Bayesian Networks
Chapter 14

Mausam

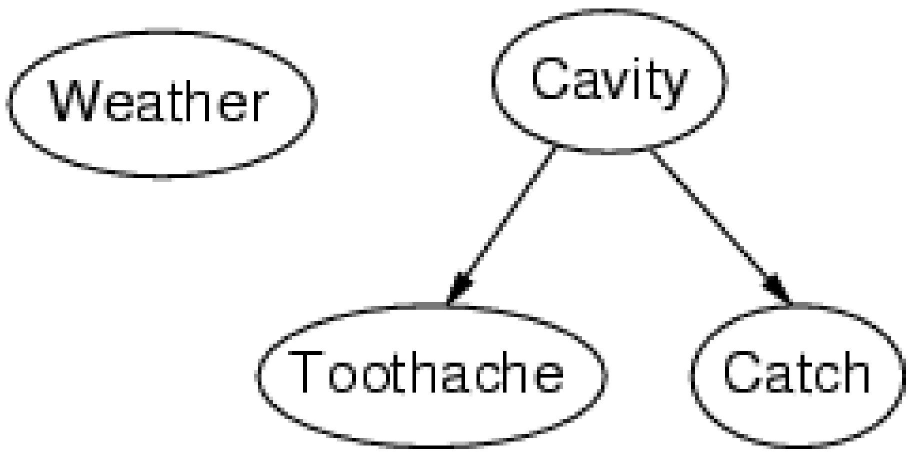
(Slides by UW-AI faculty, Stuart Russell & David Page)

Bayes Nets

- In general, joint distribution *P* over set of variables ($X_1 \times \ldots \times X_n$) 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 <u>conditionally independent</u> of each other given Cavity

