

# Bayesian Networks

## Chapter 14

Mausam

(Slides by UW-AI faculty & David Page)

# Burglars and Earthquakes

- You are at a “Done with the AI class” party.
- Neighbor John calls to say your home alarm has gone off (but neighbor Mary doesn't).
- Sometimes your alarm is set off by minor earthquakes.
- Question: Is your home being burglarized?
- 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

# Example

- Pearl lives in Los Angeles. It is a high-crime area. Pearl installed a burglar alarm. He asked his neighbors John & Mary to call him if they hear the alarm. This way he can come home if there is a burglary. Los Angeles is also earth-quake prone. Alarm goes off when there is an earth-quake.

Burglary  $\Rightarrow$  Alarm

Earth-Quake  $\Rightarrow$  Alarm

Alarm  $\Rightarrow$  John-calls

Alarm  $\Rightarrow$  Mary-calls

If there is a burglary, will Mary call?

Check  $KB \ \& \ E \models M$

If Mary didn't call, is it possible that Burglary occurred?

Check  $KB \ \& \ \sim M \text{ doesn't entail } \sim B$

# Example (Real)

- Pearl lives in Los Angeles. It is a high-crime area. Pearl installed a burglar alarm. He asked his neighbors John & Mary to call him if they hear the alarm. This way he can come home if there is a burglary. Los Angeles is also earthquake prone. Alarm goes off when there is an earthquake.
- Pearl lives in real world where (1) burglars can sometimes disable alarms (2) some earthquakes may be too slight to cause alarm (3) Even in Los Angeles, Burglaries are more likely than Earth Quakes (4) John and Mary both have their own lives and may not always call when the alarm goes off (5) Between John and Mary, John is more of a slacker than Mary.(6) John and Mary may call even without alarm going off

Burglary => Alarm

Earth-Quake => Alarm

Alarm => John-calls

Alarm => Mary-calls

If there is a burglary, will Mary call?

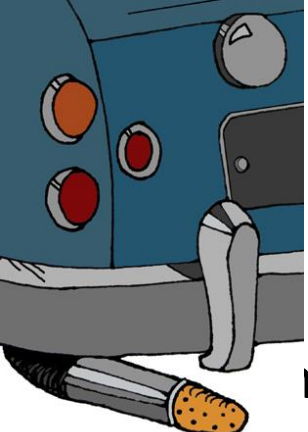
Check  $KB \ \& \ E \models M$

If Mary didn't call, is it possible that Burglary occurred?

Check  $KB \ \& \ \sim M \text{ doesn't entail } \sim B$

John already called. If Mary also calls, is it more likely that Burglary occurred?

You now also hear on the TV that there was an earthquake. Is Burglary more or less likely now?



# How do we handle Real Pearl?

## *•Potato in the tail-pipe*

naïve & Eager way:

- Model everything!
- E.g. Model exactly the conditions under which John will call
  - He shouldn't be listening to loud music, he hasn't gone on an errand, he didn't recently have a tiff with Pearl etc etc.

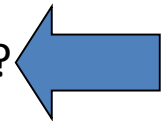
$A \ \& \ c1 \ \& \ c2 \ \& \ c3 \ \& \ ..cn \Rightarrow J$

(also the exceptions may have interactions

$c1 \ \& \ c5 \Rightarrow \sim c9$  )

- Ignorant (non-omniscient) and Lazy (non-omnipotent) way:

- Model the likelihood
- In 85% of the worlds where there was an alarm, John will actually call
- How do we do this?
  - Non-monotonic logics
  - “certainty factors”
  - “fuzzy logic”
  - “probability” theory?



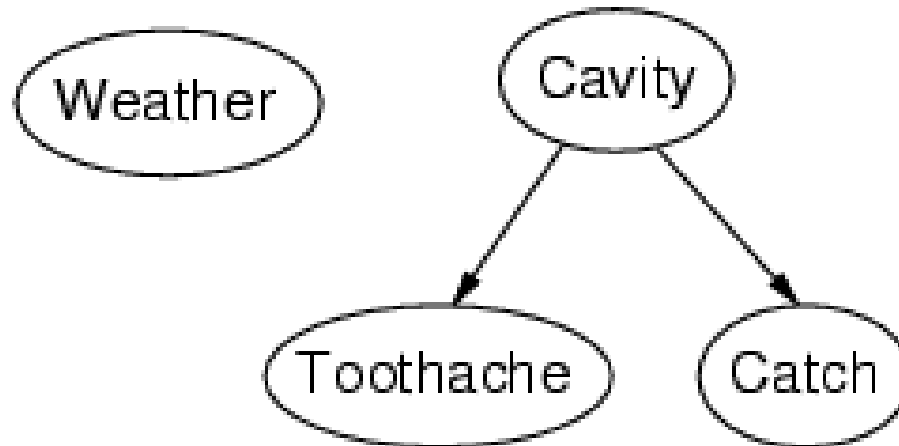
***Qualification and Ramification problems make this an infeasible enterprise***

# Bayes Nets

- In general, joint distribution  $P$  over set of variables  $(X_1 \times \dots \times X_n)$  requires exponential space for representation & inference
- BNs provide a graphical representation of *conditional independence* relations in  $P$ 
  - usually quite compact
  - requires assessment of fewer parameters, those being quite natural (e.g., causal)
  - efficient (usually) inference: query answering and belief update

# Back at the dentist's

Topology of network encodes conditional independence assertions:



Weather is independent of the other variables

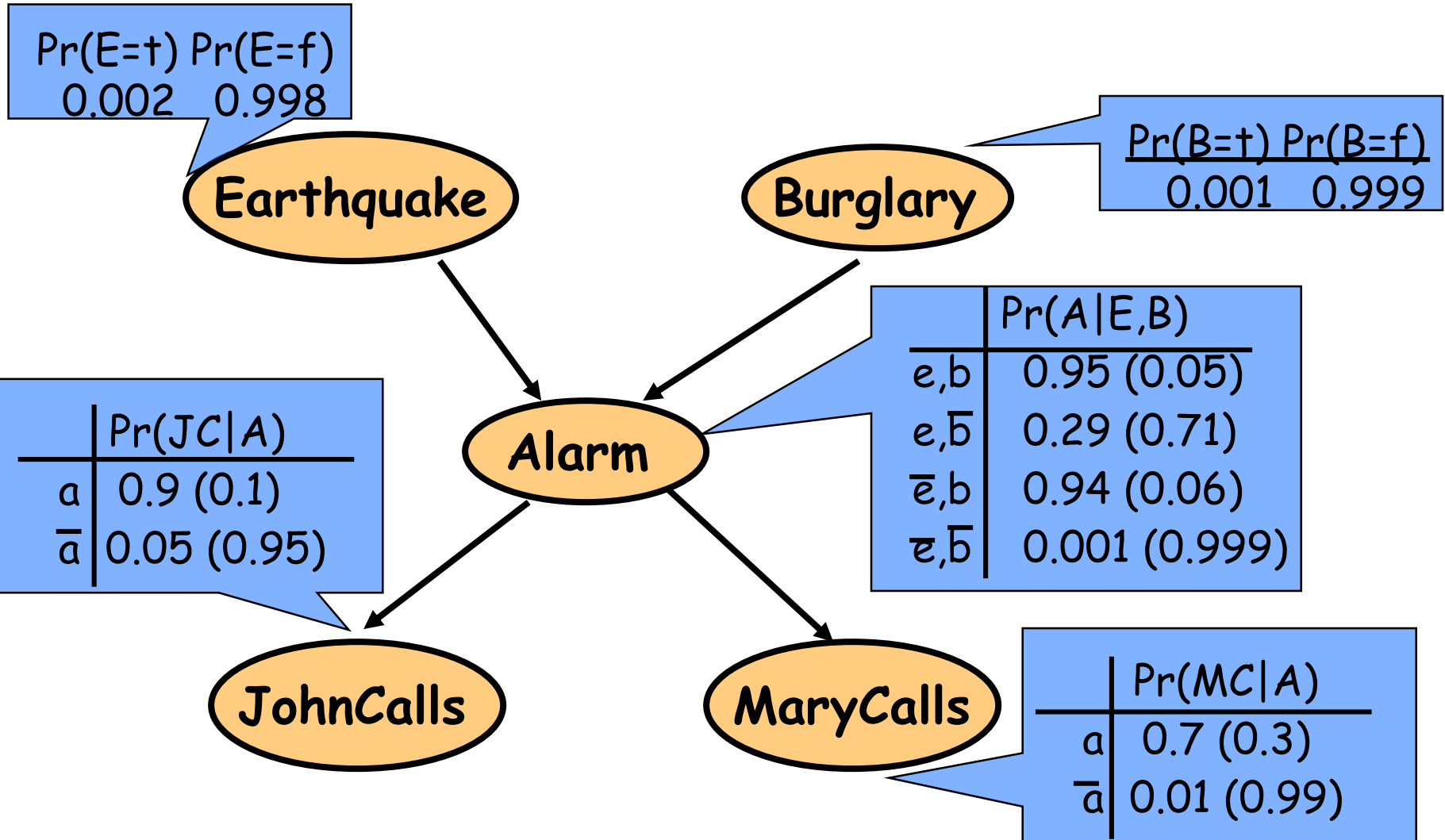
Toothache and Catch are conditionally independent of each other **given Cavity**

# Syntax

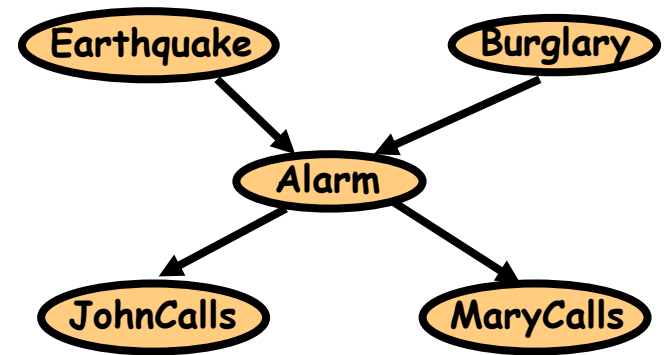
- a set of nodes, one per random variable
- a directed, acyclic graph (link  $\approx$  "directly influences")
- a conditional distribution for each node given its parents:  $P(X_i \mid \text{Parents}(X_i))$ 
  - For discrete variables, **conditional probability table (CPT)**= distribution over  $X_i$  for each combination of parent values



# Burglars and Earthquakes



# Earthquake Example (cont'd)



- If we know *Alarm*, no other evidence influences our degree of belief in *JohnCalls*

- $P(JC|MC,A,E,B) = P(JC|A)$

- also:  $P(MC|JC,A,E,B) = P(MC|A)$  and  $P(E|B) = P(E)$

- By the chain rule we have

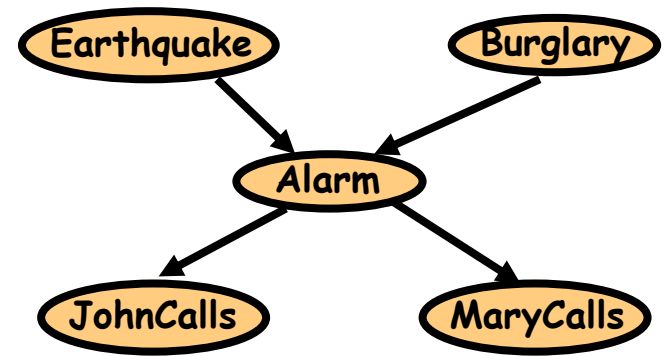
$$P(JC,MC,A,E,B) = P(JC|MC,A,E,B) \cdot P(MC|A,E,B) \cdot$$

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

$$= P(JC|A) \cdot P(MC|A) \cdot P(A|B,E) \cdot P(E) \cdot P(B)$$

- Full joint requires only 10 parameters (cf. 32)

# Earthquake Example (Global Semantics)



- We just proved

$$P(JC, MC, A, E, B) = P(JC|A) \cdot P(MC|A) \cdot P(A|B, E) \cdot P(E) \cdot P(B)$$

