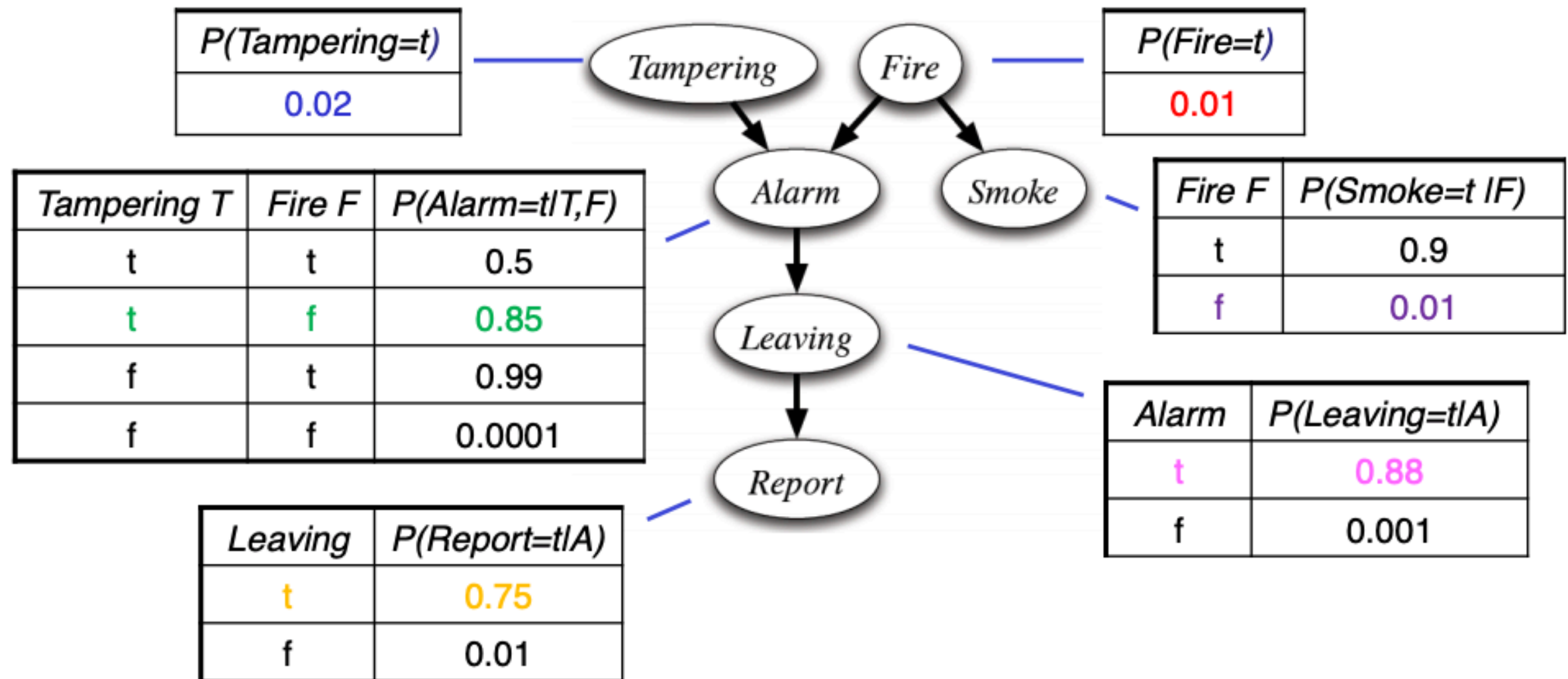
Calculate probability: Pair-share



$P(\text{Tampering} = t, \text{Fire} = f, \text{Alarm} = t, \text{Smoke} = f, \text{Leaving} = t, \text{Report} = t)$

= ?