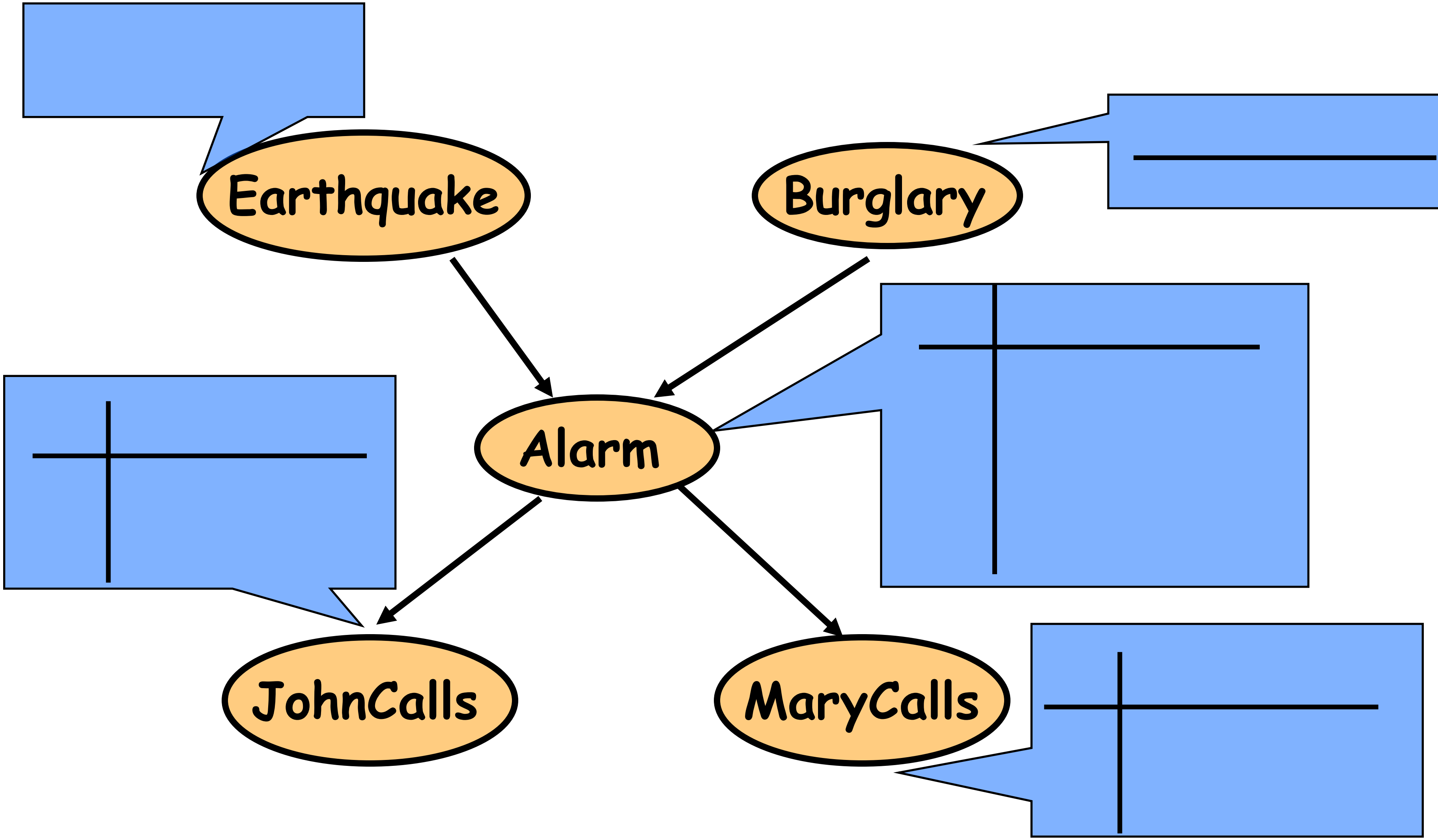
Syntax

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

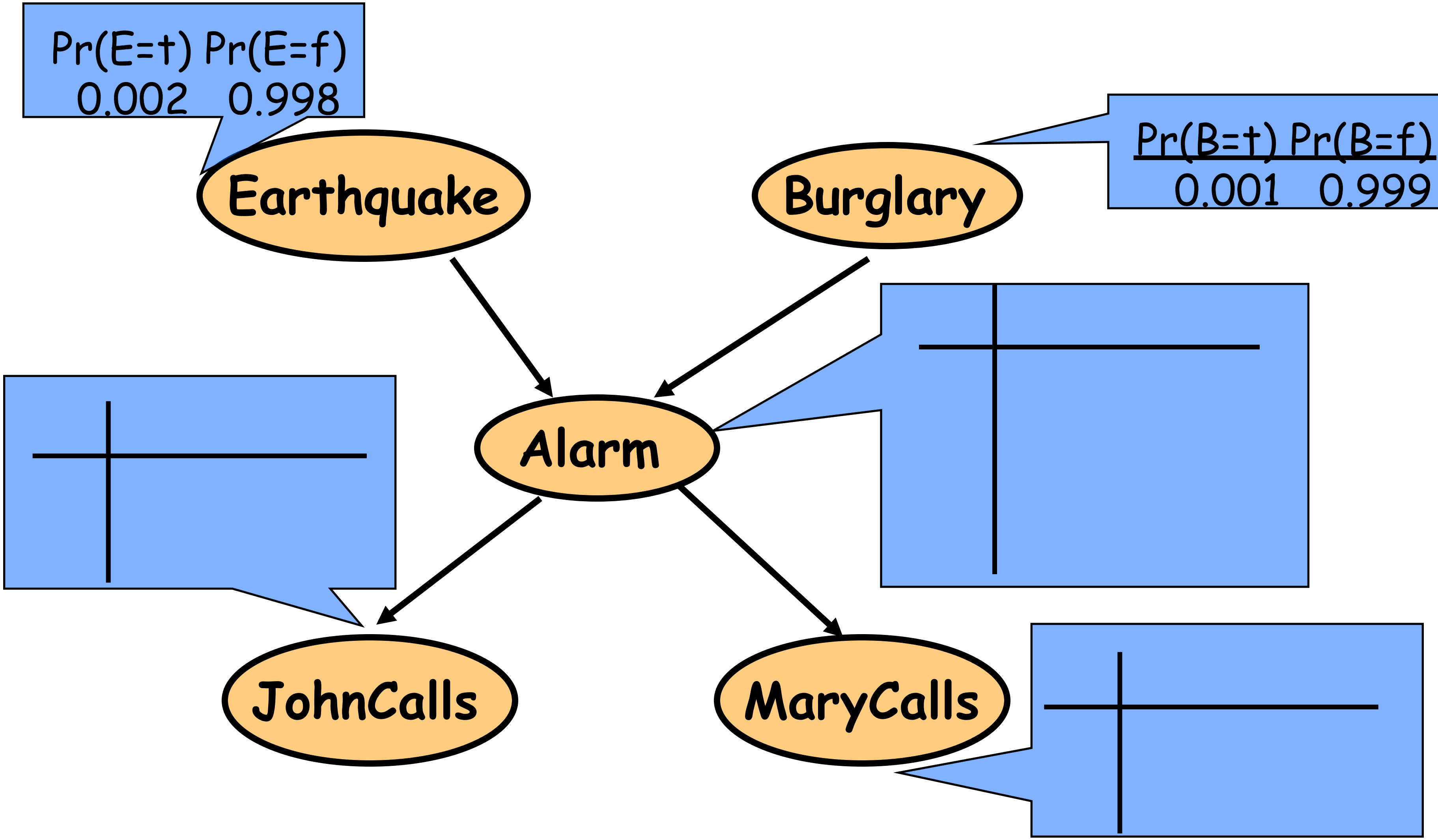
Burglars and Earthquakes

- 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

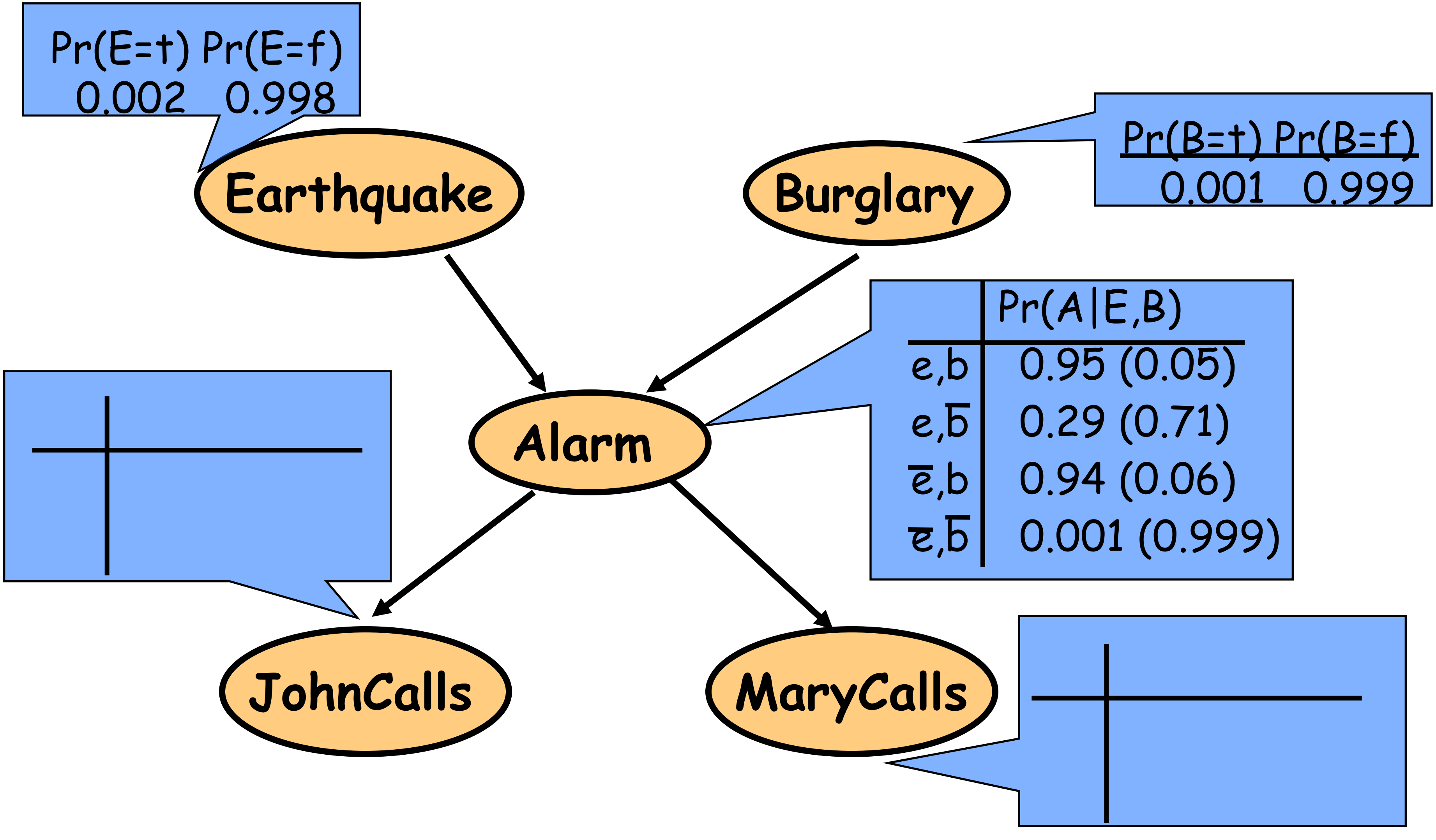
Burglars and Earthquakes



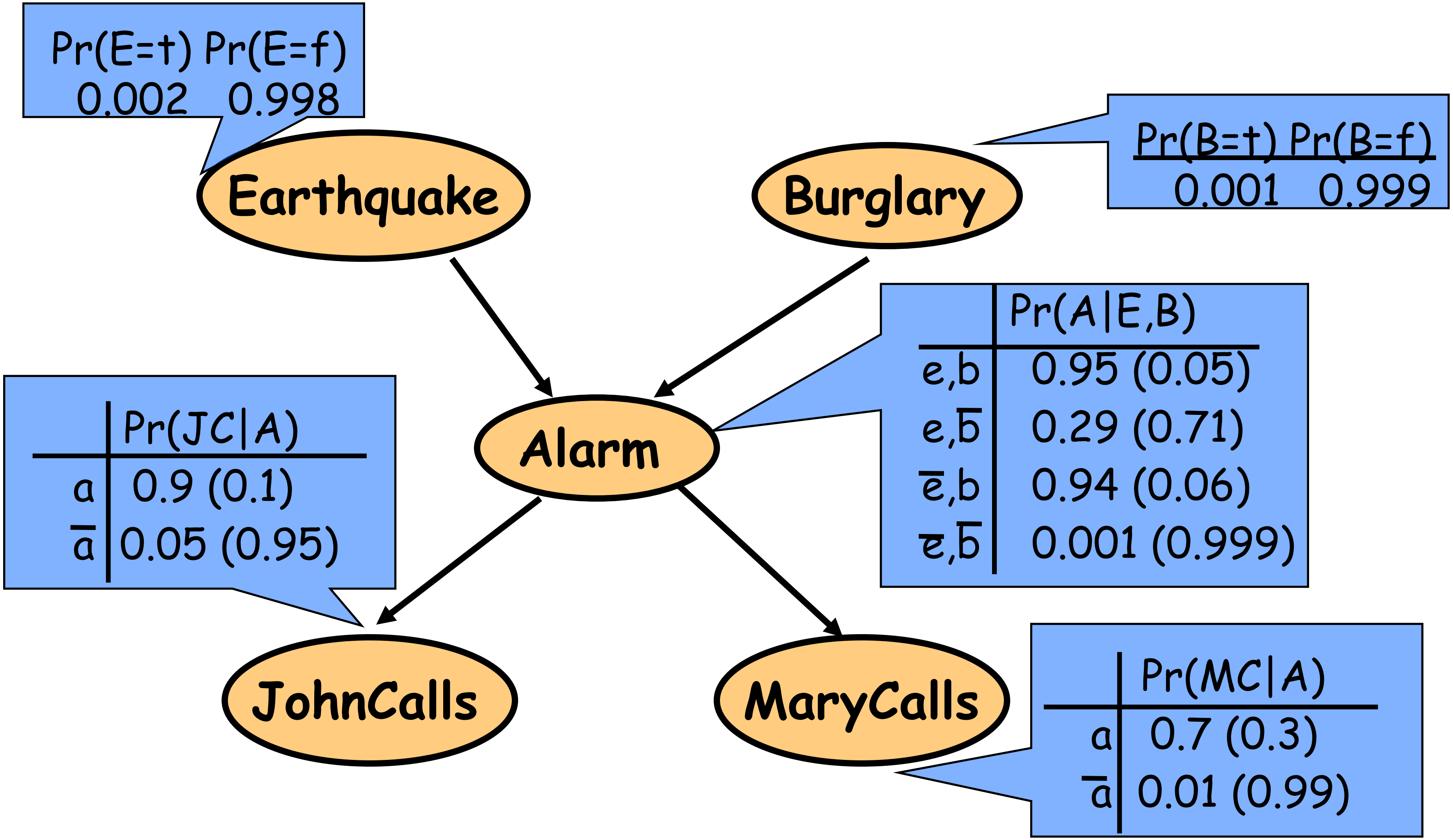
Burglars and Earthquakes



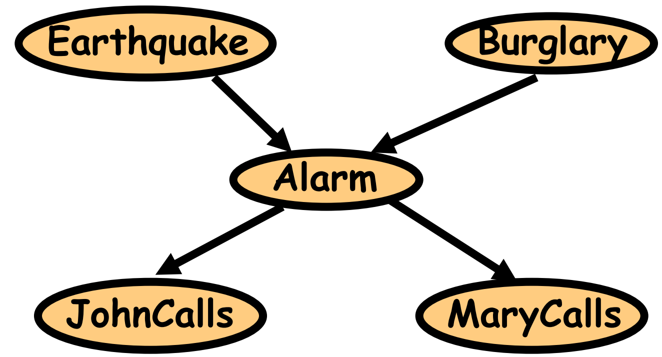
Burglars and Earthquakes



Burglars and Earthquakes



Earthquake Example (cont'd)



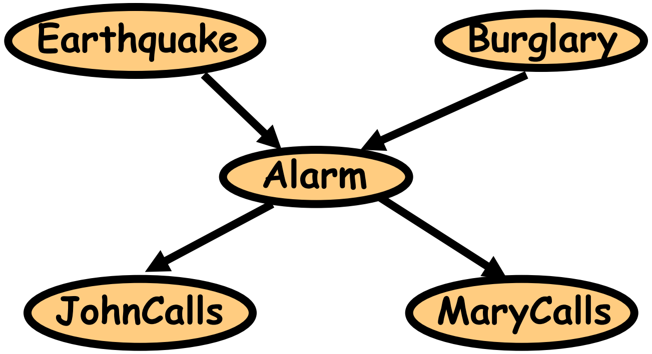
- **If we know *Alarm*, no other evidence influences our degree of belief in *JohnCalls***
  - $P(JC|MC,A,E,B) = P(JC|A)$
  - also: $P(MC|JC,A,E,B) = P(MC|A)$ and $P(E|B) = P(E)$
- **By the chain rule we have**

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

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

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

Earthquake Example (Global Semantics)



- **We just proved**

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

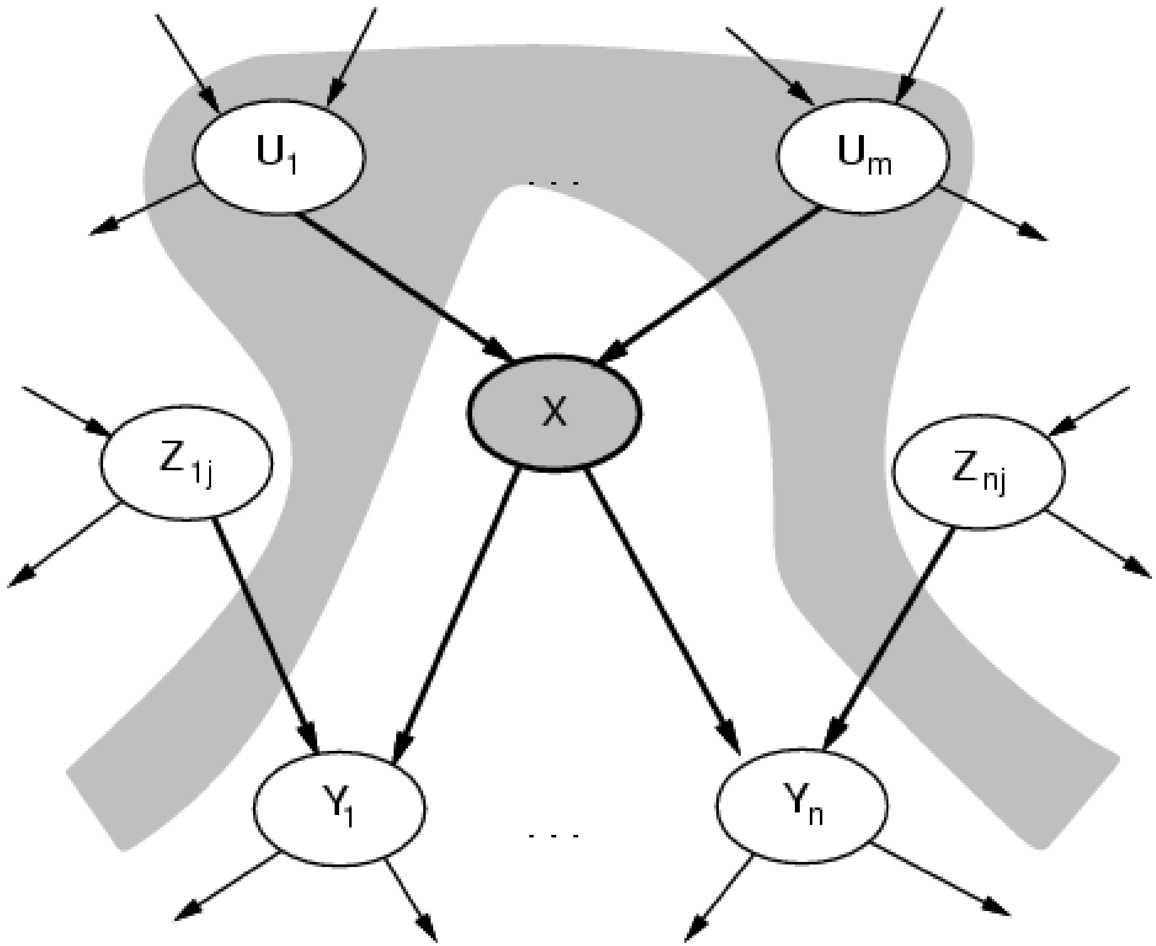
- **In general full joint distribution of a Bayes net is defined as**

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

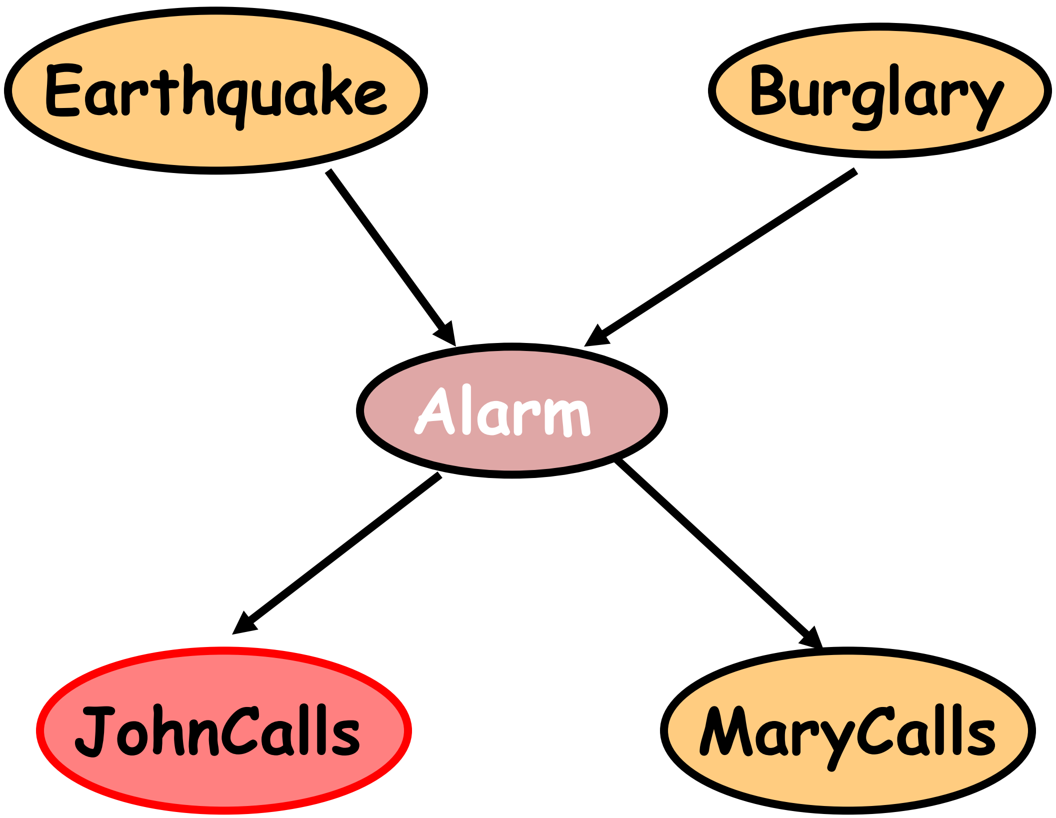
BNs: Qualitative Structure

- 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 nondescendents given its parents**
  - Graphical test exists for more general independence
  - “Markov Blanket”