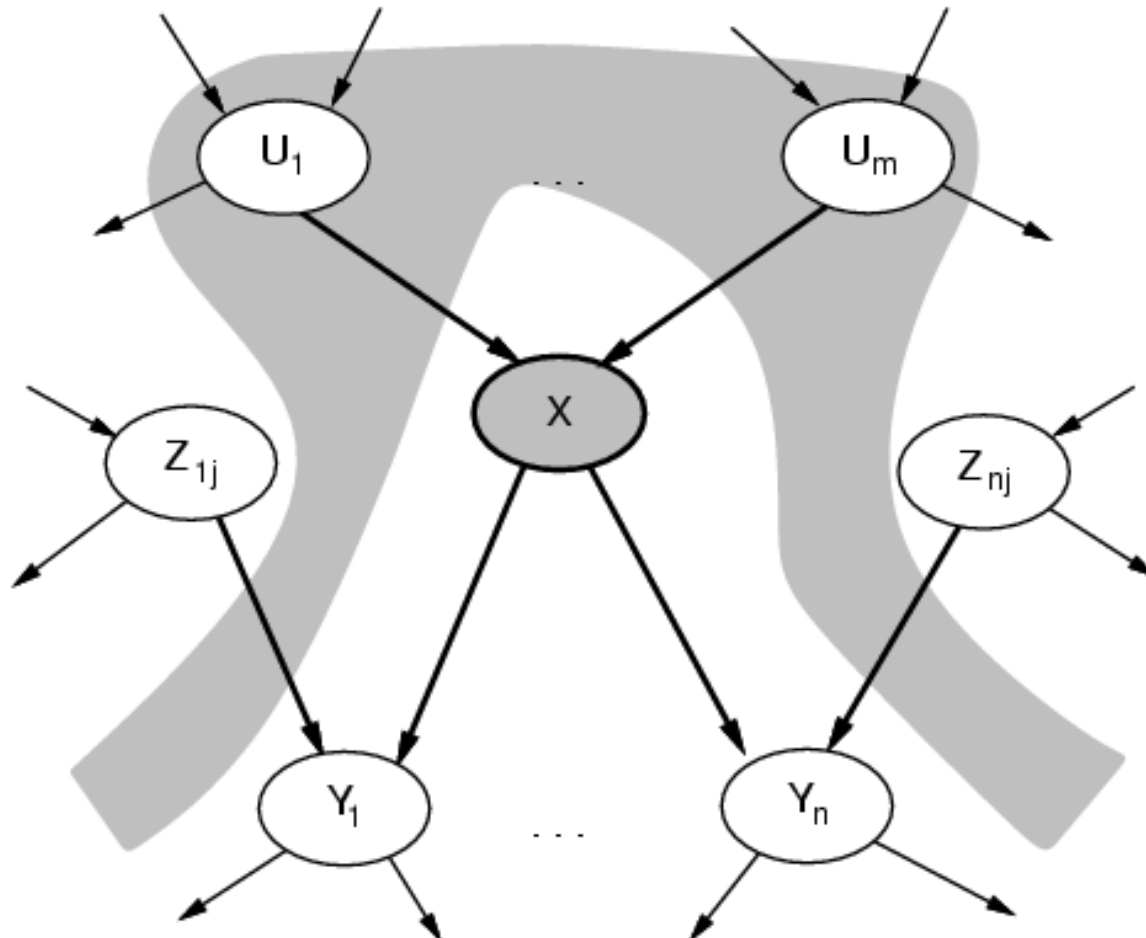
- In general full joint distribution of a Bayes net is defined as

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | Par(X_i))$$

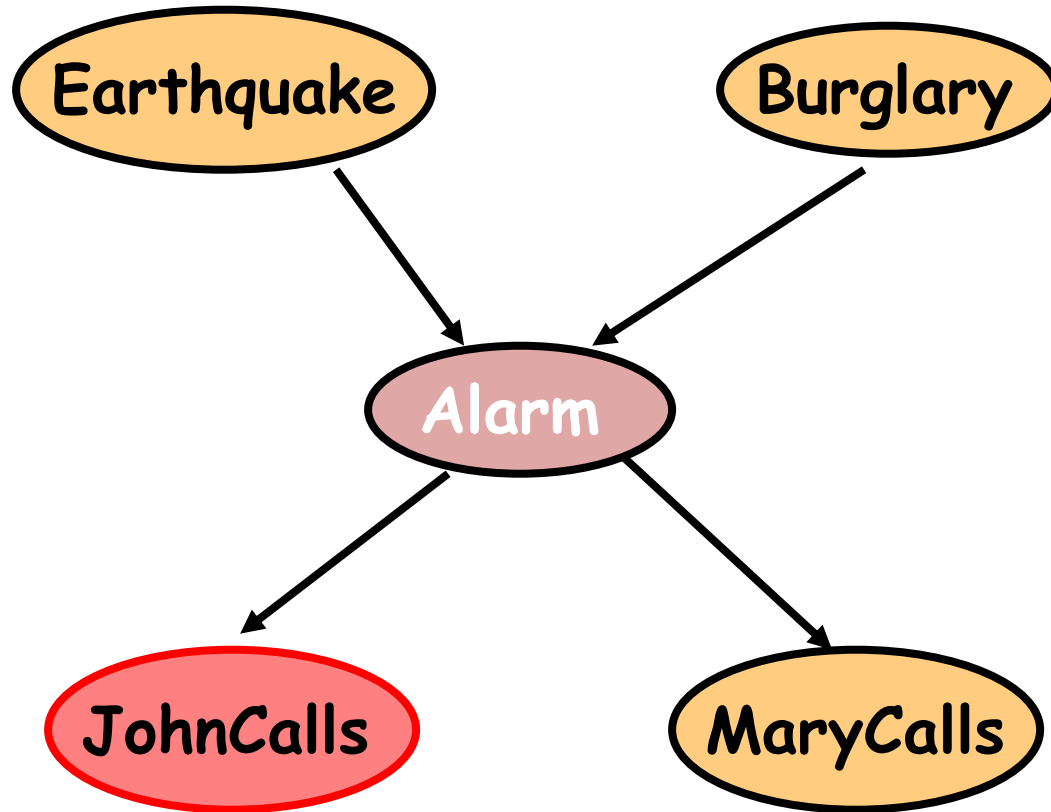
# BNs: Qualitative Structure

- Graphical structure of BN reflects conditional independence among variables
- Each variable  $X$  is a node in the DAG
- Edges denote *direct probabilistic influence*
  - usually interpreted *causally*
  - parents of  $X$  are denoted  $Par(X)$
- ***Local semantics:  $X$  is conditionally independent of all nondescendants given its parents***
  - Graphical test exists for more general independence
  - “Markov Blanket”

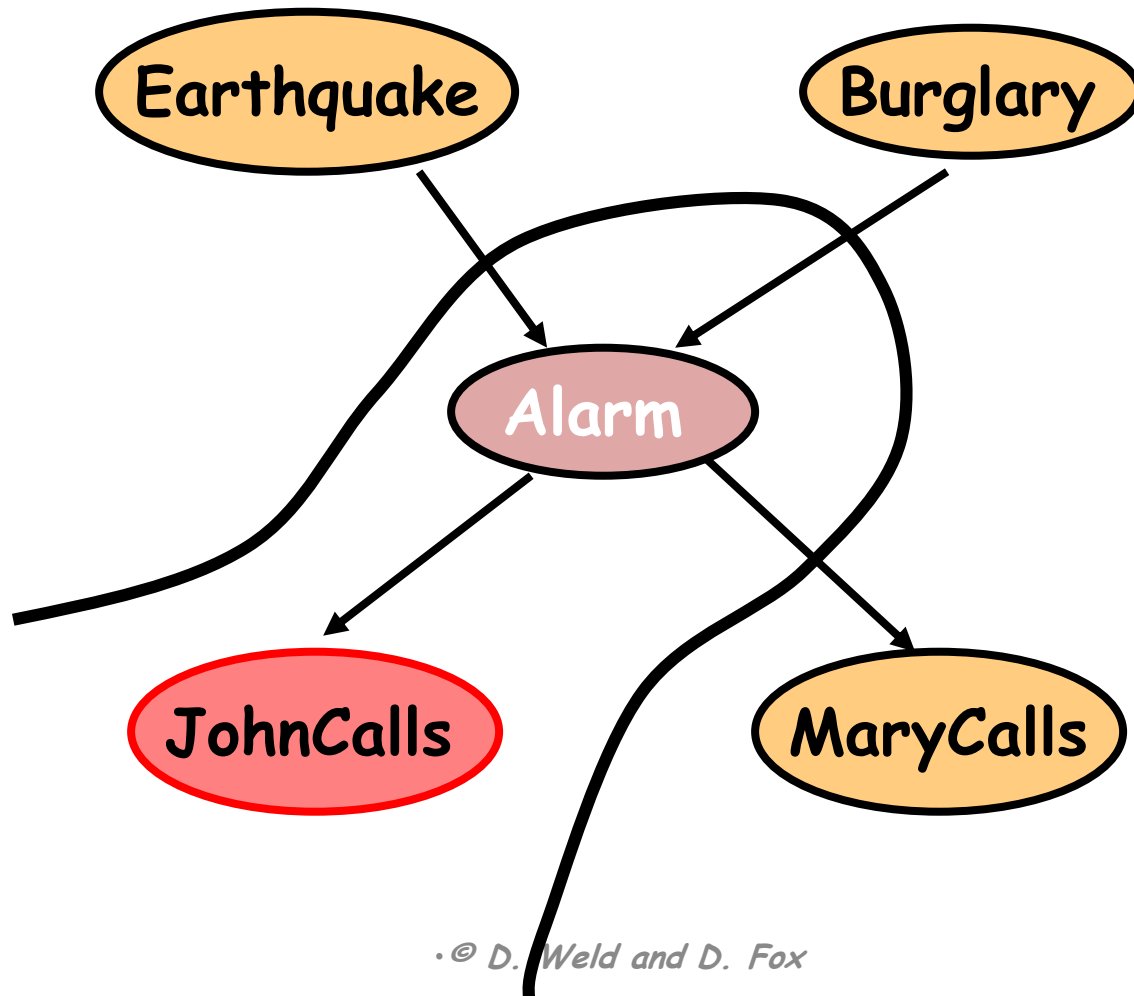
# Given Parents, $X$ is Independent of Non-Descendants