Calculate probability: Pair-share



$$P(\text{Tampering} = t, \text{Fire} = f, \text{Alarm} = t, \text{Smoke} = f, \text{Leaving} = t, \text{Report} = t)$$

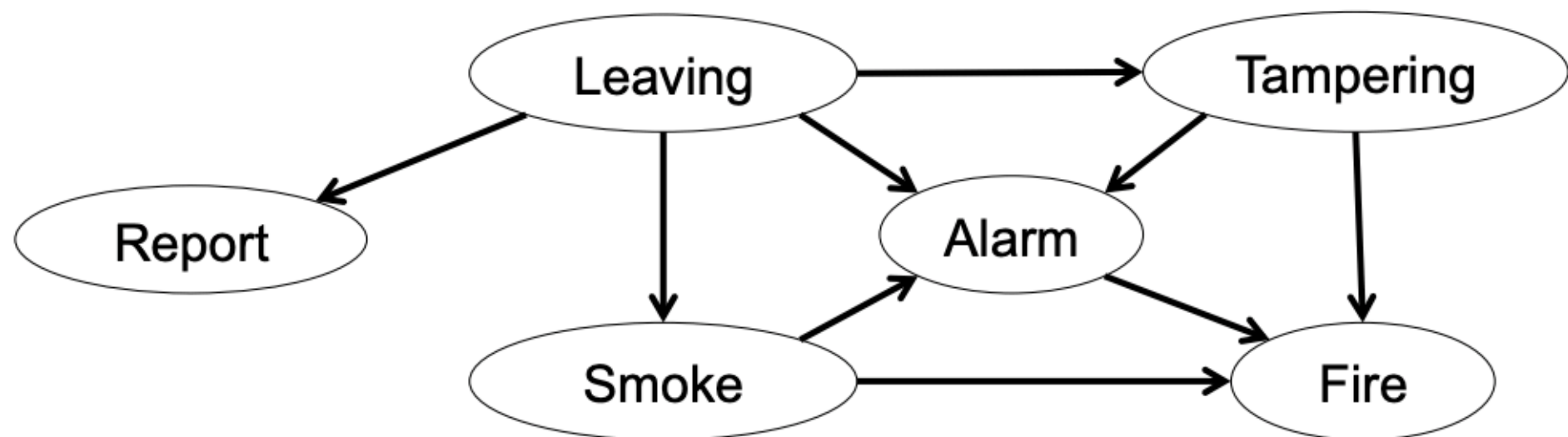
$$= P(\text{Tampering} = t) \times P(\text{Fire} = t) \times P(\text{Alarm} = t | \text{tampering} = T, \text{Fire} = f) \times P(\text{Smoke} = f | \text{Fire} = f) \times P(\text{Leaving} = t | \text{Alarm} = t)$$

$$= 0.02 \times (1 - 0.01) \times 0.85 \times (1 - 0.01) \times 0.88 \times 0.75 = 0.126$$

Variable ordering

What happens if we use different ordering ($n!$ possible orderings)
(Important for assignment 4, question 4)

Say, we use the following order: Leaving; Tampering; Report; Smoke;
Alarm; Fire.



We end up with a completely different network structure!
Which of the two structures is better (think computationally)?
The causal structure typically leads to the most compact network
Compactness typically enables more efficient reasoning

Are there wrong network structures?

Some variable orderings yield more compact, some less compact structures. Compact ones are better.

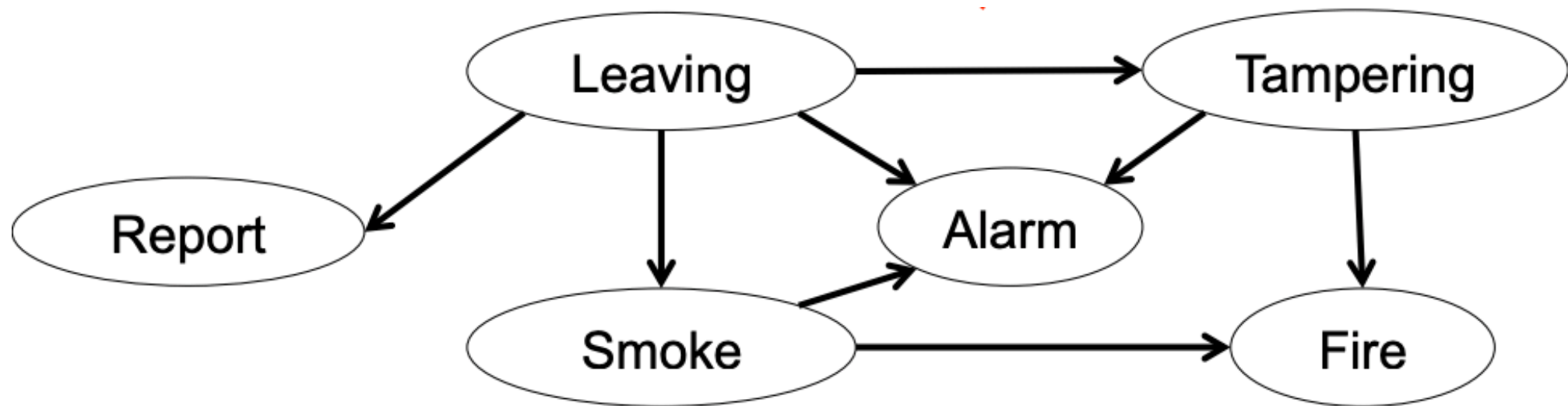
But all representations resulting from this process are correct.