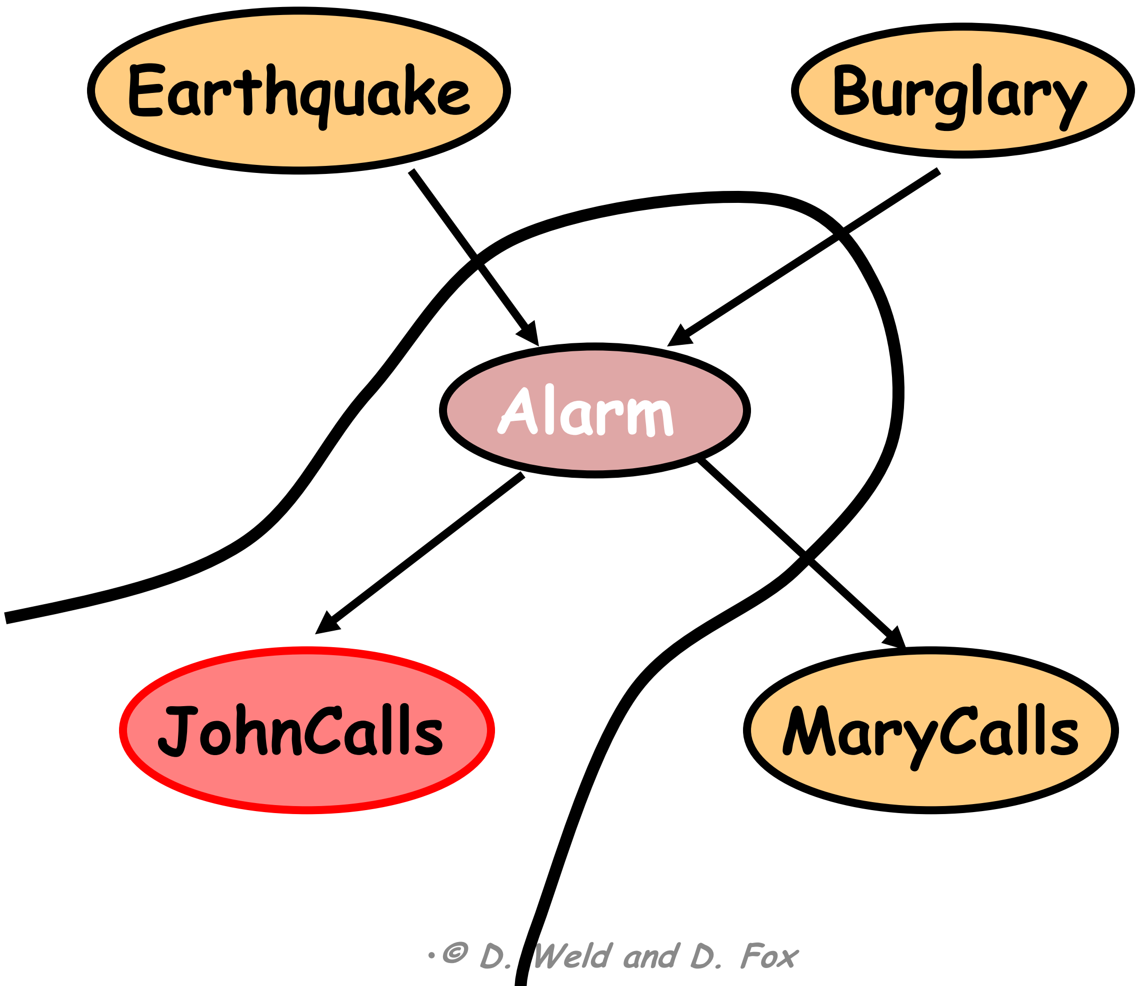
Given Parents, X is Independent of Non-Descendants