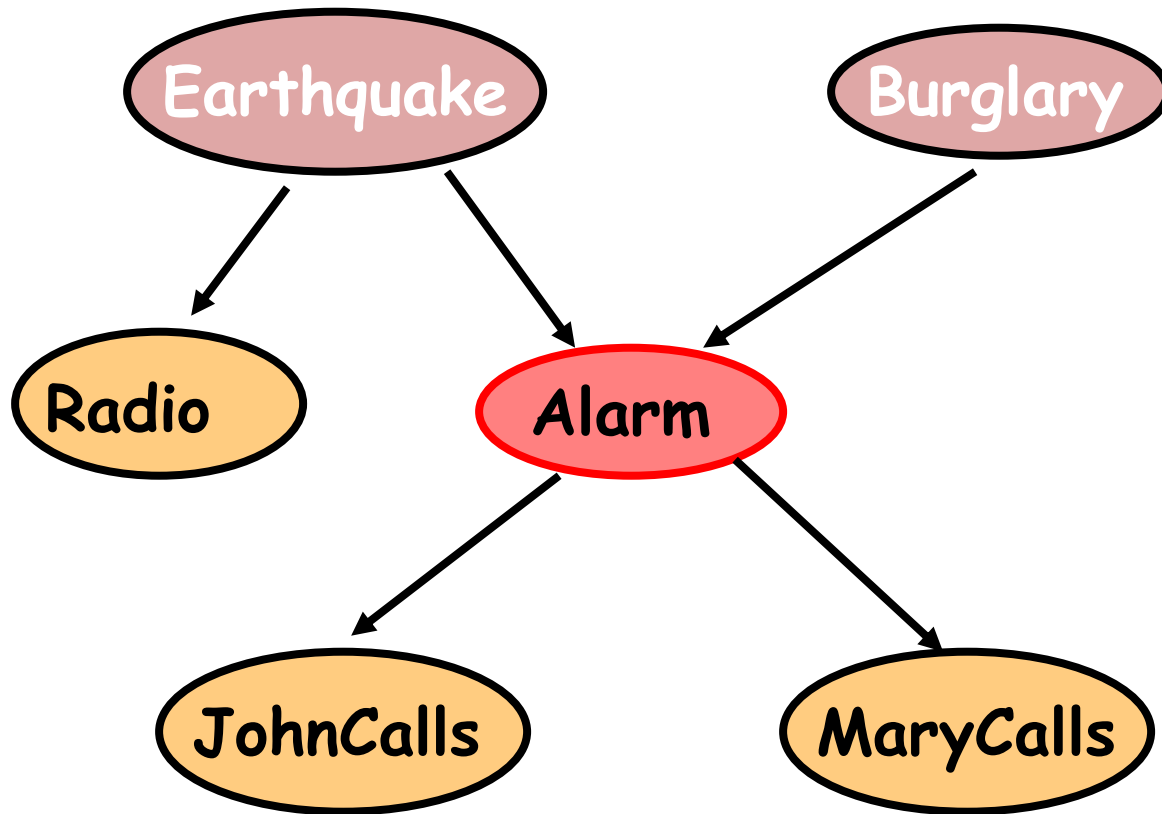
# Examples



# For Example

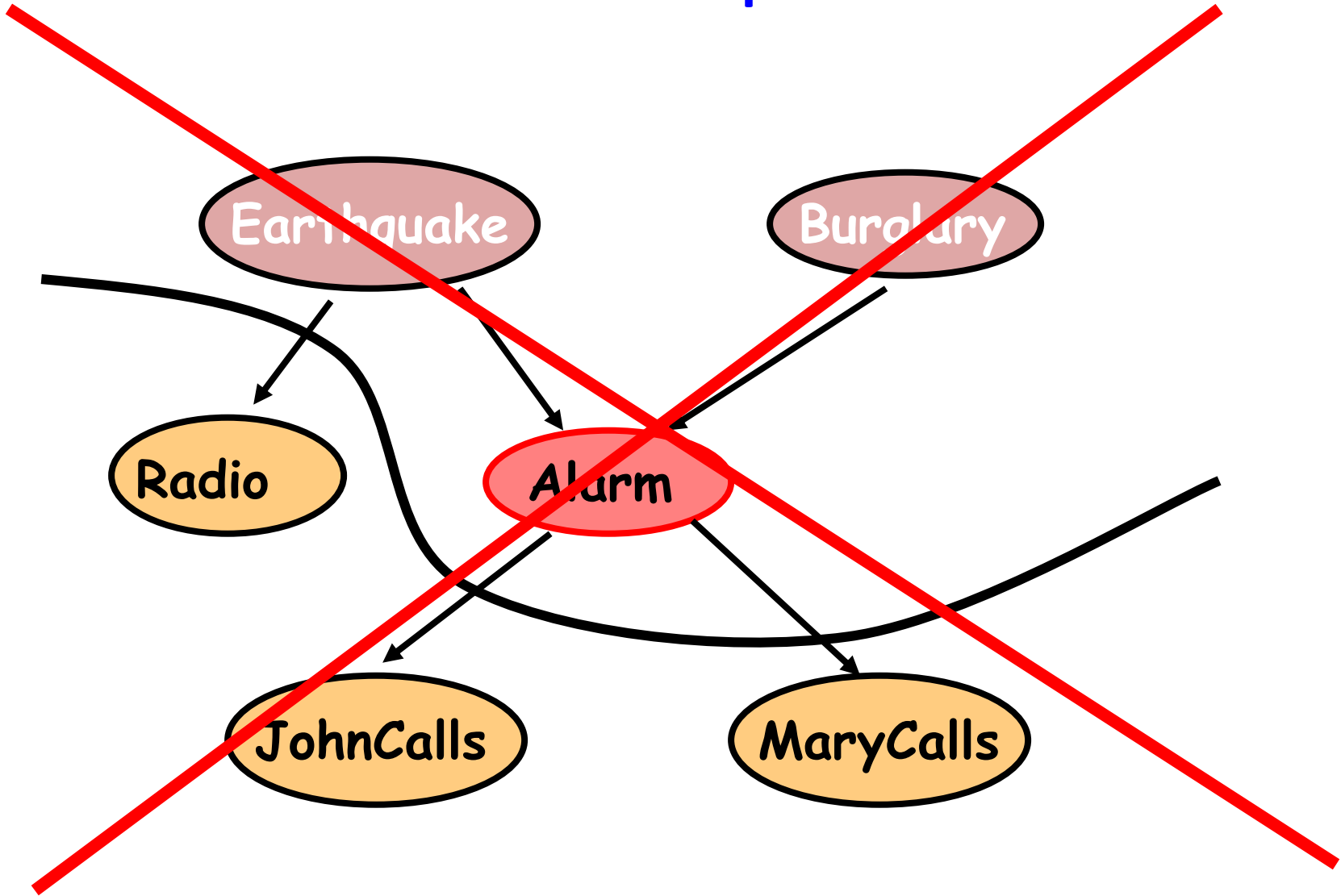


# For Example

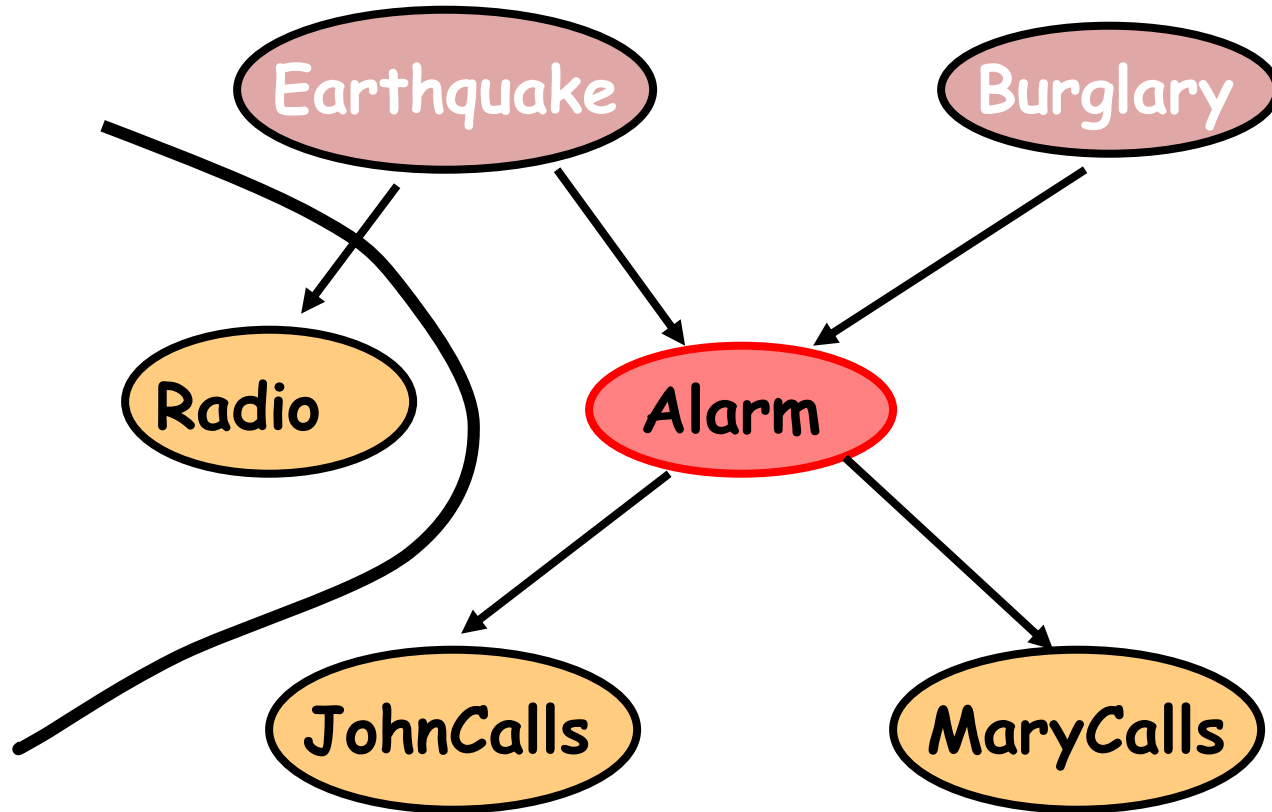




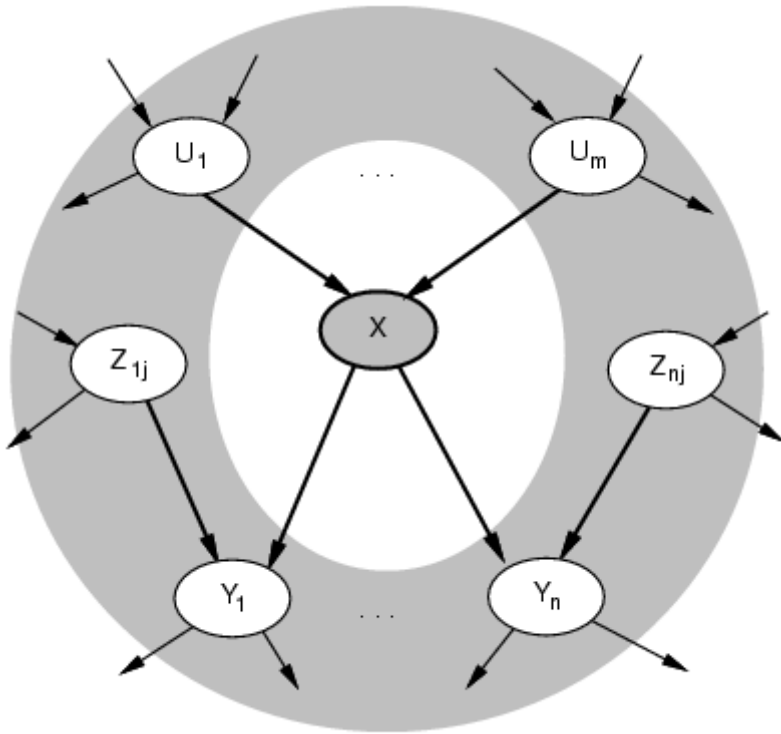
# For Example



# For Example

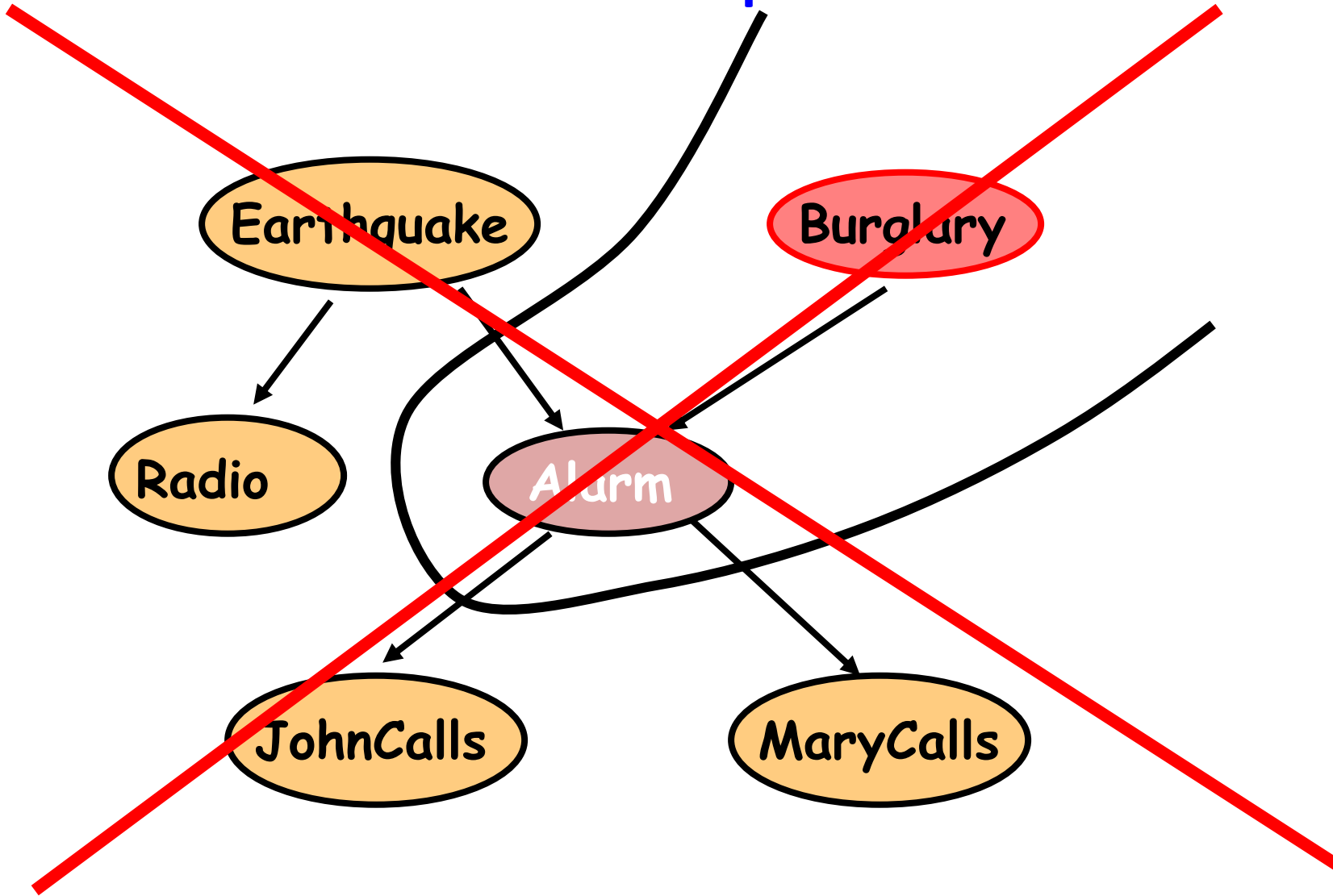


Given Markov Blanket, X is Independent of  
All Other Nodes

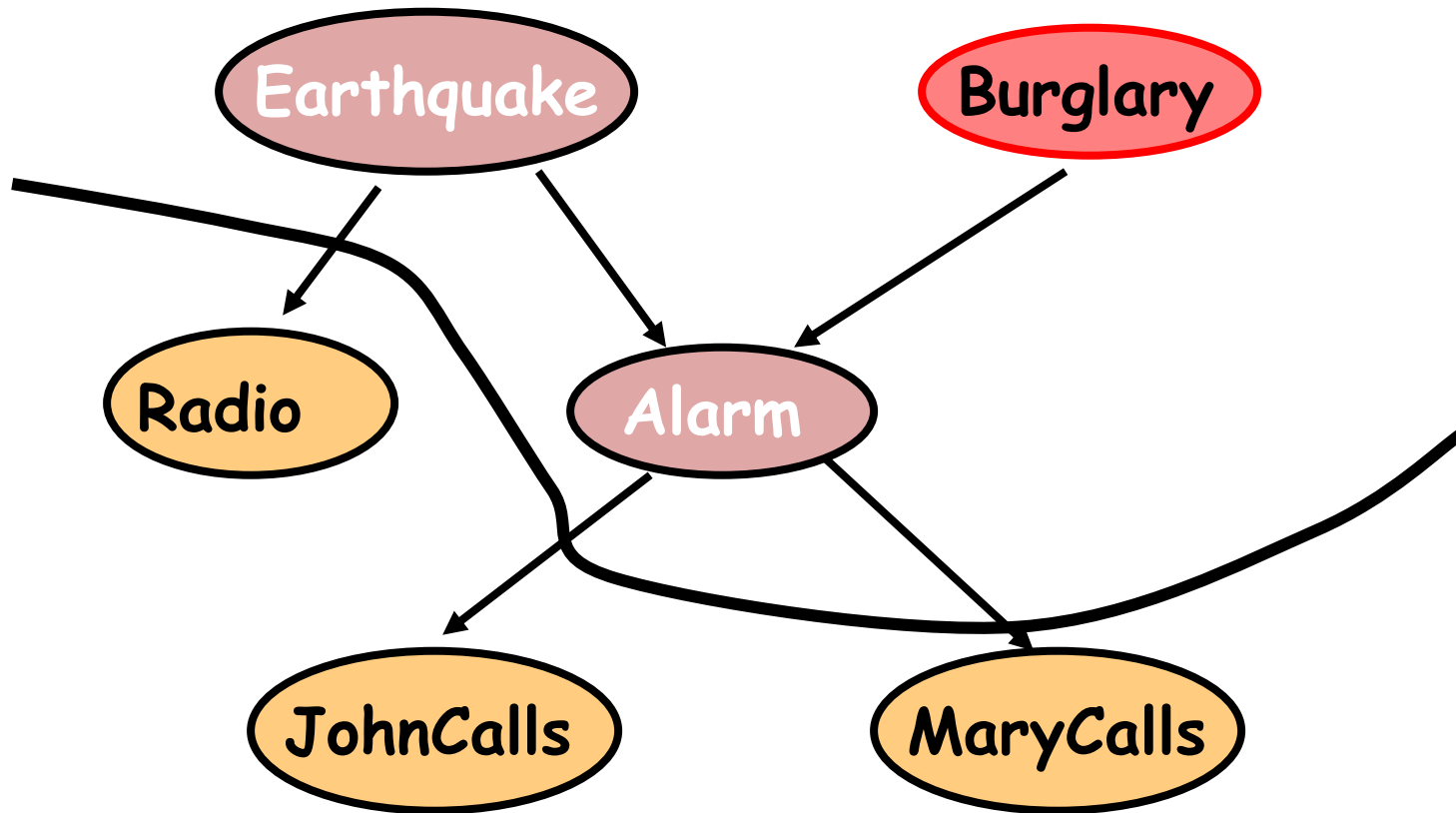


$$\text{MB}(X) = \text{Par}(X) \cup \text{Childs}(X) \cup \text{Par}(\text{Childs}(X))$$

# For Example



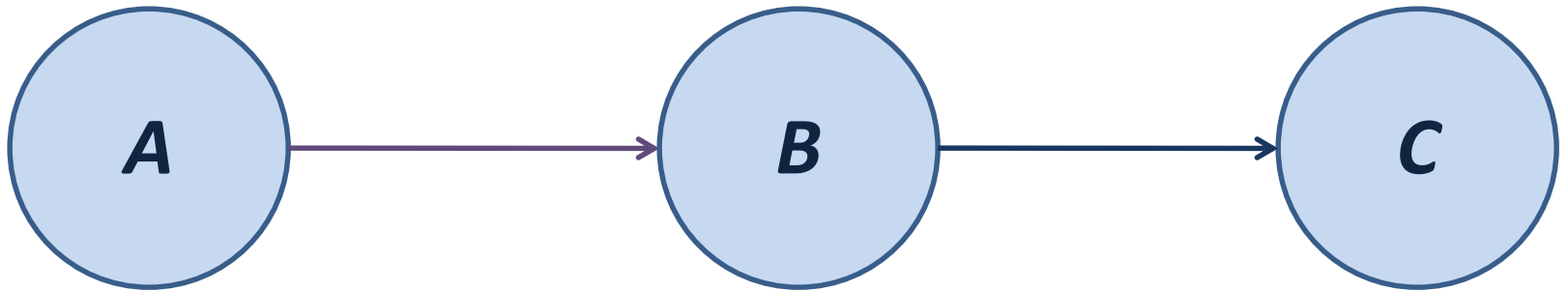