One extreme: the fully connected network is always correct but rarely the best choice.

Are there wrong network structures?

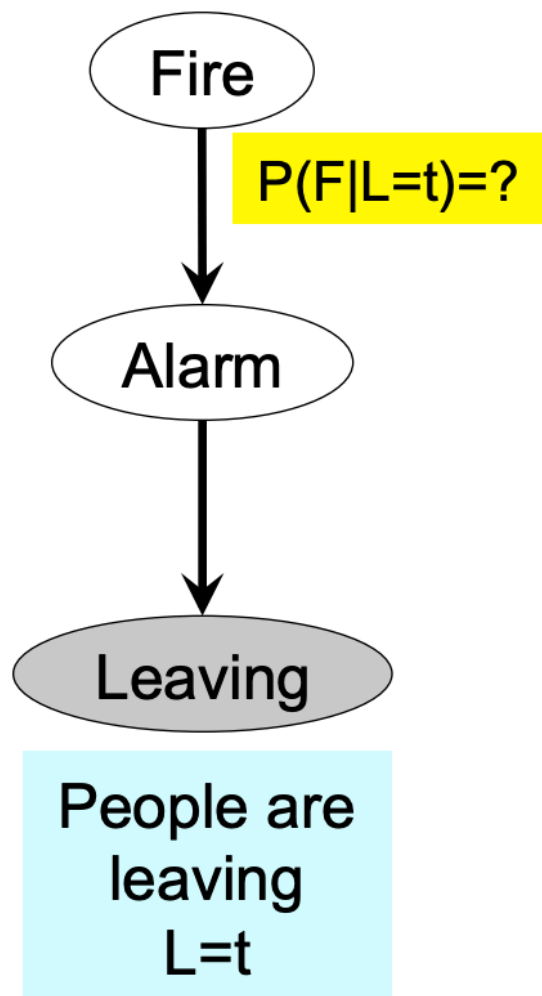
How can a network structure be wrong?

- If it misses directed edges that are required
- E.g., an edge is missing below: $\text{Fire} \not\perp\!\!\!\perp \text{Alarm} \mid \{\text{Tampering, Smoke}\}$