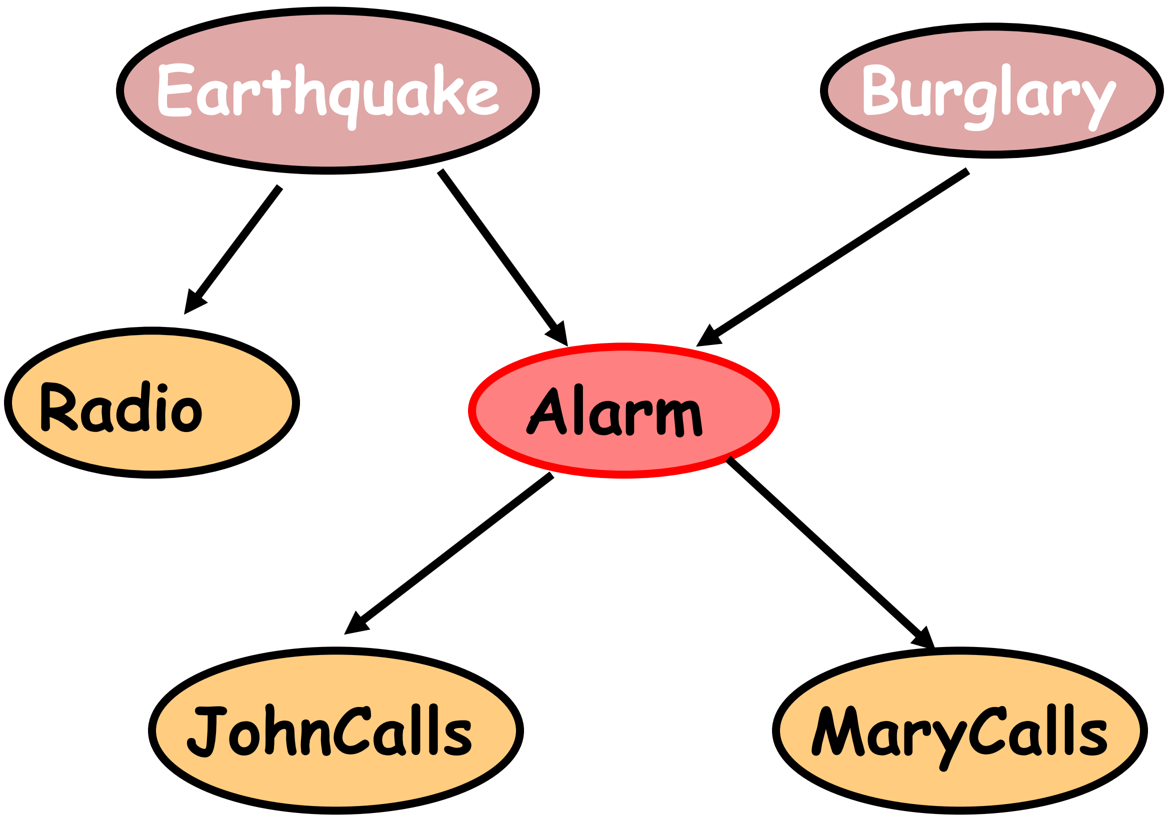
Examples



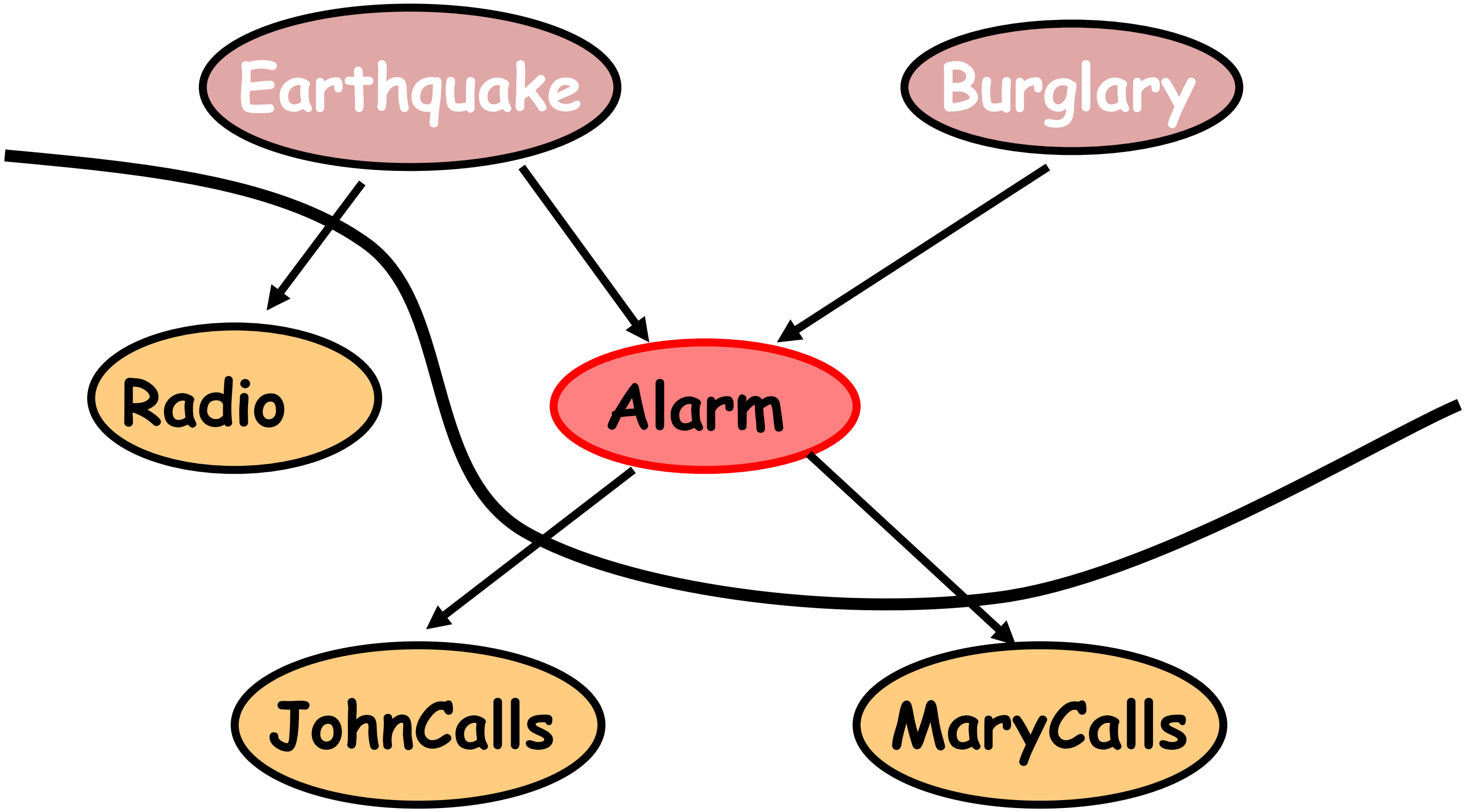
For Example



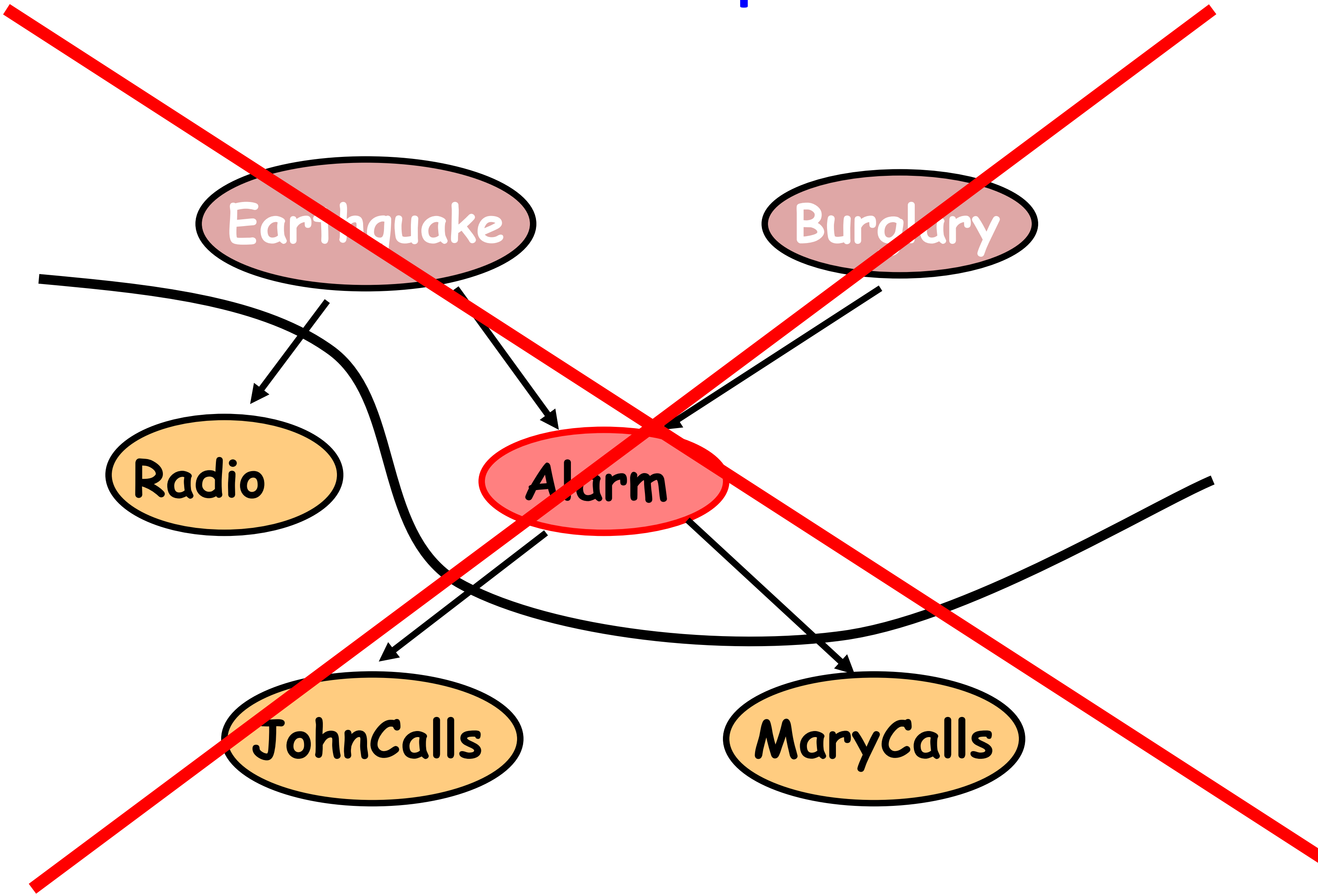
For Example



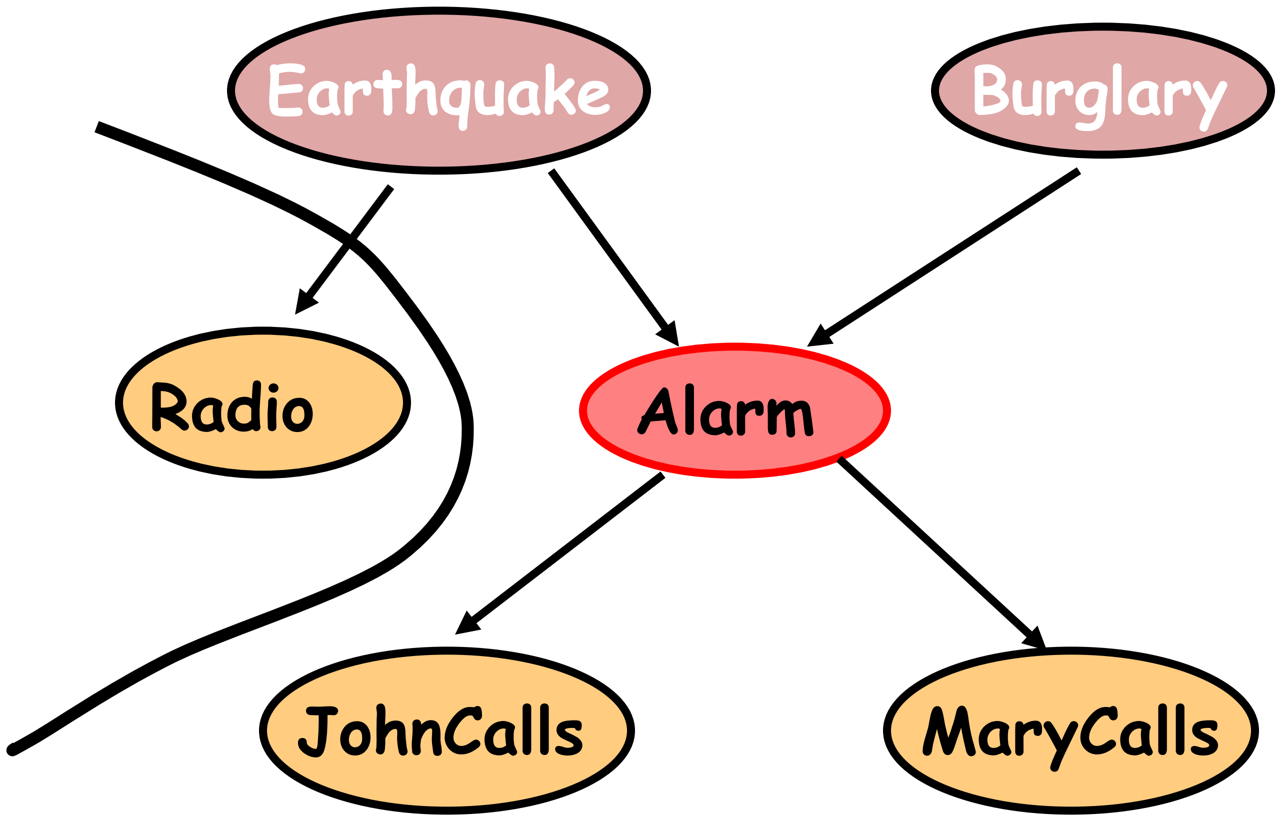
For Example



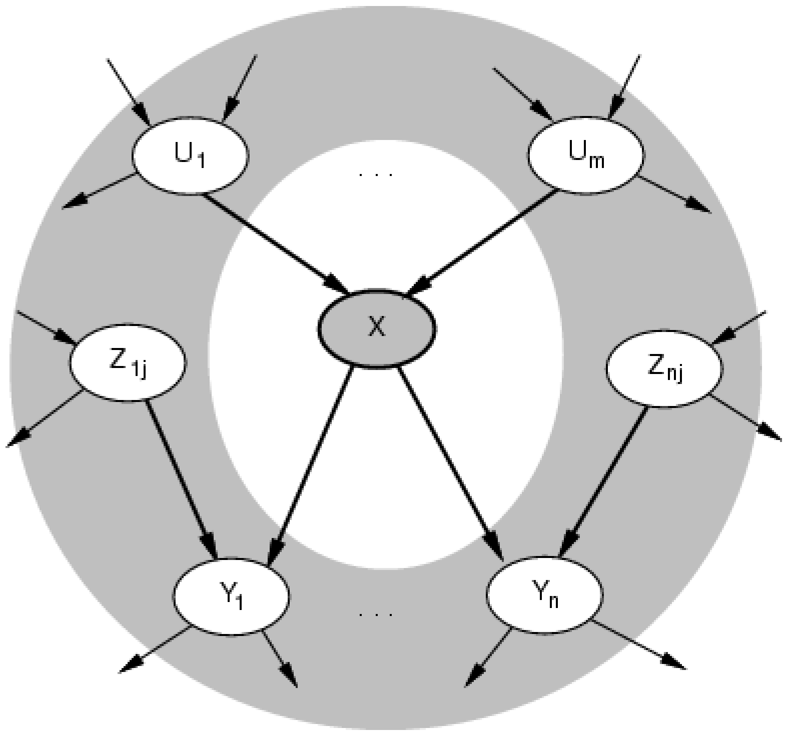
For Example



For Example

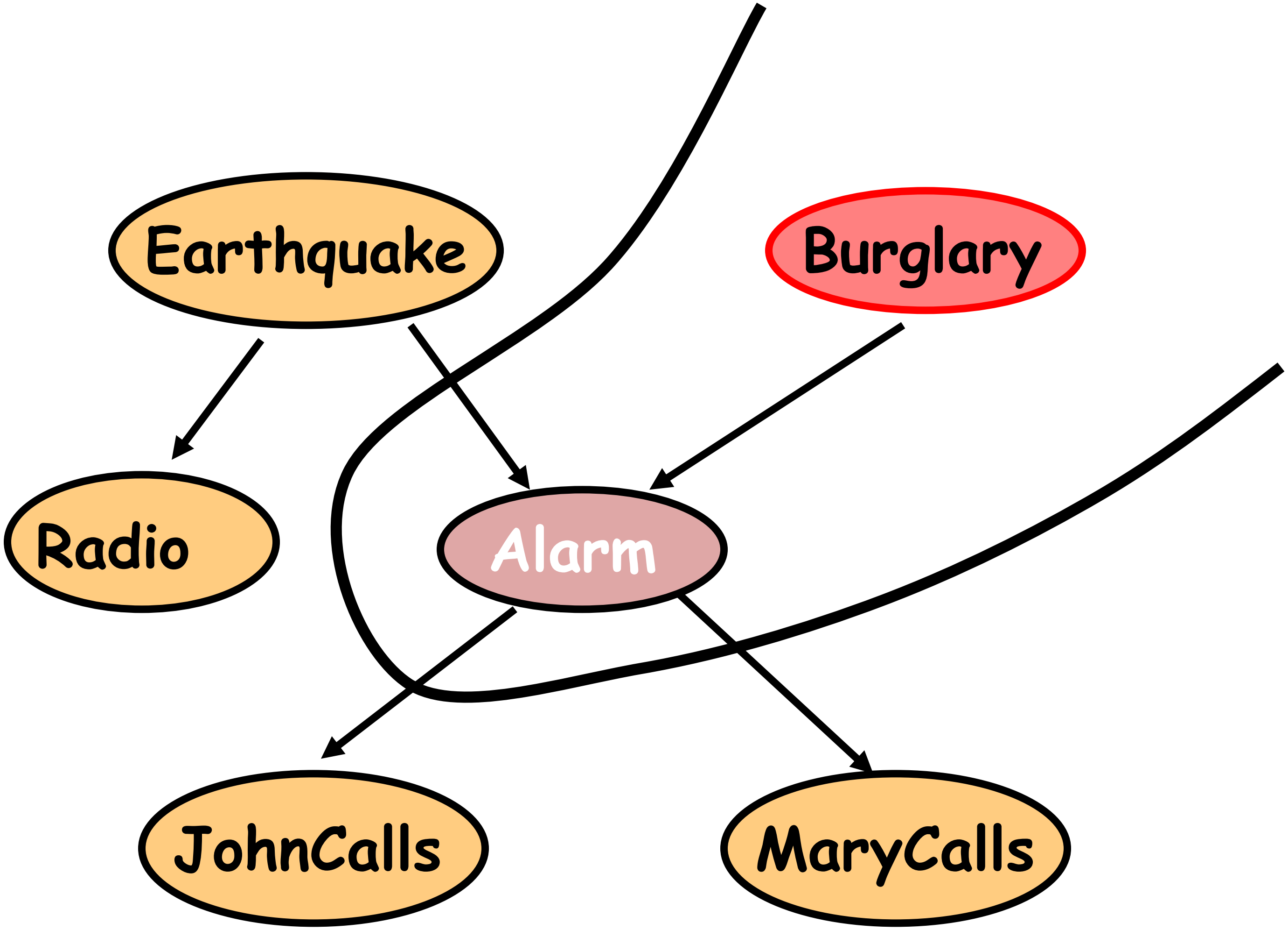


Given Markov Blanket, X is Independent of All Other Nodes

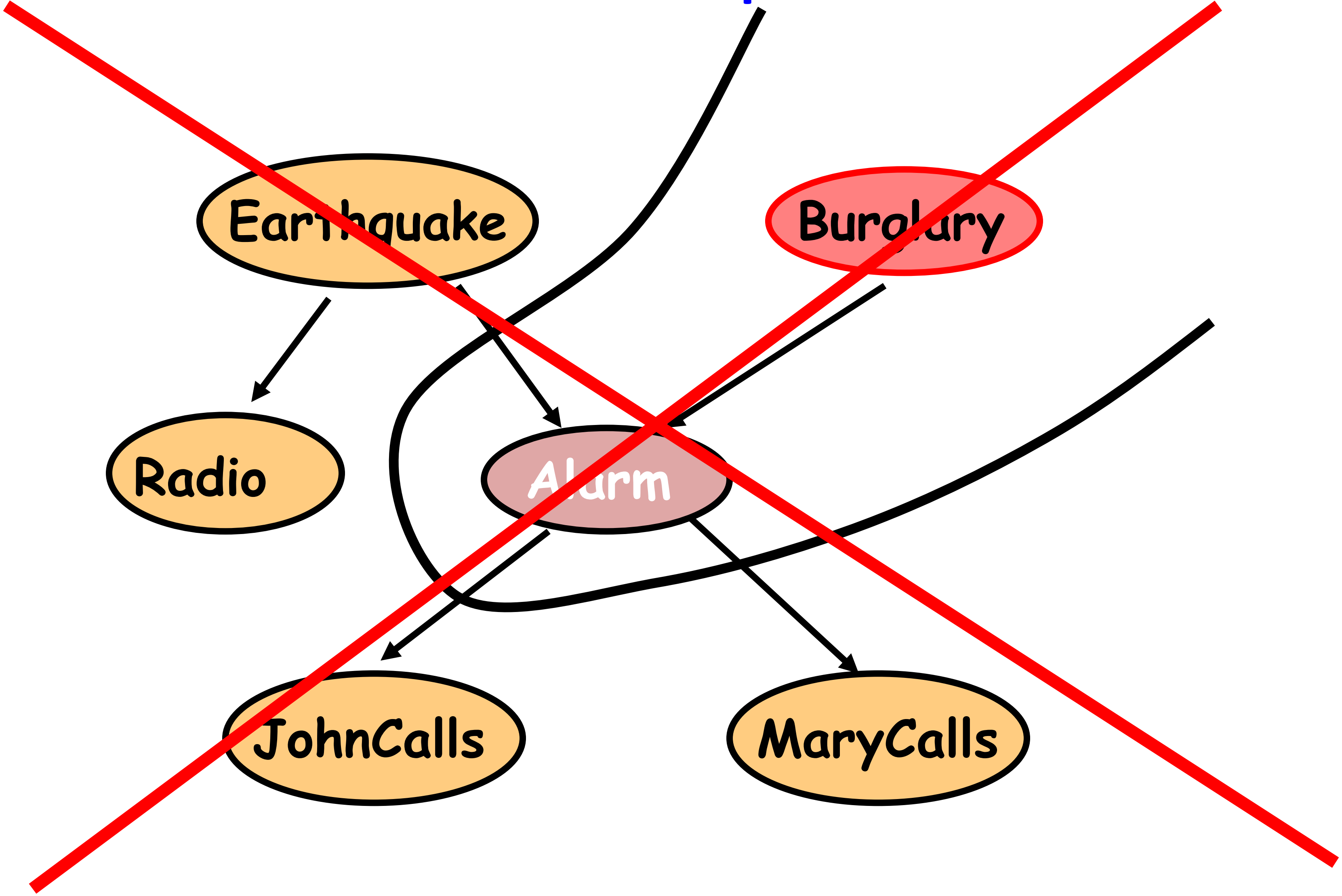


MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

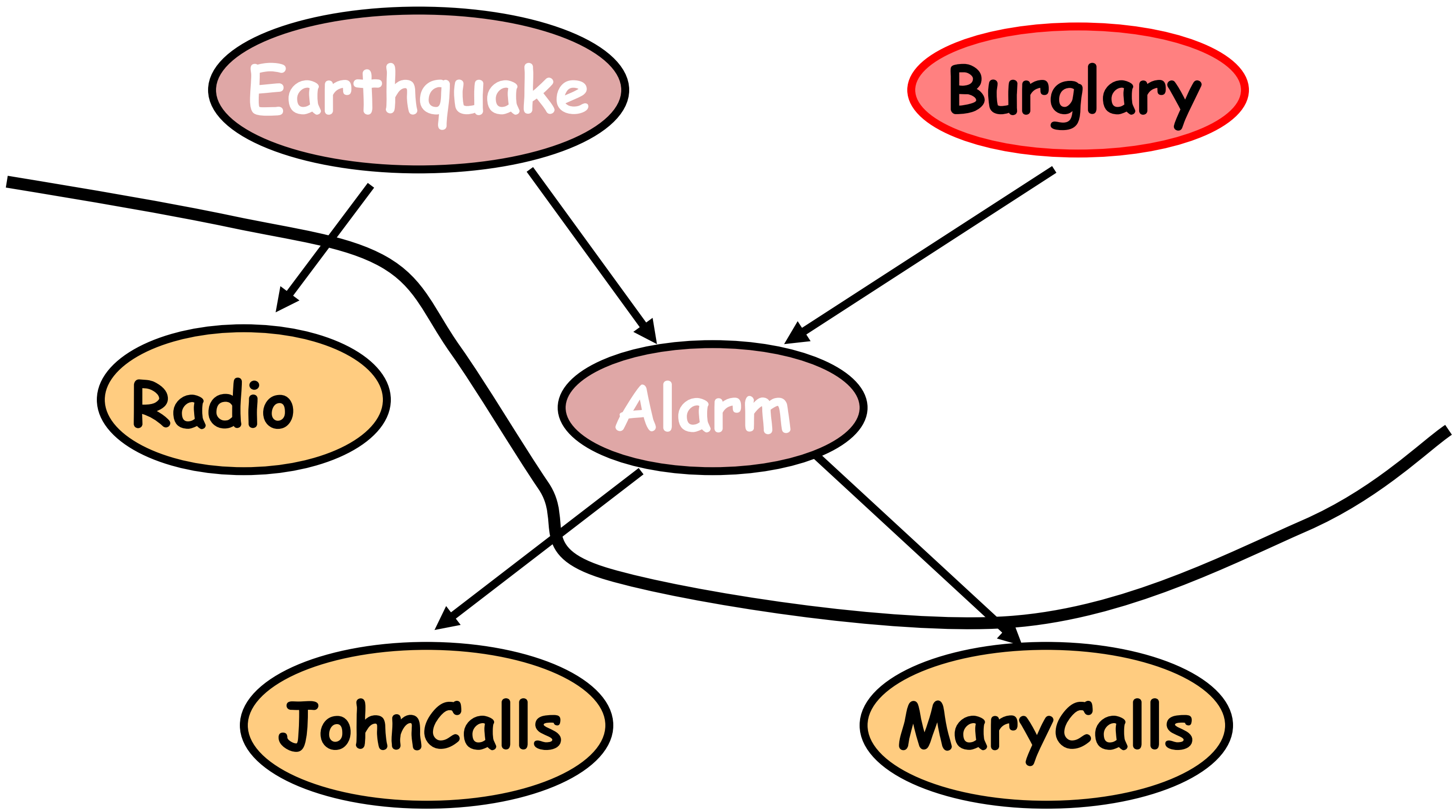
For Example



For Example