# For Example



# d-Separation

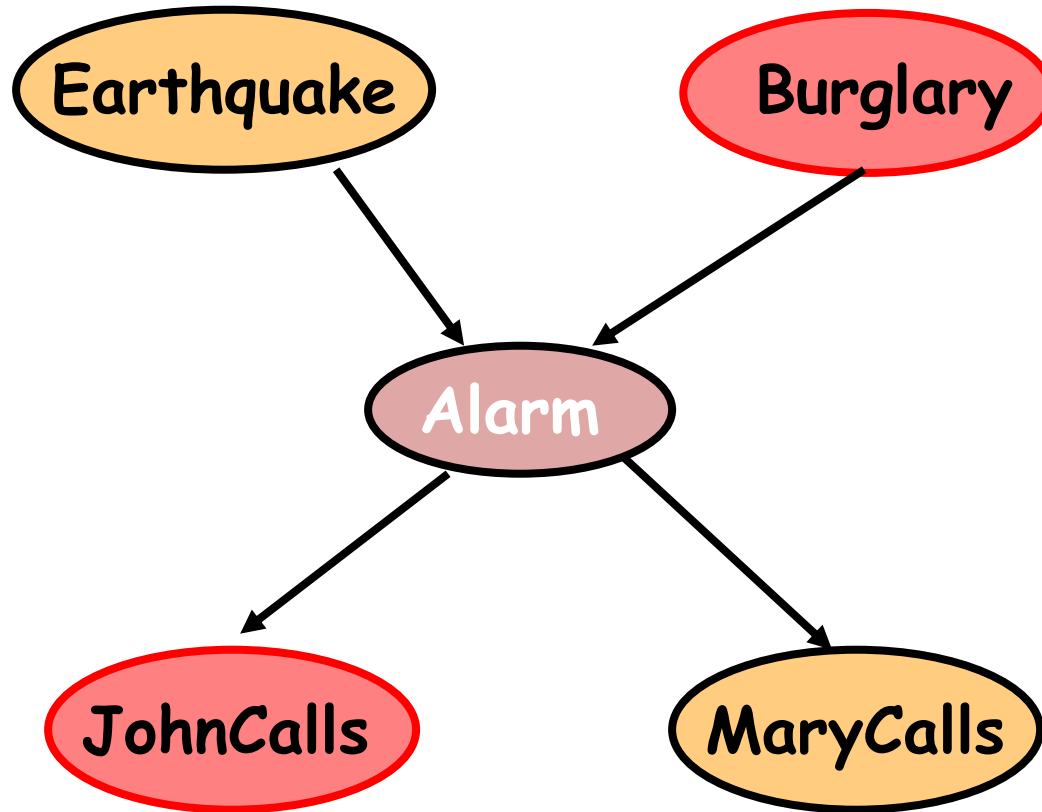
- An undirected path between two nodes is “cut off” if information cannot flow across one of the nodes in the path
- Two nodes are d-separated if every undirected path between them is cut off
- Two sets of nodes are d-separated if every pair of nodes, one from each set, is d-separated

# d-Separation



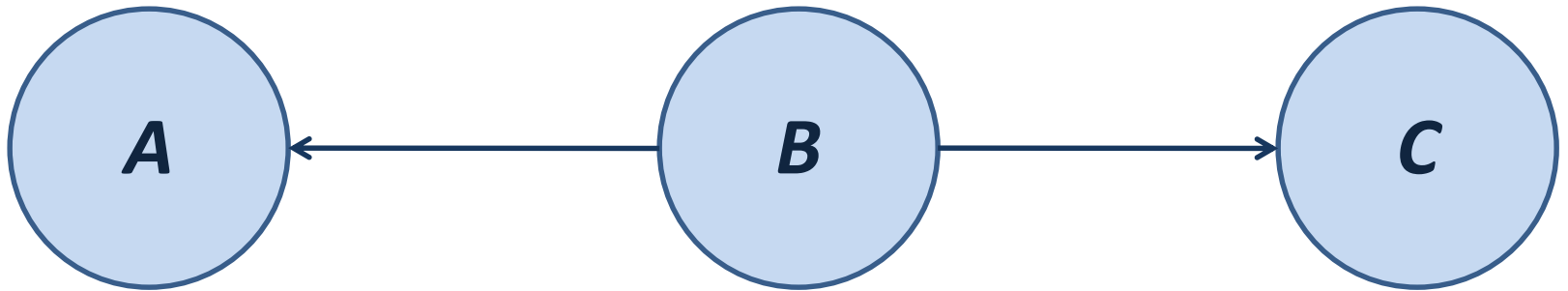
*Linear connection: Information can flow between A and C if and only if we do not have evidence at B*