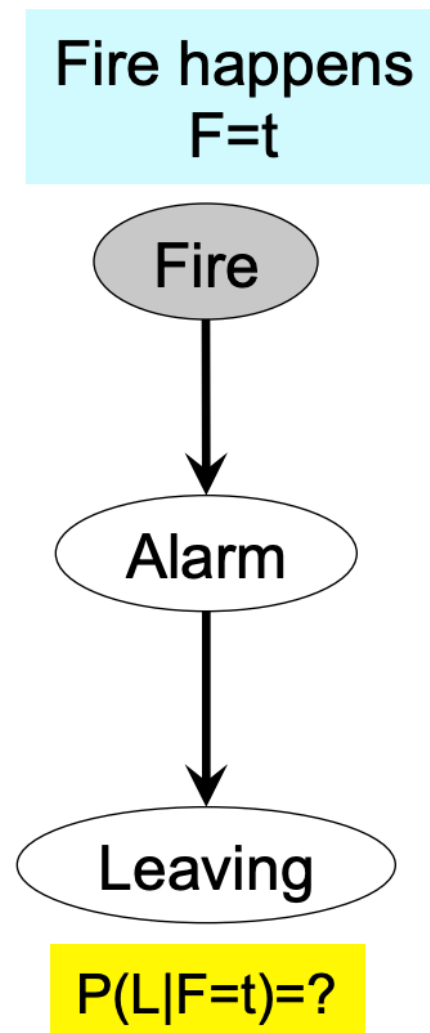


BNs: Types of inference

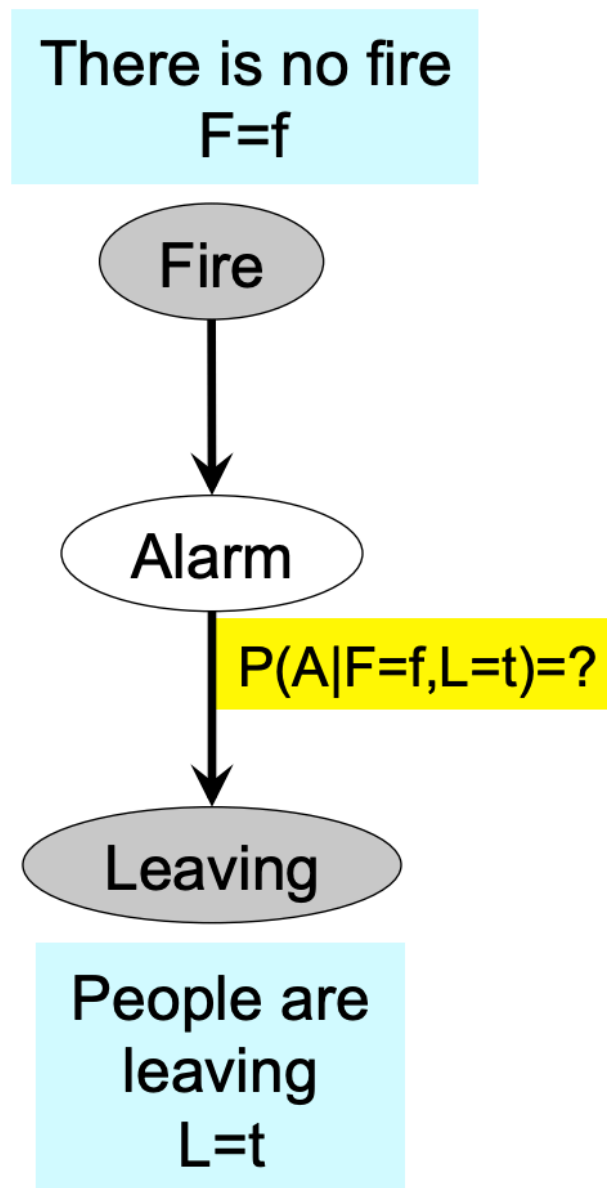
Diagnostic



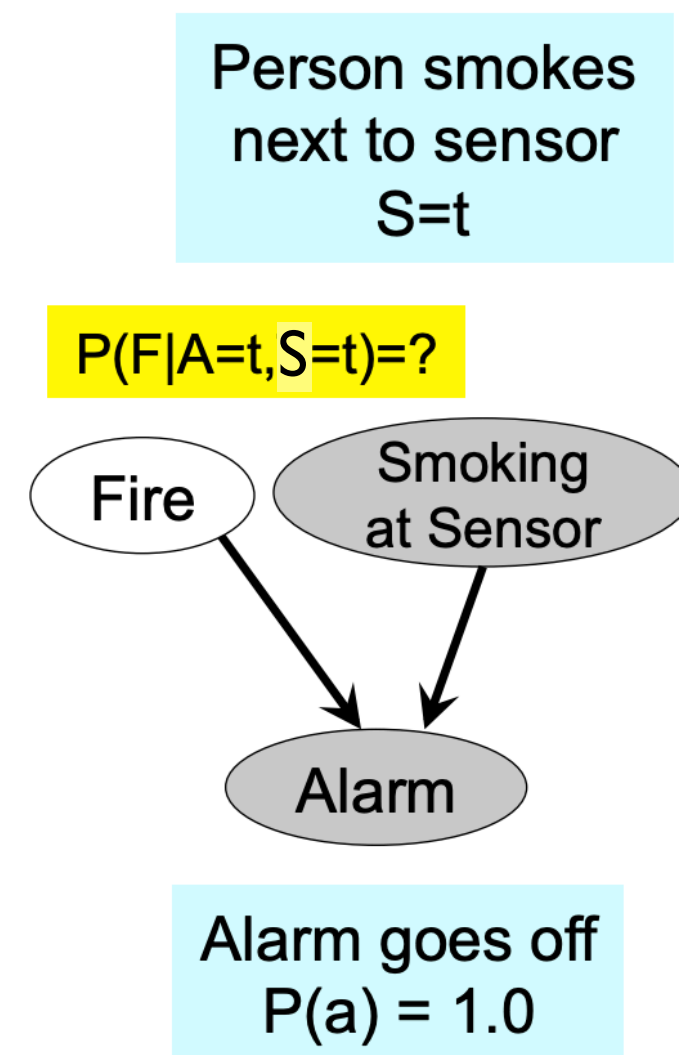
Predictive



Mixed



Intercausal



We will use the same reasoning procedure for all of these types

Inference in Bayesian networks

Given:

A Bayesian Network BN

Observations of a subset of its variables $E : E = e$

A subset of its variables Y that is queried

Compute: The conditional probability $P(Y | E = e)$

How: Run **variable elimination algorithm**

Note: We can already do all this with Inference by Enumeration. The BN represents the JPD. Could just multiply out the BN to get full JPD and then do Inference by Enumeration BUT that's **extremely inefficient**; it does not scale.

Inference in general BNs

- The variable elimination algorithm (VE)
- The VE algorithm manipulates conditional probabilities in the form of “factors”. So first we have to introduce **factors** and the operations we can perform on them.

Factors

A factor is a function from a tuple of random variables to the real numbers R . We write a factor on variables X_1, \dots, X_j as $f(X_1, \dots, X_j)$

A **factor** can denote:

- One distribution
- One *partial* distribution
- Several distributions
- Several *partial* distributions over the given tuple of variables

Factors

A factor is a function from a tuple of random variables to the real numbers R . We write a factor on variables X_1, \dots, X_j as $f(X_1, \dots, X_j)$

Operations on factors

- Assigning variables
- Summing out variables
- Multiplication of factors

Factors and operations on them

A factor is a function from a tuple of random variables to the real numbers R .

Operation I: assigning a variable in a factor

E.g., $X=t$

Factor of Y,X,Z			
X	Y	Z	$f_1(X,Y,Z)$
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
f	t	t	0.4
f	t	f	0.6
f	f	t	0.3
f	f	f	0.7

$f_1(X,Y,Z)_{X=t} = f_2(Y,Z)$

Y	Z	$f_2(Y,Z)$
t	t	0.1
t	f	0.9
f	t	0.2
f	f	0.8

Factor of Y,Z

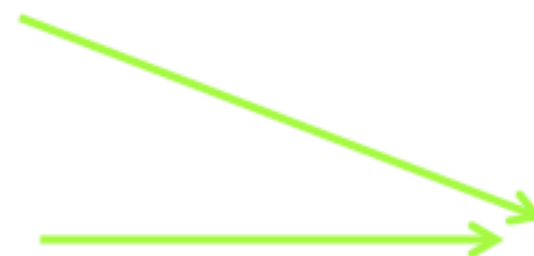
Factors and operations on them

A factor is a function from a tuple of random variables to the real numbers R .

Operation 2: marginalize out a variable from a factor

B	A	C	$f_3(A,B,C)$
t	t	t	0.03
t	t	f	0.07
f	t	t	0.54
f	t	f	0.36
t	f	t	0.06
t	f	f	0.14
f	f	t	0.48
f	f	f	0.32