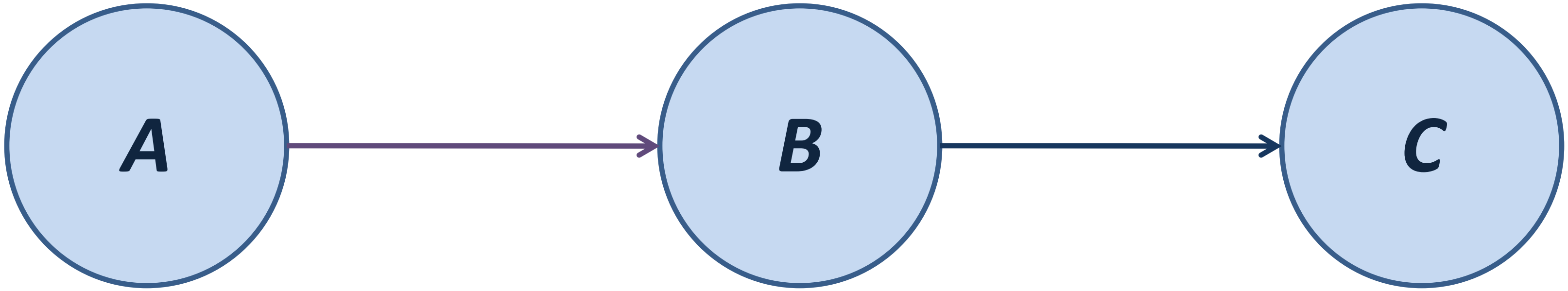
For Example



d-Separation

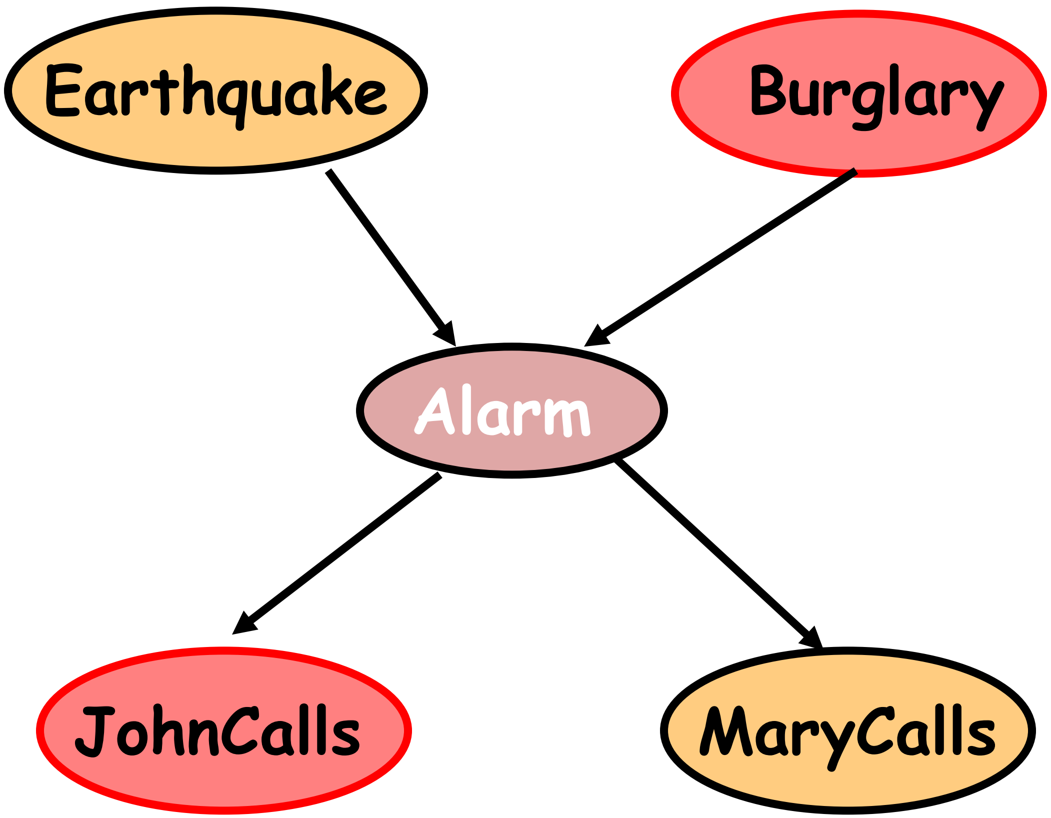
- An undirected path between two nodes is "cut off" if information cannot flow across one of the nodes in the path
- Two nodes are d-separated if every undirected path between them is cut off
- Two sets of nodes are d-separated if every pair of nodes, one from each set, is d-separated

d-Separation

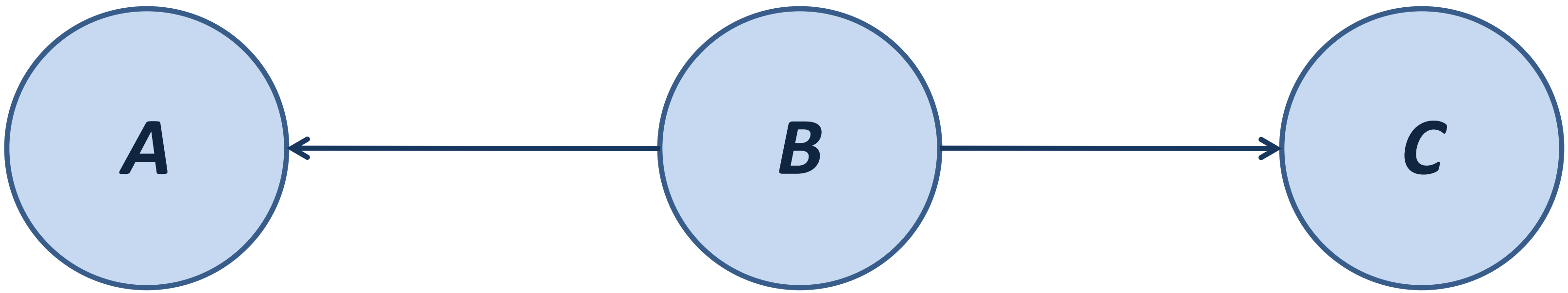


Linear connection: Information can flow between A and C if and only if we do not have evidence at B