# For Example



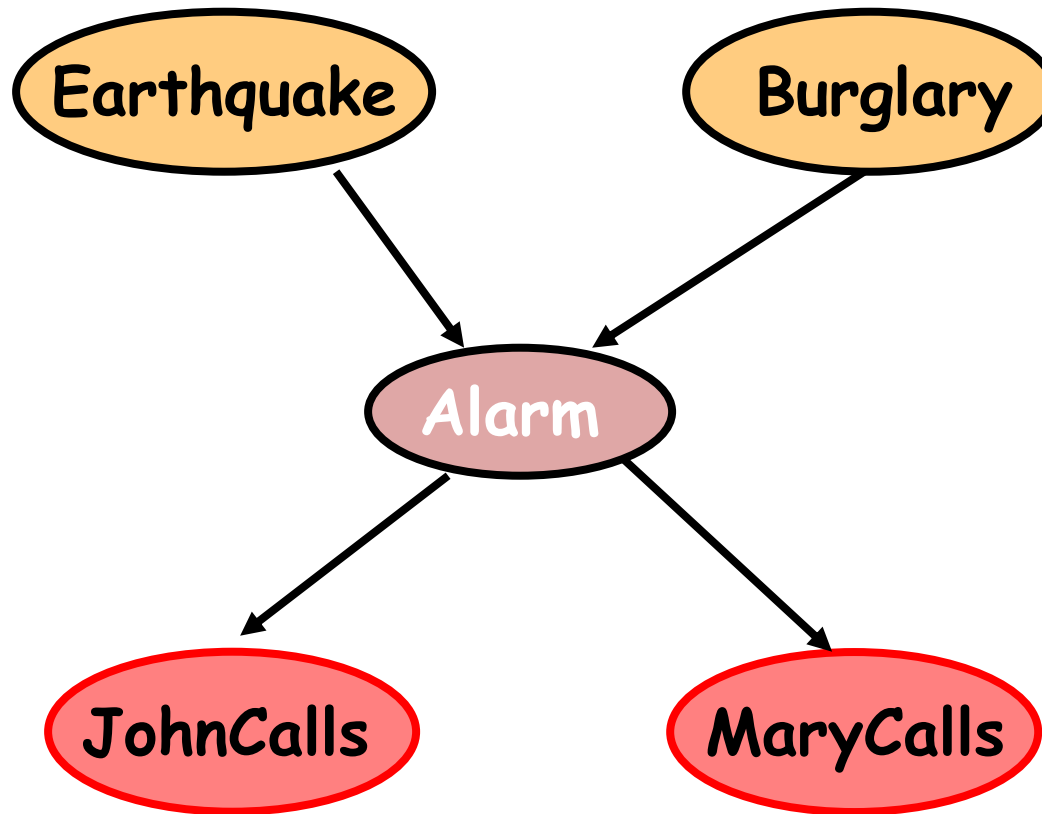


## d-Separation (continued)

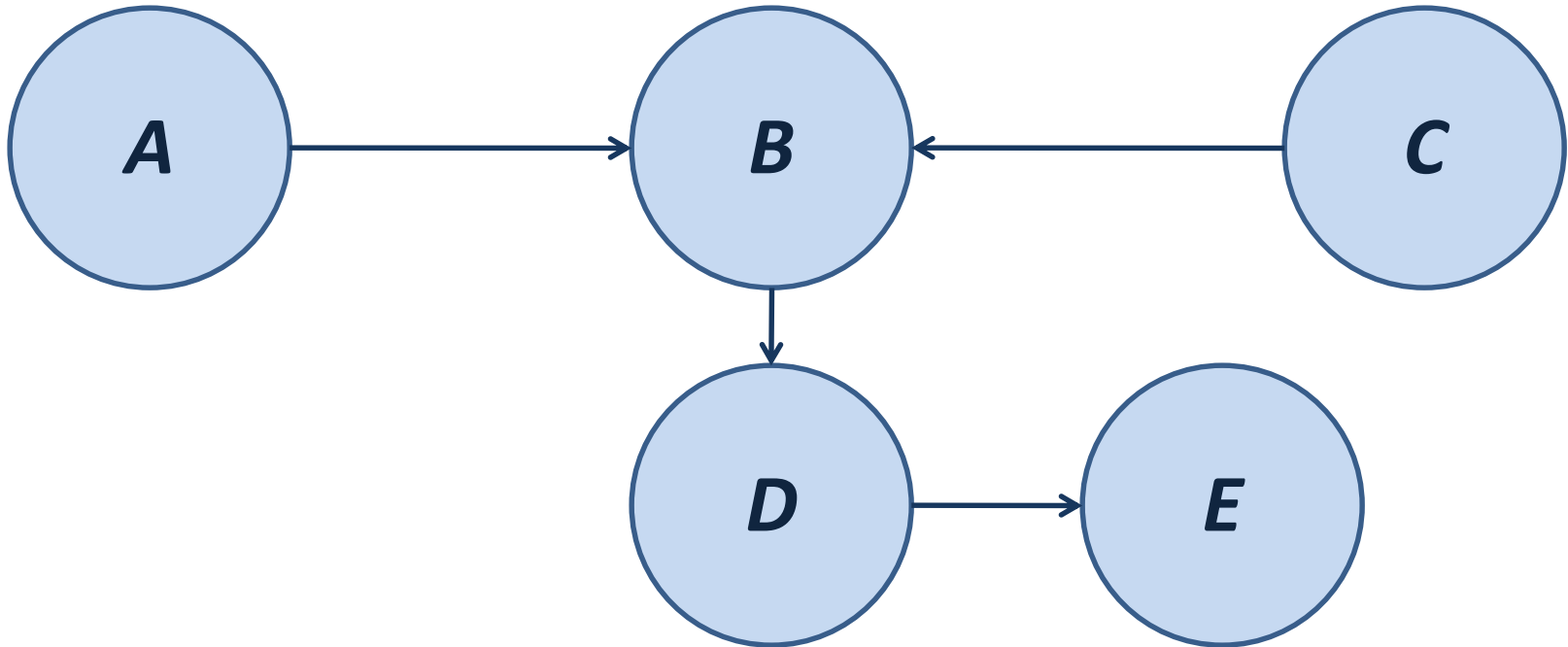


*Diverging connection: Information can flow between A and C if and only if we do not have evidence at B*

# For Example

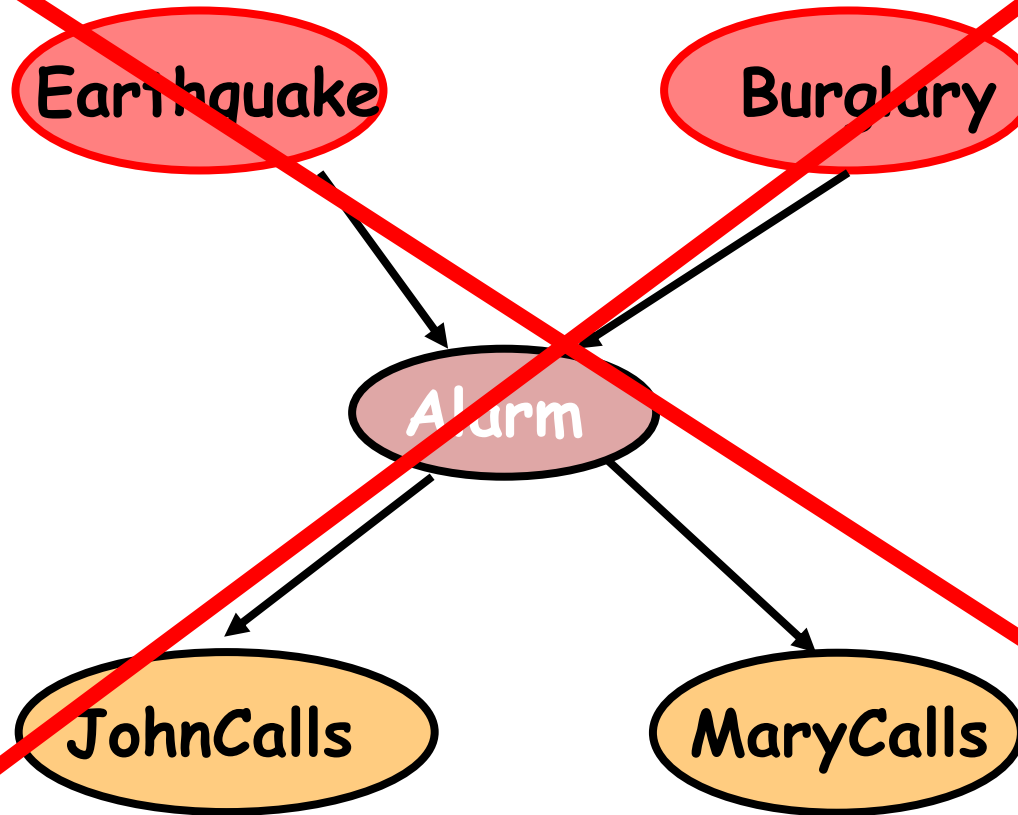


## d-Separation (continued)



*Converging connection: Information can flow between A and C if and only if we do have evidence at B or any descendent of B (such as D or E)*


# For Example



# d-Separation

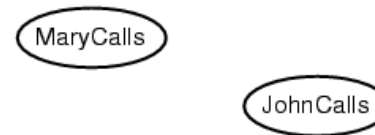
- An undirected path between two nodes is “cut off” if information cannot flow across one of the nodes in the path
- Two nodes are d-separated if every undirected path between them is cut off
- Two sets of nodes are d-separated if every pair of nodes, one from each set, is d-separated

## Note: For Some CPT Choices, More Conditional Independences May Hold

- Suppose we have: 
- Then only conditional independence we have is:  
$$P(A \perp C \mid B)$$
- Now choose CPTs such that A must be *True*, B must take same value as A, and C must take same value as B
- In the resulting distribution P, all pairs of variables are conditionally independent given the third

# Bayes Net Construction Example

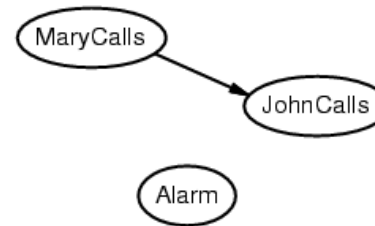
Suppose we choose the ordering  $M, J, A, B, E$



$$P(J \mid M) = P(J)?$$

# Example

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



$$P(J \mid M) = P(J)?$$

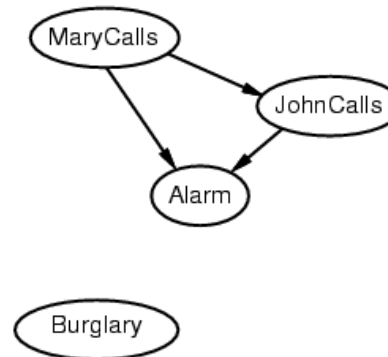
**No**

$$P(A \mid J, M) = P(A \mid J)? \quad P(A \mid M)? \quad P(A)?$$



# Example

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



$$P(J \mid M) = P(J)?$$

**No**

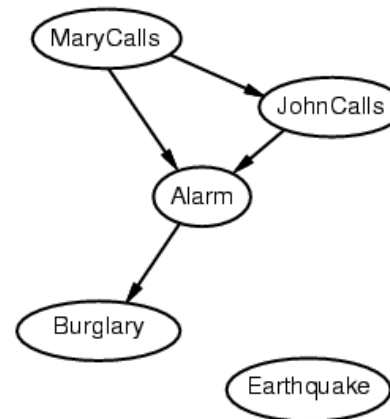
$$P(A \mid J, M) = P(A \mid J)? \quad P(A \mid J, M) = P(A)? \quad \mathbf{No}$$