$$\sum_B f_3(A,B,C) = f_4(A,C)$$

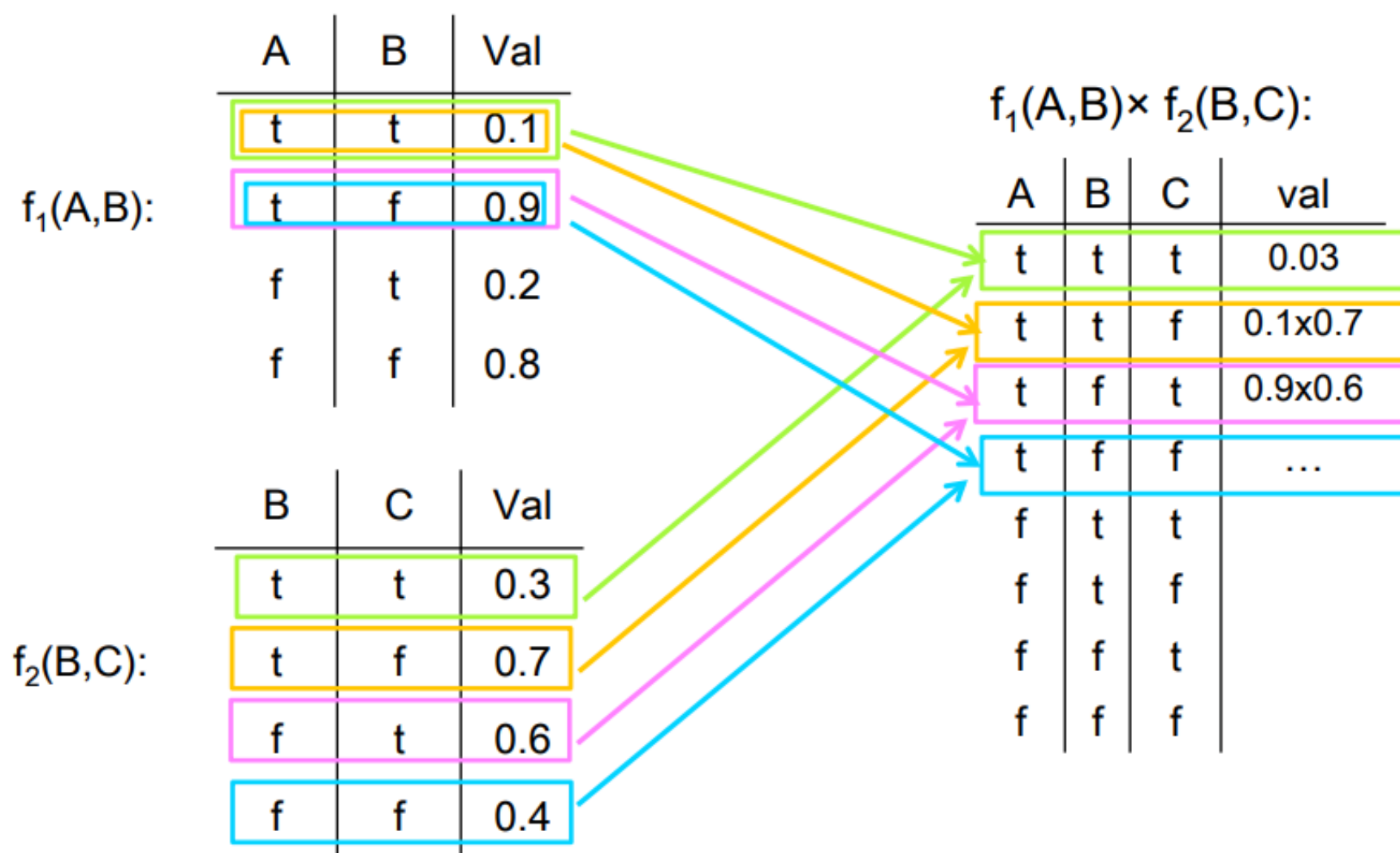


A	C	$f_4(A,C)$
t	t	0.57
t	f	0.43
f	t	0.54
f	f	0.46

Factors and operations on them

A factor is a function from a tuple of random variables to the real numbers R .

Operation 3: Multiplying factors



$$f_1(A, B) \times f_2(B, C) \\ = f_3(A, B, C)$$

$$= f_1(A = a, B = b) \times \\ f_2(B = b, C = c) \\ = f_3(A = a, B = b, C = c)$$

Summary: Factors and operations on them

A **factor** is a function from a tuple of random variables to the real numbers R .

Operation 1: **assigning** a variable in a factor

$$\text{E.g., } f_2(Y, Z) = f_1(X, Y, Z)_{X=t}$$

Operation 2: **marginalize out** a variable from a factor

$$\text{E.g., } f_4(A, C) = \sum_B f_3(A, B, C)$$

Operation 3: **multiply** two factors

$$\text{E.g., } f_7(A, B, C) = f_5(A, B) \times f_6(B, C)$$

$$\text{E.g., } f_7(A = a, B = b, C = c) = f_5(A = a, B = b) \times f_6(B = b, C = c)$$

Revisit: Learning outcomes

From this lecture, students are expected to be able to:

- Build a belief network for a simple domain
- Compute the representational savings in terms of number of probabilities required
- Classify the types of inference: diagnostic, predictive, inter-causal, mixed
- Define factors and apply operations to factors, including assigning, summing out and multiplying factors

Coming up

8.4 Probabilistic Inference

8.5 Sequential Probability Models