For Example

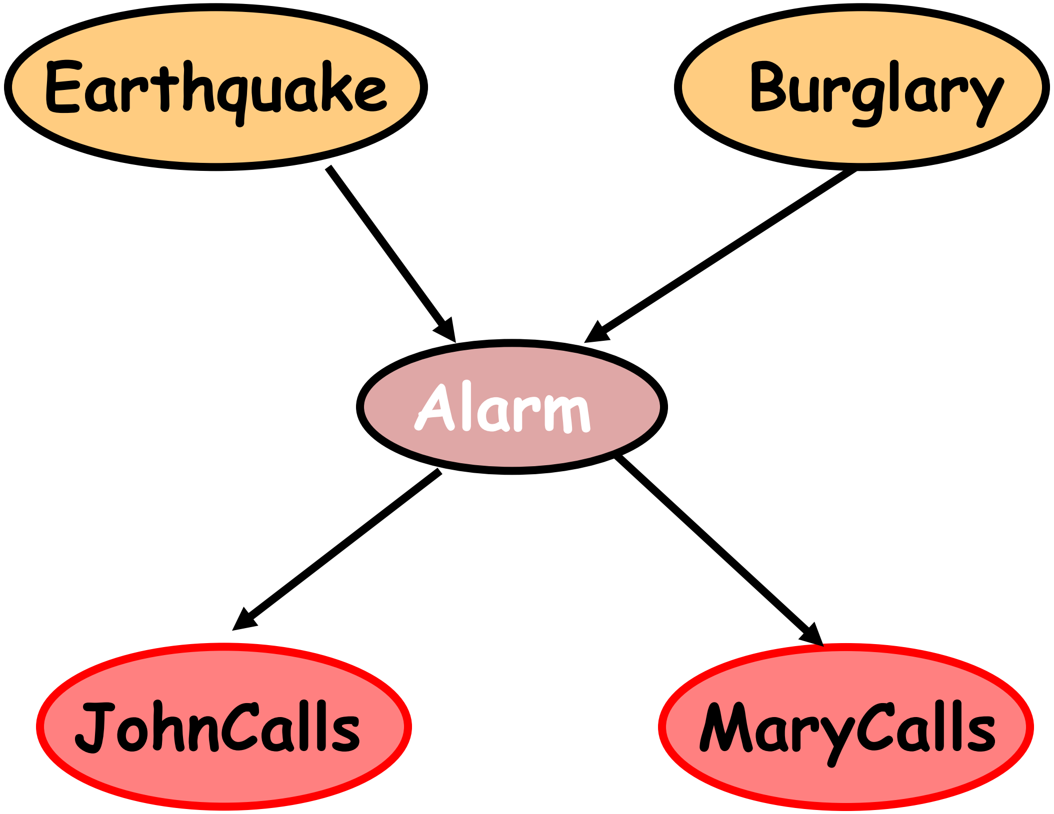


d-Separation (continued)

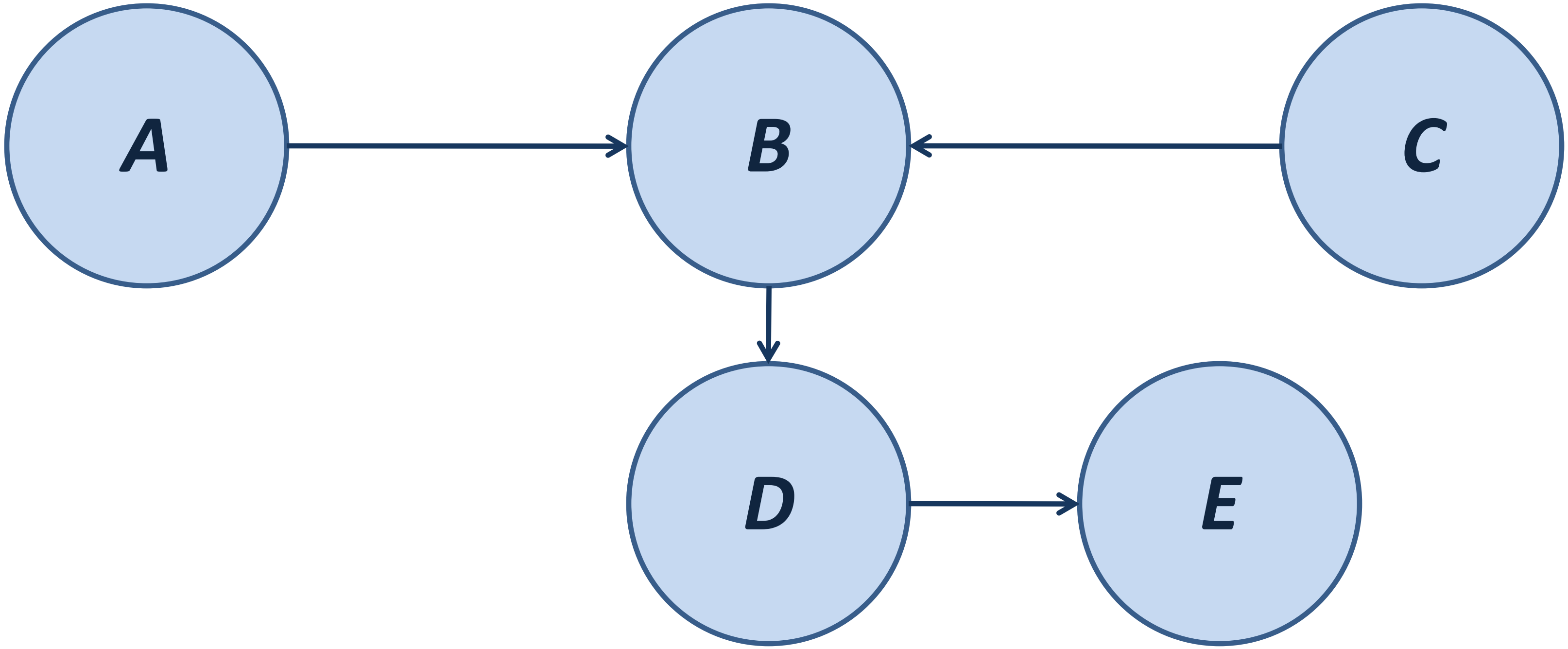


Diverging connection: Information can flow between A and C if and only if we do not have evidence at B

For Example

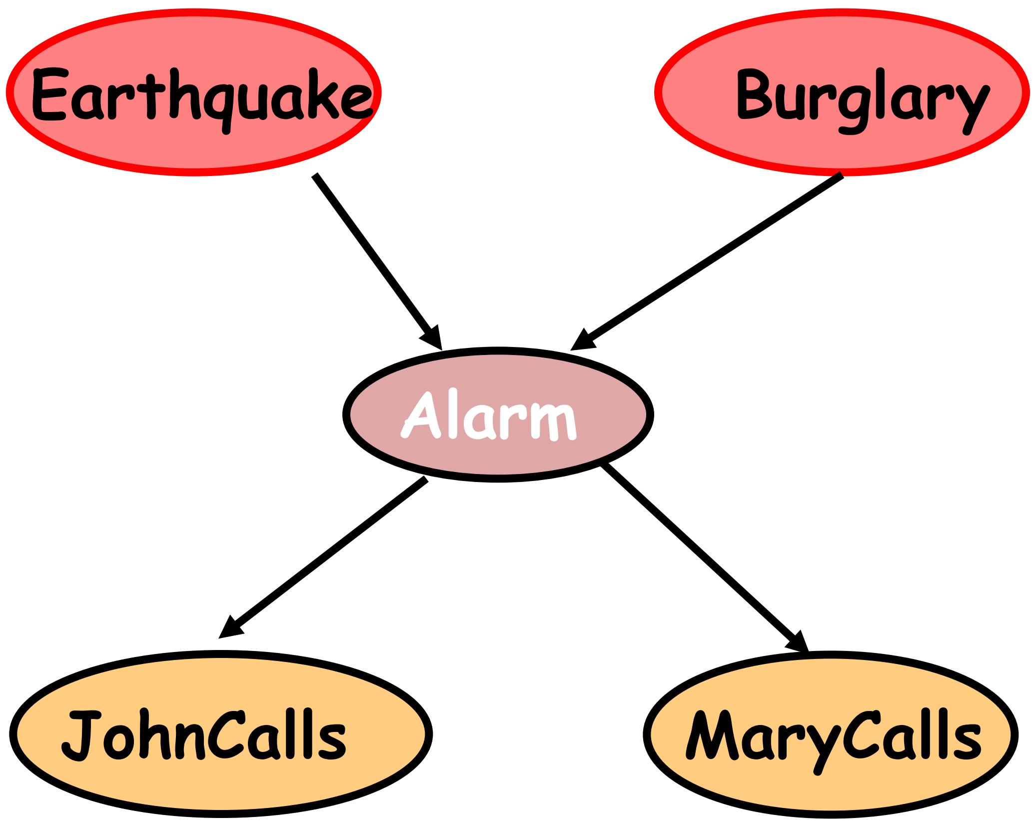


d-Separation (continued)

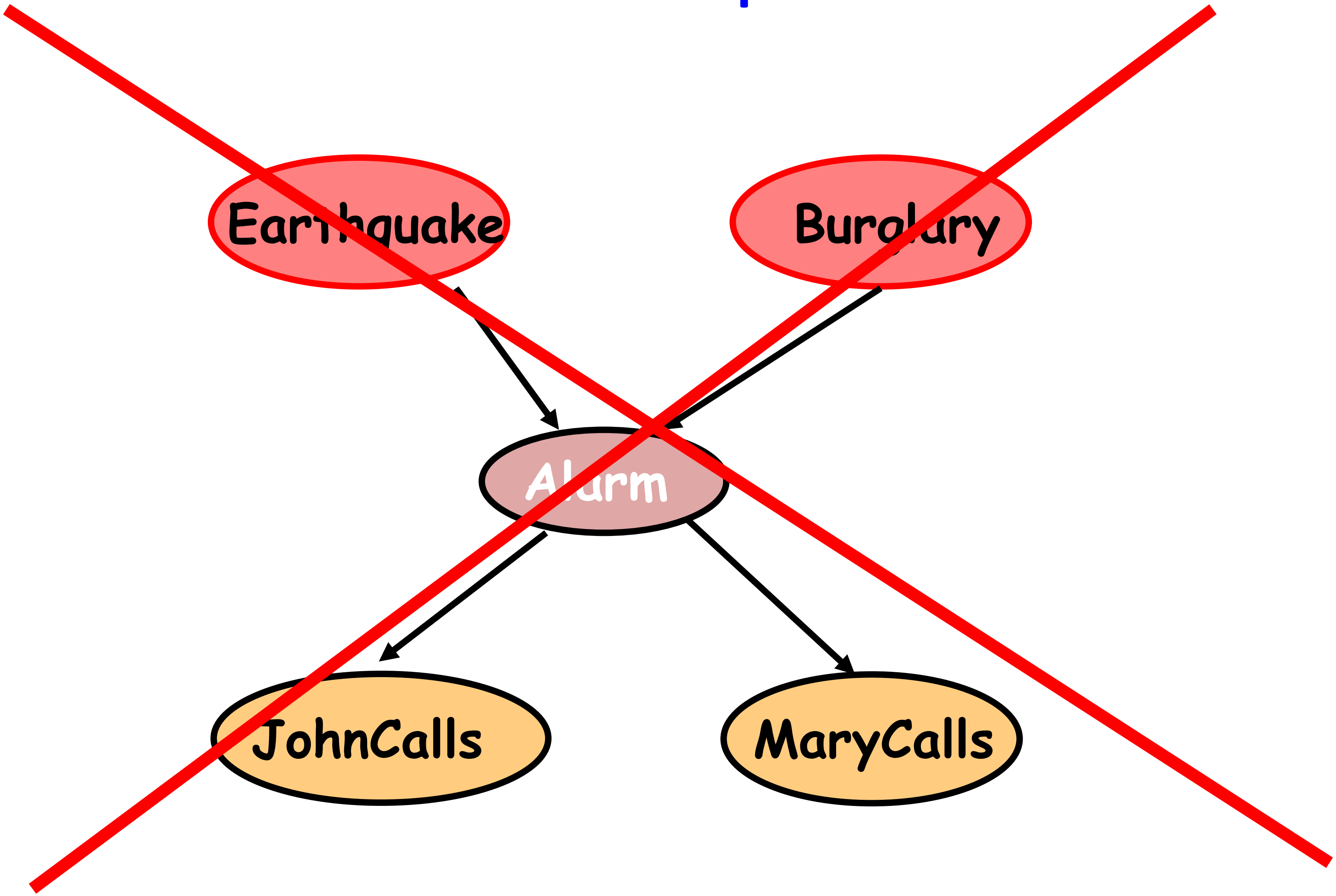


Converging connection: Information can flow between A and C if and only if we do have evidence at B or any descendent of B (such as D or E)