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

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

# Example

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



$$P(J \mid M) = P(J)?$$

**No**

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

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

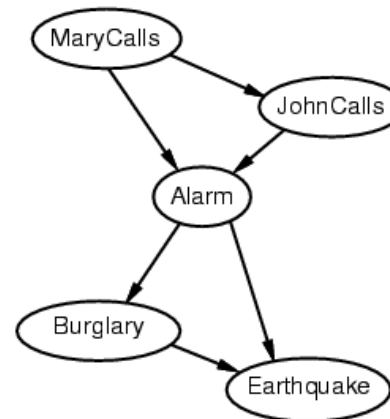
$$P(B \mid A, J, M) = P(B)? \quad \mathbf{No}$$

$$P(E \mid B, A, J, M) = P(E \mid A)?$$

$$P(E \mid B, A, J, M) = P(E \mid A, B)?$$

# Example

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



$$P(J \mid M) = P(J)?$$

**No**

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

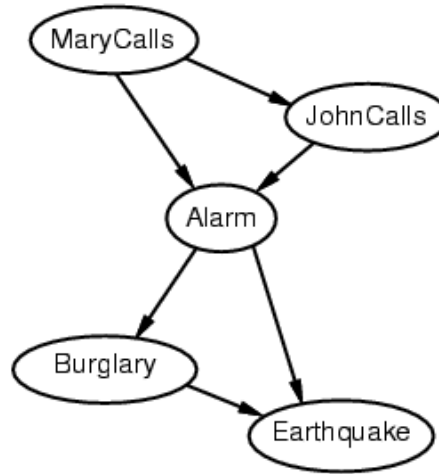
$$P(B \mid A, J, M) = P(B \mid A)? \quad \mathbf{Yes}$$

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

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

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

# Example contd.



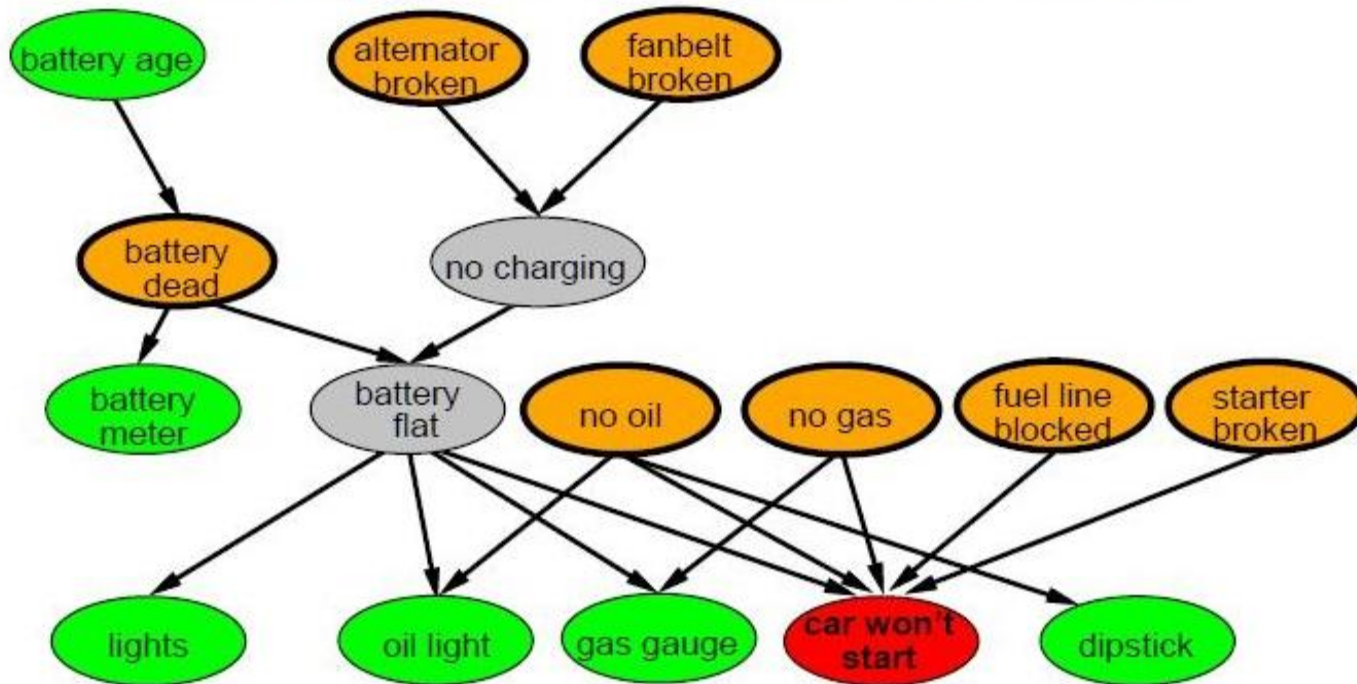
- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact:  $1 + 2 + 4 + 2 + 4 = 13$  numbers needed

# Example: Car Diagnosis

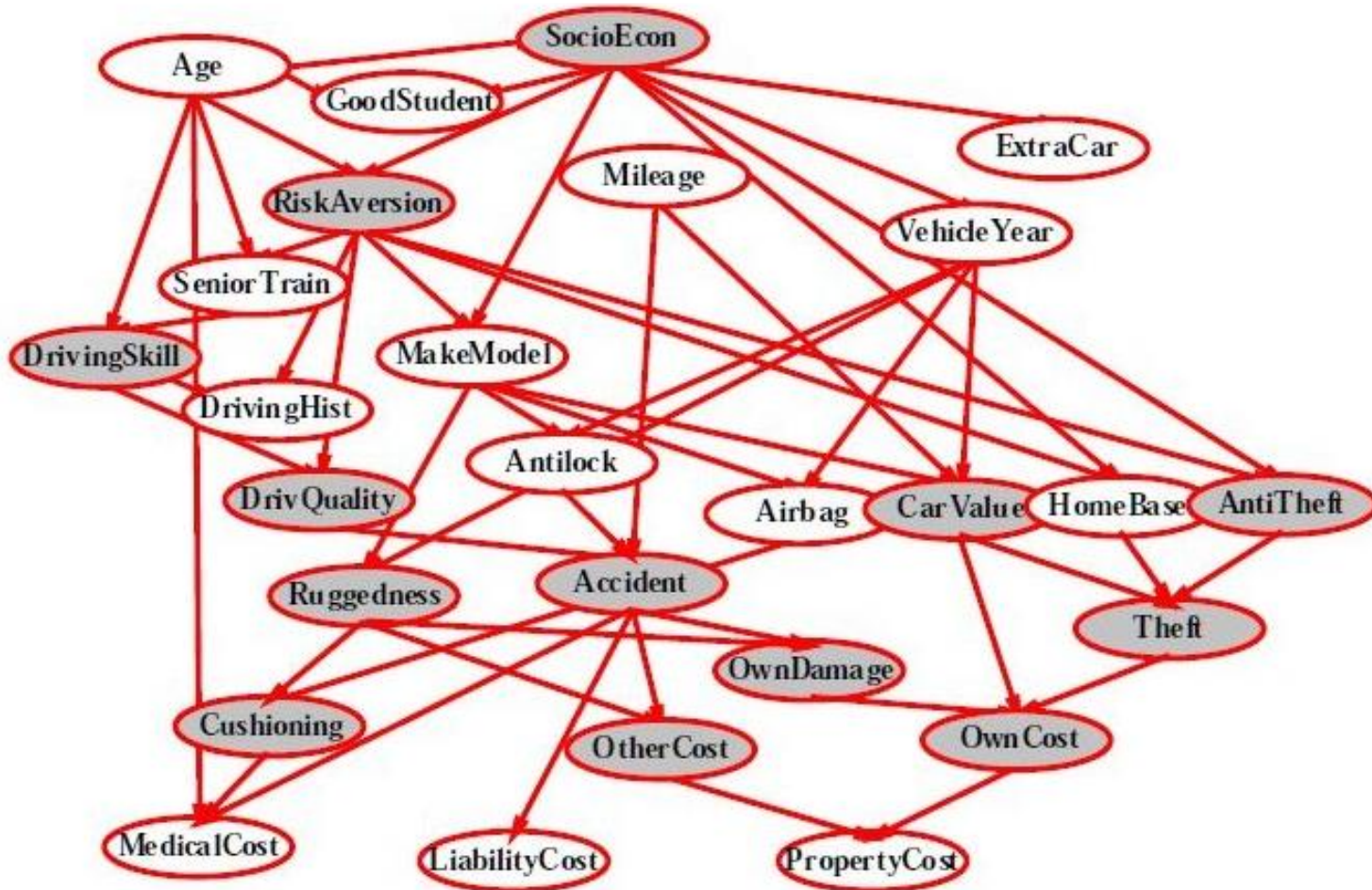
Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



# Example: Car Insurance



# Other Applications

- Medical Diagnosis
- Computational Biology and Bioinformatics
- Natural Language Processing
- Document classification
- Image processing
- Decision support systems
- Ecology & natural resource management
- Robotics
- Forensic science...

# Compact Conditionals

CPT grows exponentially with number of parents

CPT becomes infinite with continuous-valued parent or child

Solution: canonical distributions that are defined compactly

Deterministic nodes are the simplest case:

$$X = f(\text{Parents}(X)) \text{ for some function } f$$

E.g., Boolean functions

$$\text{NorthAmerican} \Leftrightarrow \text{Canadian} \vee \text{US} \vee \text{Mexican}$$

E.g., numerical relationships among continuous variables

$$\frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation}$$



# Compact Conditionals

Noisy-OR distributions model multiple noninteracting causes

- 1) Parents  $U_1 \dots U_k$  include all causes (can add leak node)
- 2) Independent failure probability  $q_i$  for each cause alone

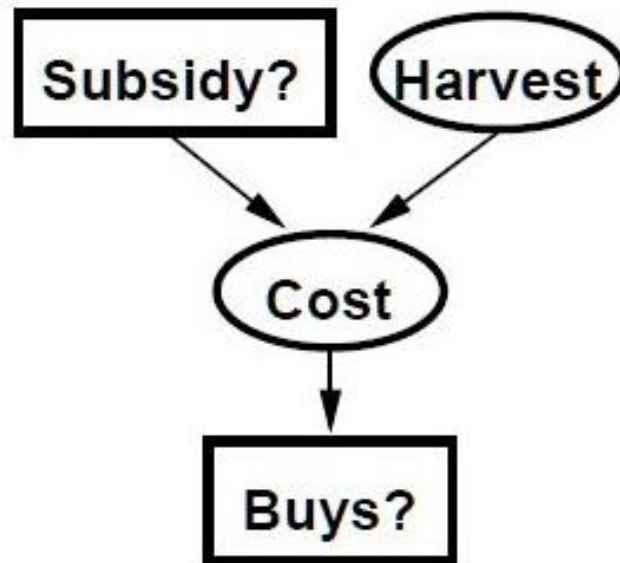
$$\Rightarrow P(X|U_1 \dots U_j, \neg U_{j+1} \dots \neg U_k) = 1 - \prod_{i=1}^j q_i$$

<i>Cold</i>	<i>Flu</i>	<i>Malaria</i>	$P(\text{Fever})$	$P(\neg \text{Fever})$
F	F	F	<b>0.0</b>	1.0
F	F	T	0.9	<b>0.1</b>
F	T	F	0.8	<b>0.2</b>
F	T	T	0.98	0.02 = 0.2 × 0.1
T	F	F	0.4	<b>0.6</b>
T	F	T	0.94	0.06 = 0.6 × 0.1
T	T	F	0.88	0.12 = 0.6 × 0.2
T	T	T	0.988	0.012 = 0.6 × 0.2 × 0.1

Number of parameters **linear** in number of parents

# Hybrid (discrete+cont) Networks

Discrete (*Subsidy?* and *Buys?*); continuous (*Harvest* and *Cost*)



Option 1: discretization—possibly large errors, large CPTs

Option 2: finitely parameterized canonical families

- 1) Continuous variable, discrete+continuous parents (e.g., *Cost*)
- 2) Discrete variable, continuous parents (e.g., *Buys?*)