For Example

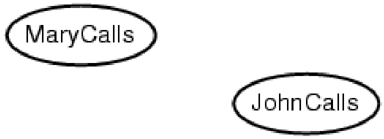


For Example



Bayes Net Construction Example

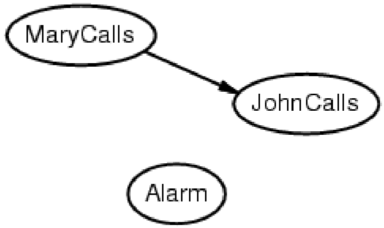
Suppose we choose the ordering *M, J, A, B, E*



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

Example

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



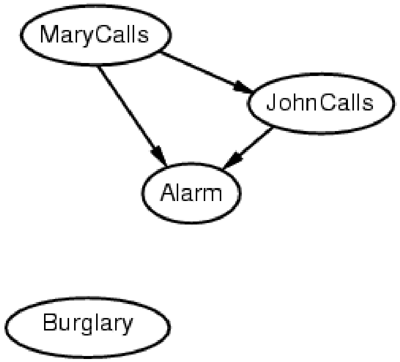
P*(J | M)* = ***P****(J)?*

No

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

Example

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



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

No

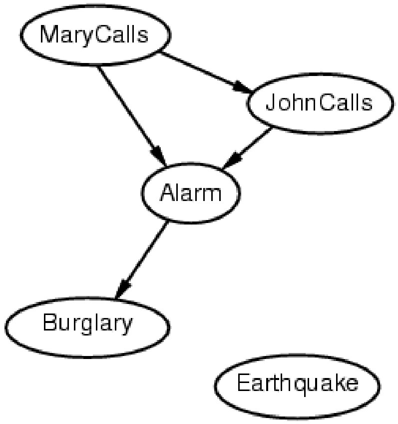
***P**(A | J, M) = **P**(A | J)*? ***P**(A | J, M) = **P**(A)*? **No**

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

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

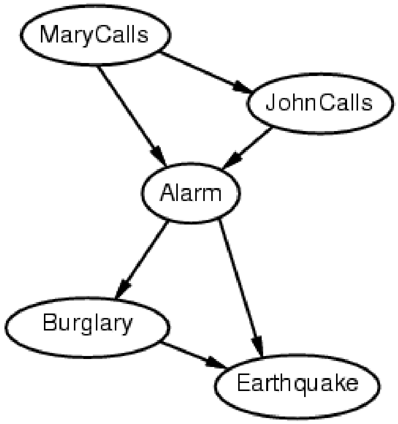
P*(B | A, J, M)* = ***P****(B)*? **No**

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**

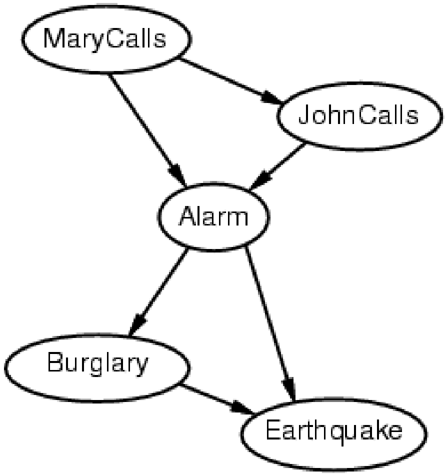
P*(B | A, J, M)* = ***P****(B | A)*? **Yes**

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.



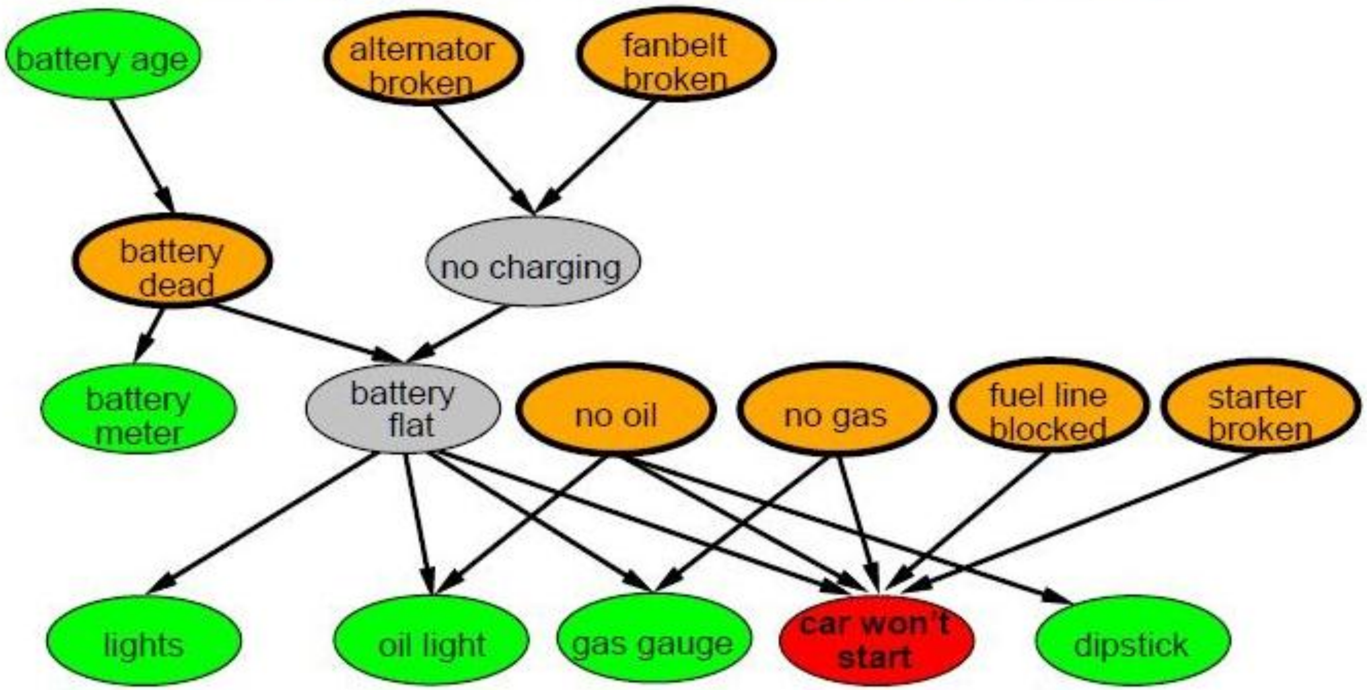
- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Example: Car Diagnosis

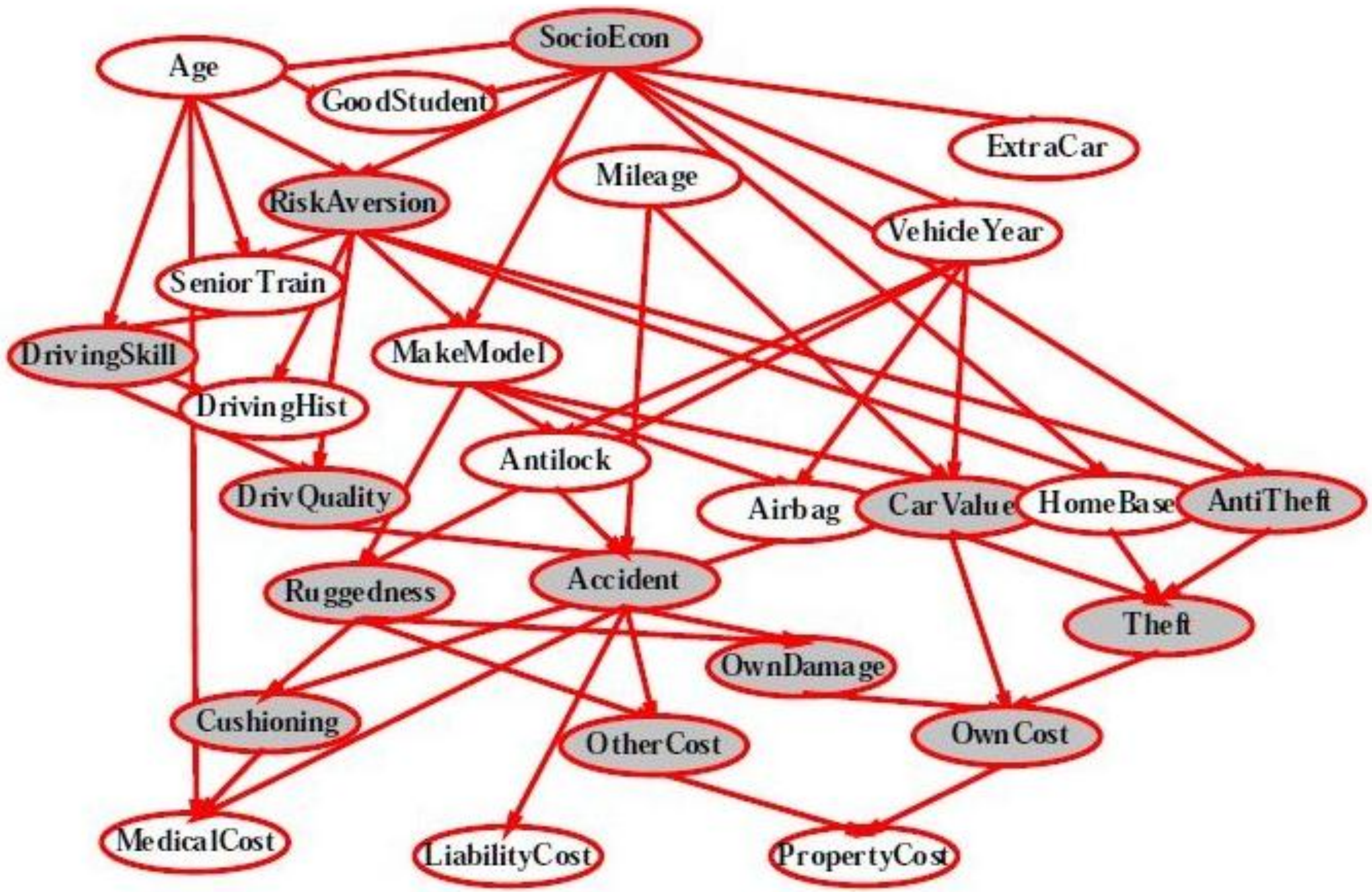
Initial evidence: car won't start

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
- Traffic Monitoring
- 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:
$X = f(Parents(X))$ for some function f

E.g., Boolean functions
$NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$

E.g., numerical relationships among continuous variables

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

Compact Conditionals

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)

2) Independent failure probability q_i for each cause alone

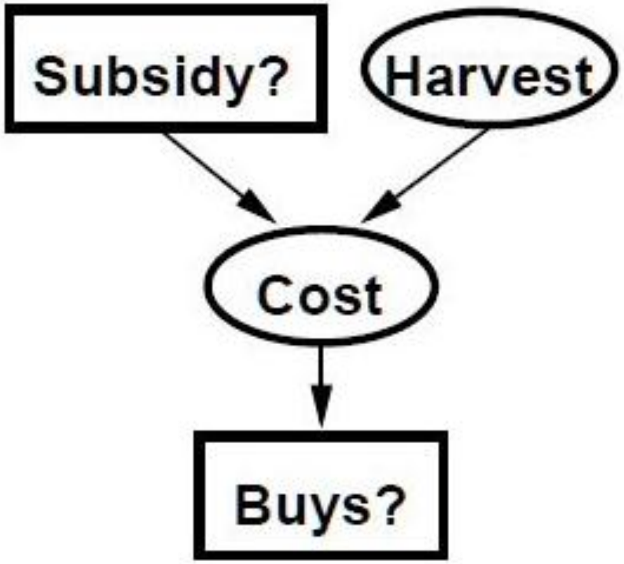
$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