# #1: Continuous Child Variables

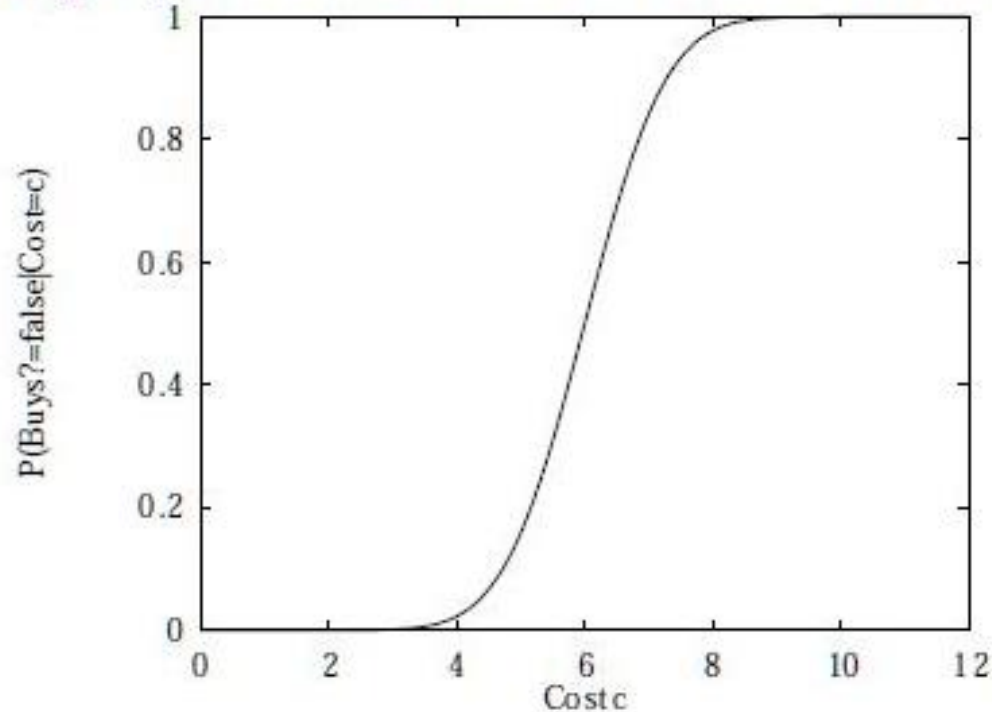
Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

Most common is the linear Gaussian model, e.g.,:

$$\begin{aligned} P(\text{Cost} = c | \text{Harvest} = h, \text{Subsidy?} = \text{true}) \\ &= N(a_t h + b_t, \sigma_t)(c) \\ &= \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{c - (a_t h + b_t)}{\sigma_t}\right)^2\right) \end{aligned}$$

# #2 Discrete child – cont. parents

Probability of *Buys?* given *Cost* should be a “soft” threshold:



Probit distribution uses integral of Gaussian:

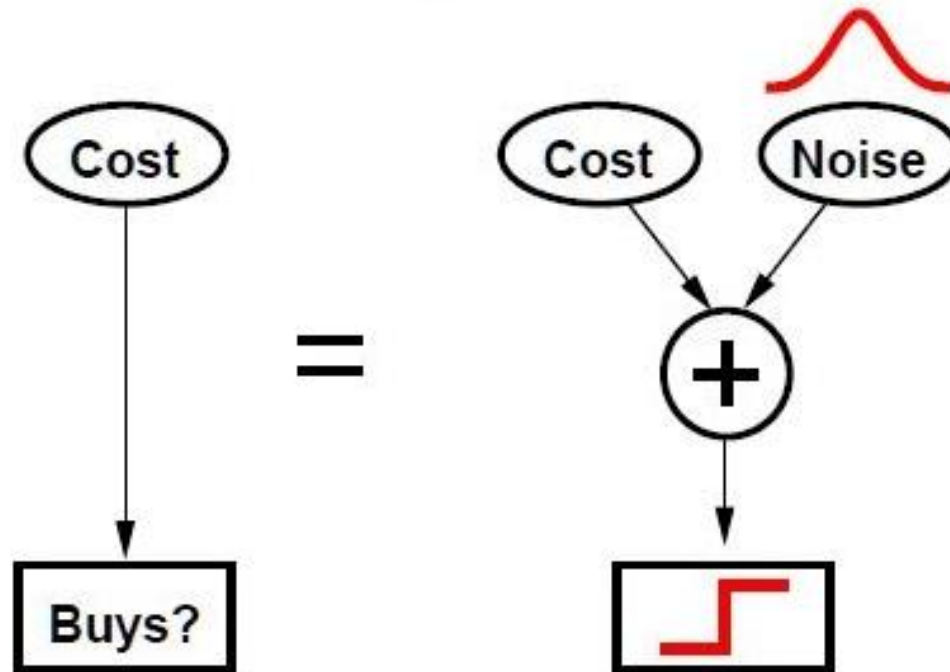
$$\Phi(x) = \int_{-\infty}^x N(0, 1)(x) dx$$

$$P(\text{Buys?} = \text{true} \mid \text{Cost} = c) = \Phi((-c + \mu)/\sigma)$$

• © D. Weld and D. Fox

# Why probit?

1. It's sort of the right shape
2. Can view as hard threshold whose location is subject to noise

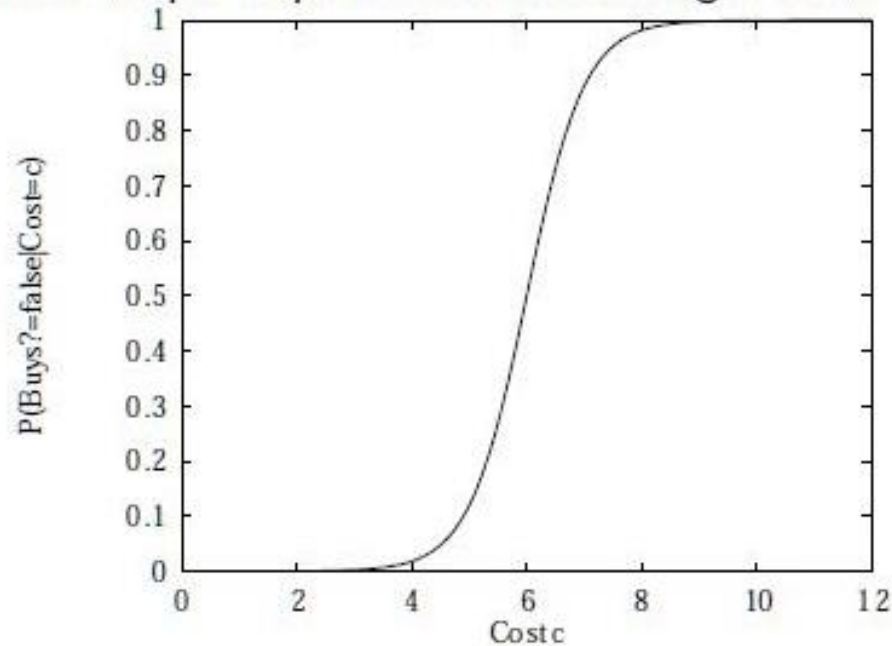


# Sigmoid Function

Sigmoid (or logit) distribution also used in neural networks:

$$P(\text{Buys?} = \text{true} \mid \text{Cost} = c) = \frac{1}{1 + \exp\left(-2\frac{-c + \mu}{\sigma}\right)}$$

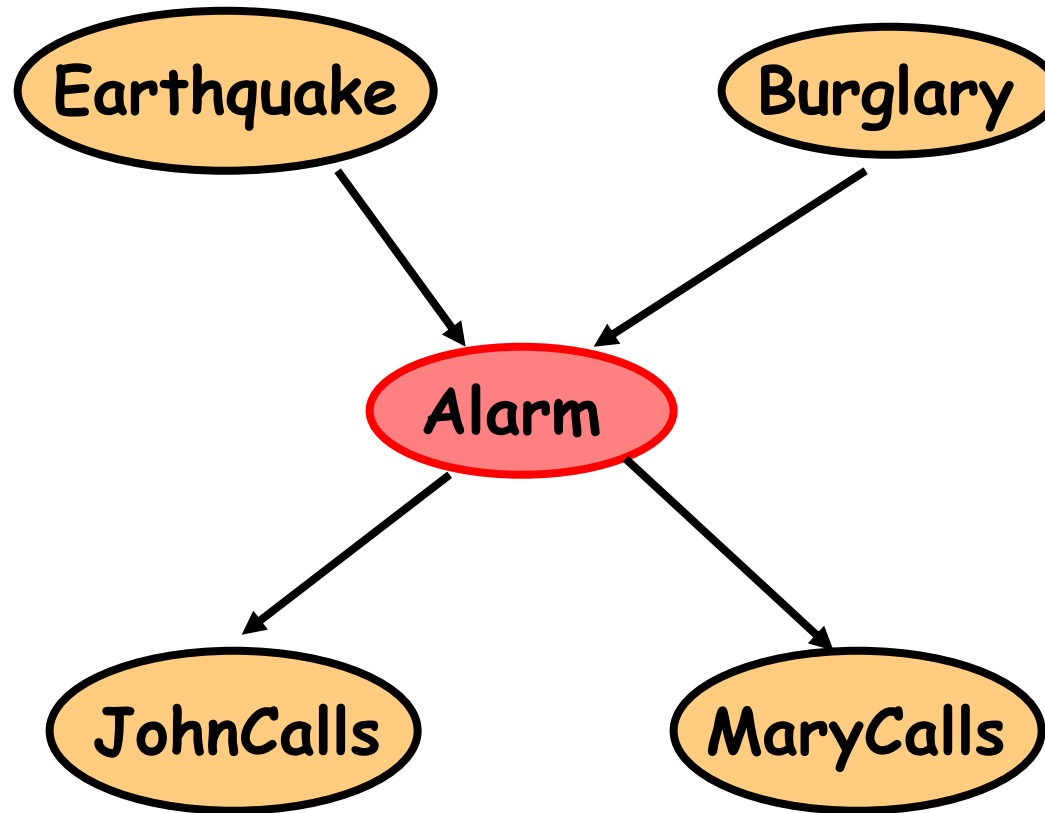
Sigmoid has similar shape to probit but much longer tails:



# Inference in BNs

- The graphical independence representation
  - yields efficient inference schemes
- We generally want to compute
  - Marginal probability:  $Pr(Z)$ ,
  - $Pr(Z/\mathbf{E})$  where  $\mathbf{E}$  is (conjunctive) evidence
    - Z: query variable(s),
    - E: evidence variable(s)
    - everything else: hidden variable
- Computations organized by network topology

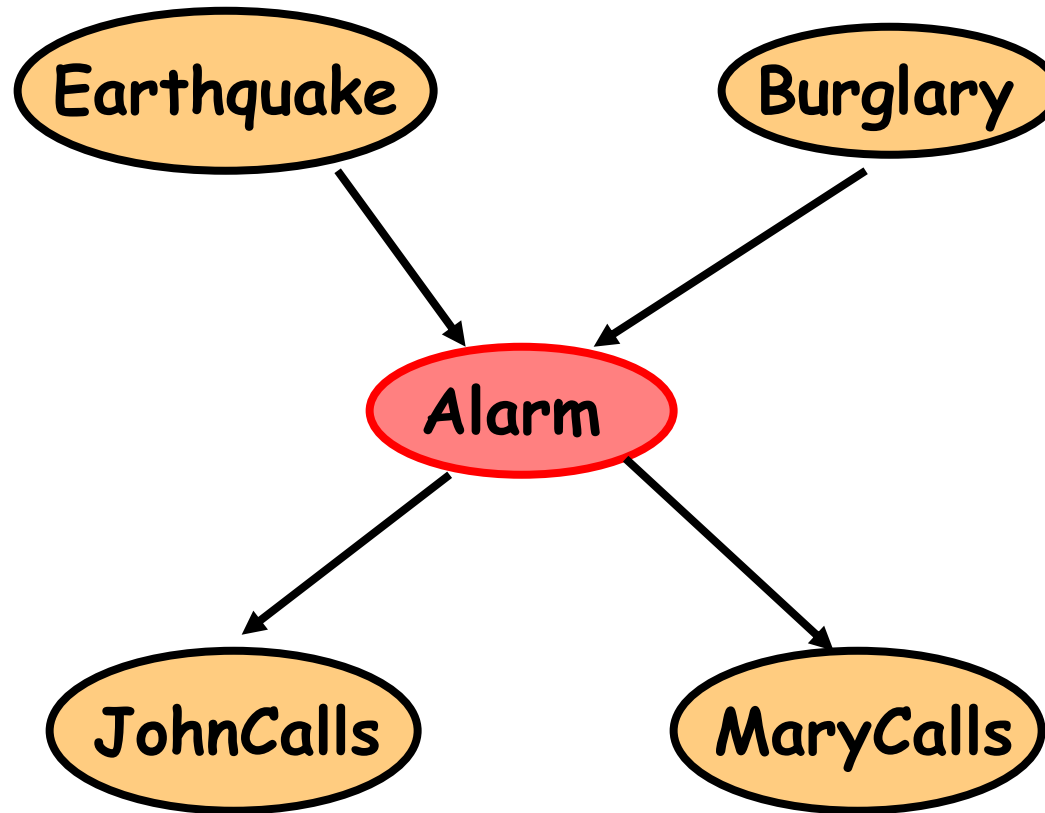
$P(B \mid J=\text{true}, M=\text{true})$



$$P(b|j,m) = \alpha \sum_{e,a} P(b,j,m,e,a)$$



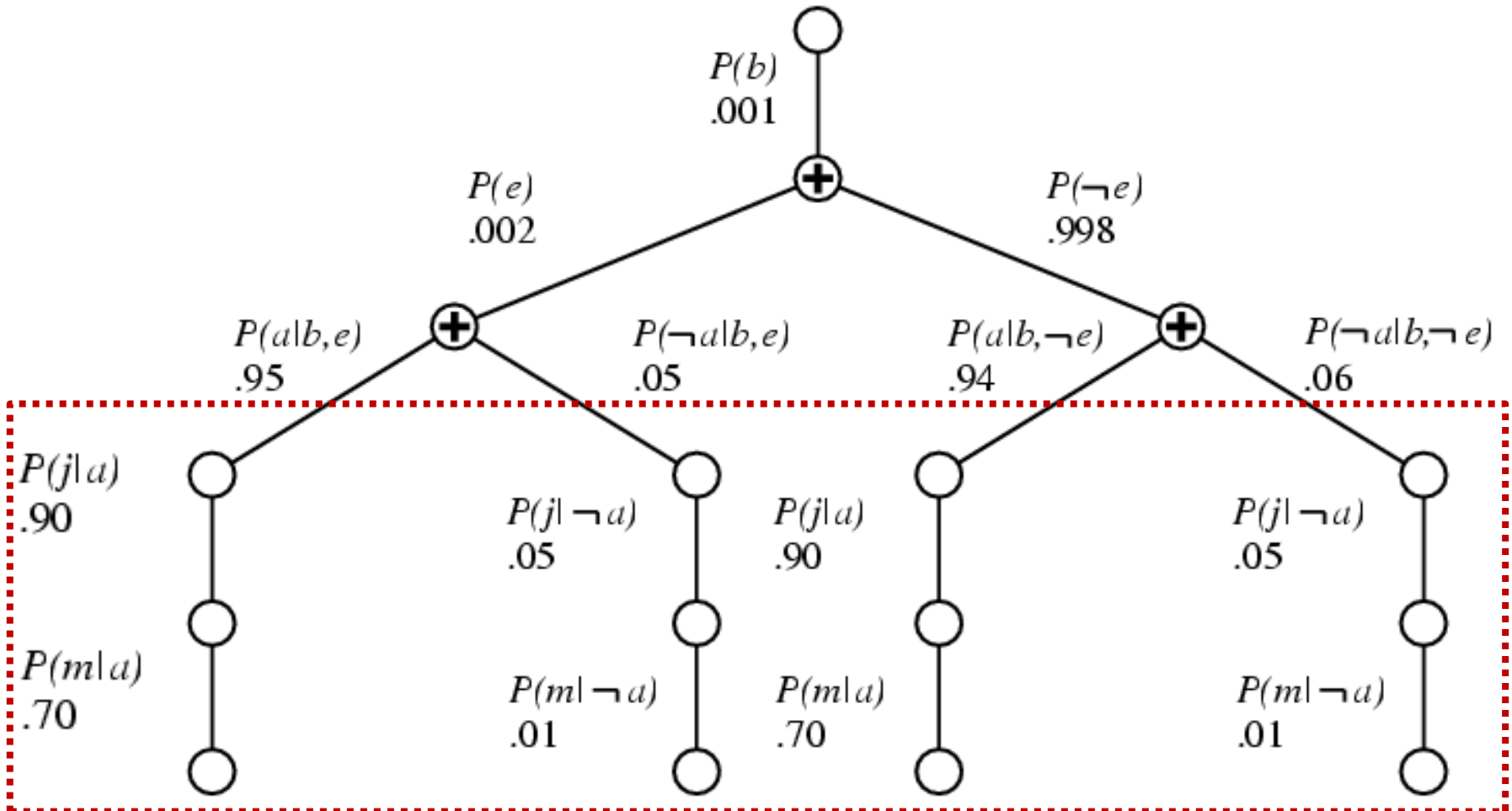
$P(B \mid J=\text{true}, M=\text{true})$



$$P(b \mid j, m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b, e) P(j \mid a) P(m \mid a)$$

# Variable Elimination

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



Repeated computations → Dynamic Programming

# Variable Elimination

- A *factor* is a function from some set of variables into a specific value: e.g.,  $f(E,A,N1)$ 
  - CPTs are factors, e.g.,  $P(A/E,B)$  function of  $A,E,B$
- VE works by *eliminating* all variables in turn until there is a factor with only query variable
- To eliminate a variable:
  - *join* all factors containing that variable (like DB)
  - *sum out* the influence of the variable on new factor
  - exploits product form of joint distribution

# Example of VE: $P(JC)$

$P(J)$

$$= \sum_{M,A,B,E} P(J,M,A,B,E)$$

$$= \sum_{M,A,B,E} P(J|A)P(M|A) P(B)P(A|B,E)P(E)$$

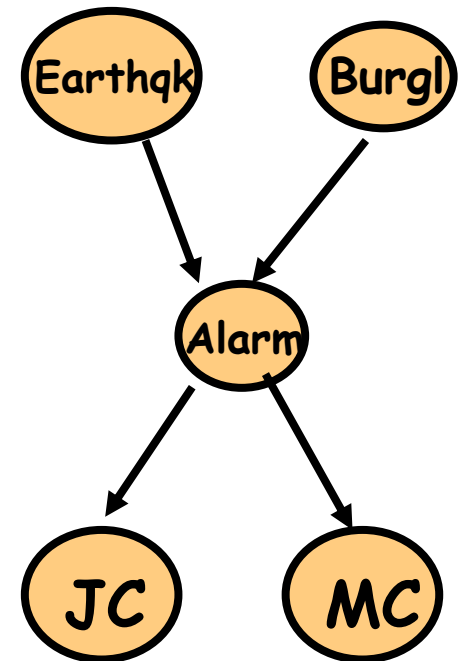
$$= \sum_A P(J|A) \sum_M P(M|A) \sum_B P(B) \sum_E P(A|B,E)P(E)$$

$$= \sum_A P(J|A) \sum_M P(M|A) \sum_B P(B) f1(A,B)$$

$$= \sum_A P(J|A) \sum_M P(M|A) f2(A)$$

$$= \sum_A P(J|A) f3(A)$$

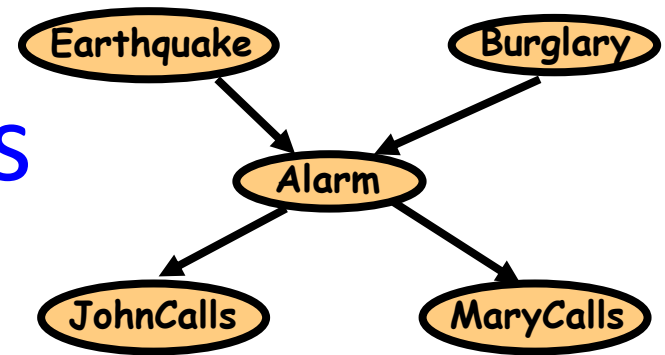
$$= f4(J)$$



# Notes on VE

- Each operation is a simple multiplication of factors and summing out a variable
- Complexity determined by size of largest factor
  - in our example, 3 vars (not 5)
  - linear in number of vars,
  - exponential in largest factor elimination ordering greatly impacts factor size
  - optimal elimination orderings: NP-hard
  - heuristics, special structure (e.g., polytrees)
- Practically, inference is much more tractable using structure of this sort

# Irrelevant variables



$$P(J)$$

$$= \sum_{M,A,B,E} P(J,M,A,B,E)$$

$$= \sum_{M,A,B,E} P(J|A)P(B)P(A|B,E)P(E)P(M|A)$$

$$= \sum_A P(J|A) \sum_B P(B) \sum_E P(A|B,E)P(E) \sum_M P(M|A)$$

$$= \sum_A P(J|A) \sum_B P(B) \sum_E P(A|B,E)P(E)$$

$$= \sum_A P(J|A) \sum_B P(B) f_1(A,B)$$

$$= \sum_A P(J|A) f_2(A)$$

$$= f_3(J)$$

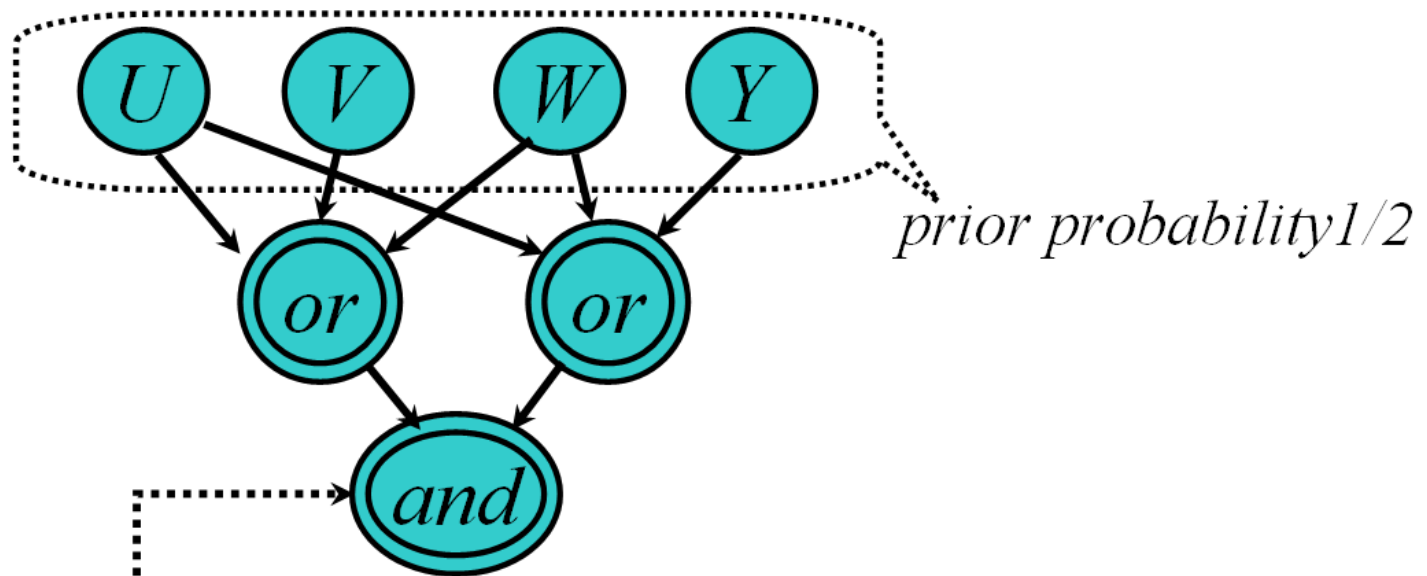
M is irrelevant to the computation

Thm:  $Y$  is irrelevant unless  $Y \in \text{Ancestors}(Z \cup E)$

# Reducing 3-SAT to Bayes Nets

- **Theorem:** Inference in a multi-connected Bayesian network is NP-hard.

Boolean 3CNF formula  $\phi = (u \vee \bar{v} \vee w) \wedge (\bar{u} \vee \bar{w} \vee y)$



Probability ( ) =  $1/2^n \cdot \#$  satisfying assignments of  $\phi$

© D. Weld and D. Fox

•56 70

# Complexity of Exact Inference

- Exact inference is NP hard
  - 3-SAT to Bayes Net Inference
  - It can count no. of assignments for 3-SAT: #P complete
- Inference in tree-structured Bayesian network
  - Polynomial time
  - compare with inference in CSPs
- Approximate Inference
  - Sampling based techniques