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

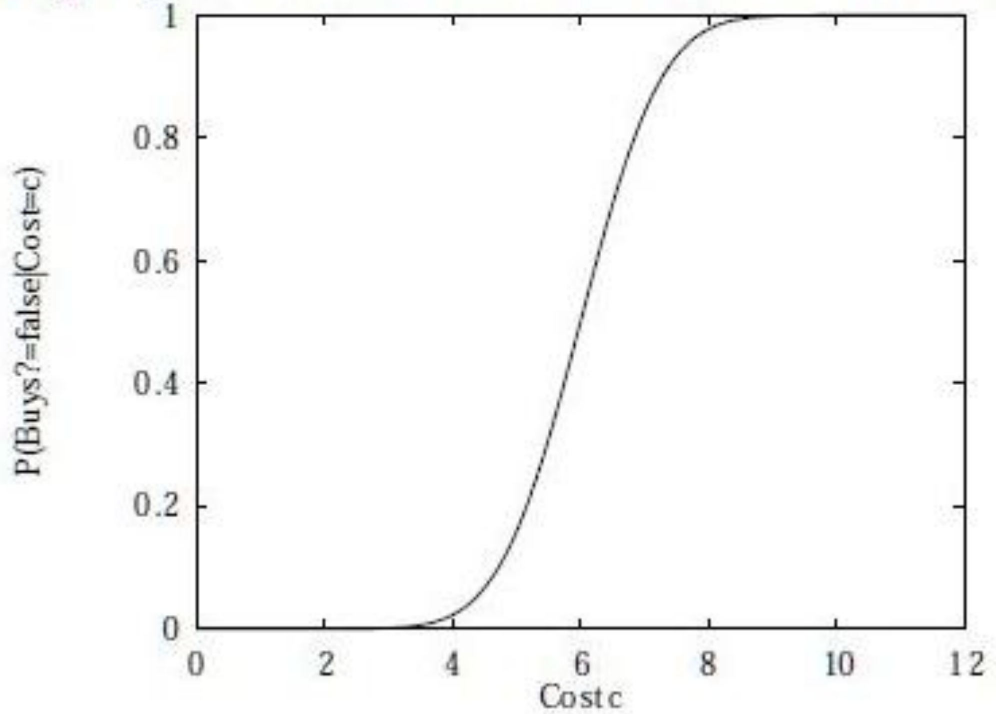
Need one conditional density function for child variable given continuous parents, for each possible assignment to discrete parents

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

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

#2 Discrete child – cont. parents

Probability of $Buys?$ given $Cost$ should be a "soft" threshold:



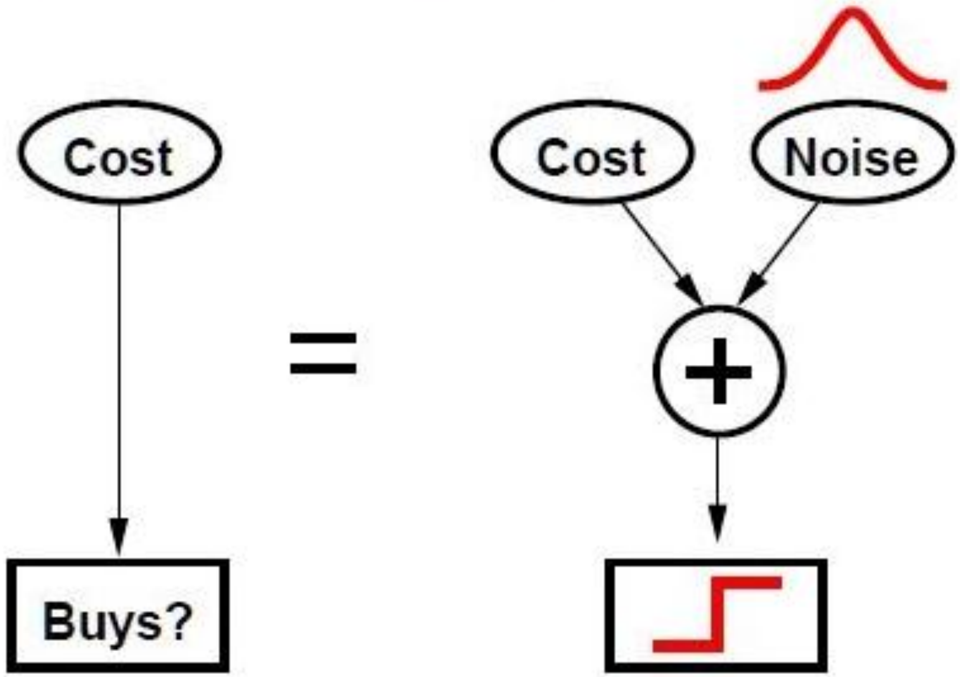
Probit distribution uses integral of Gaussian:

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

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

Why probit?

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

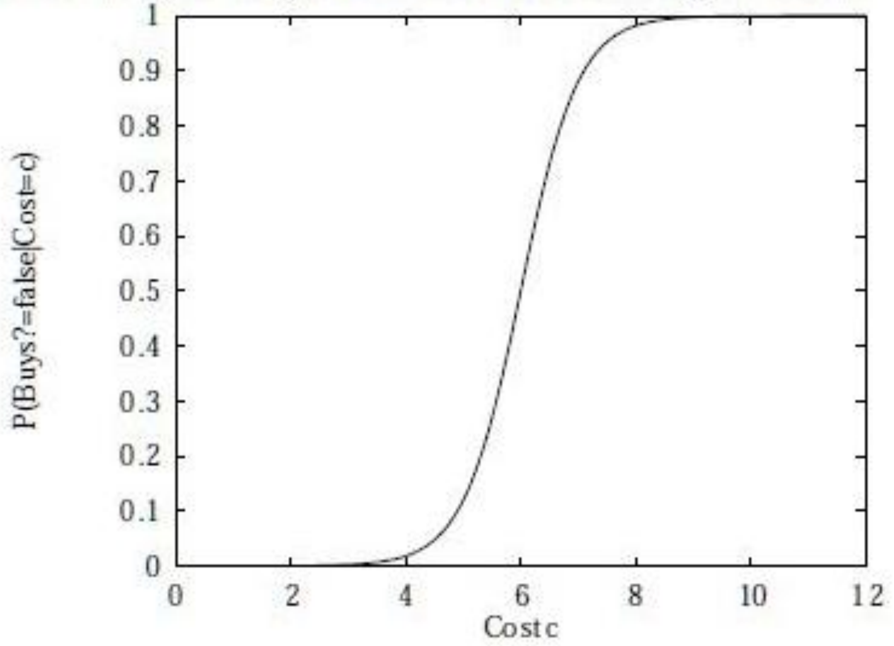


Sigmoid Function

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

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

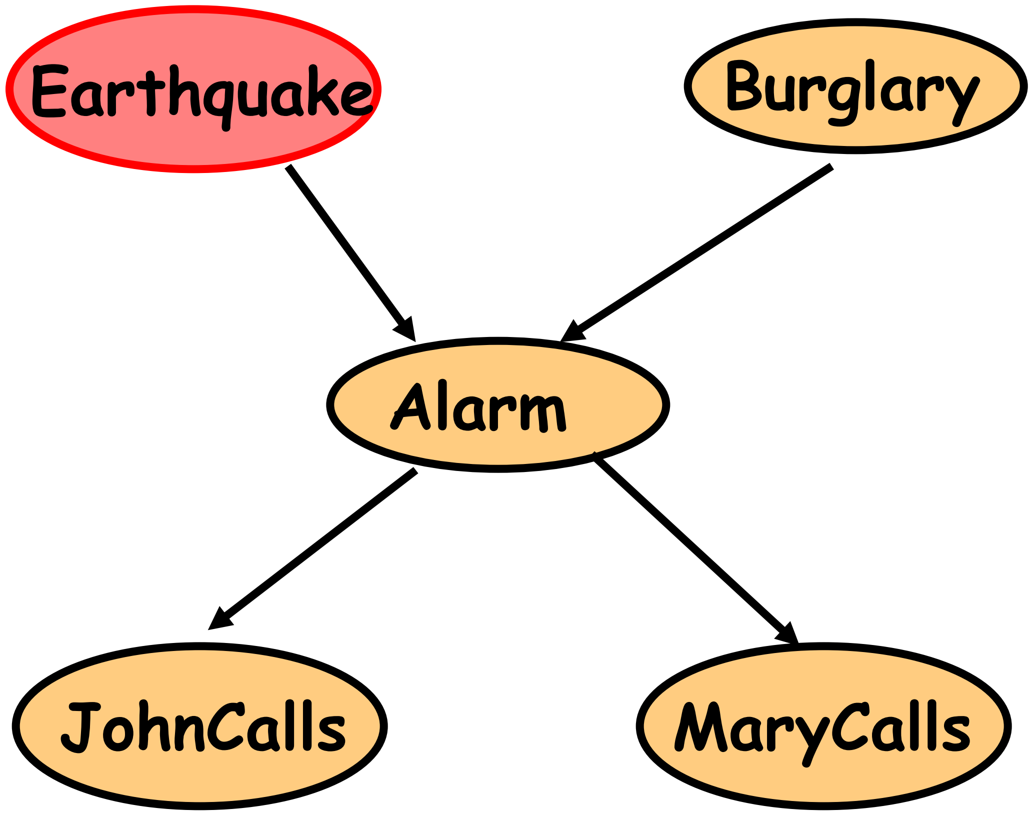
Sigmoid has similar shape to probit but much longer tails:



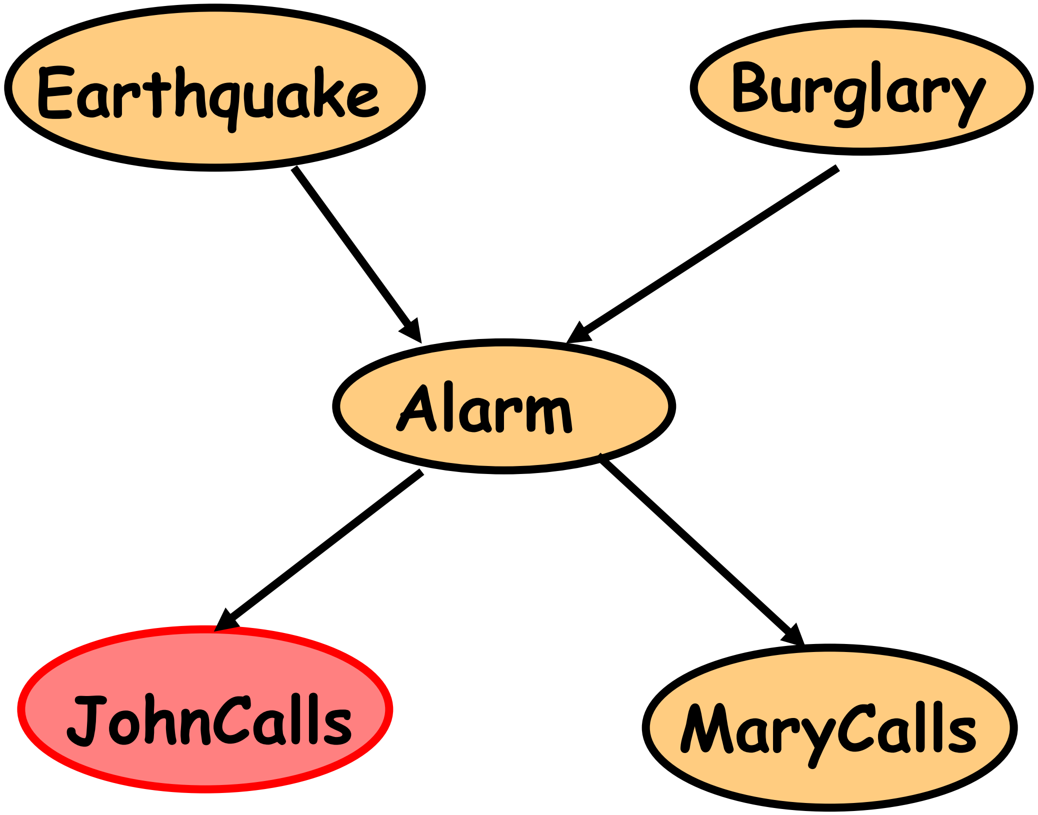
Inference in BNs

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

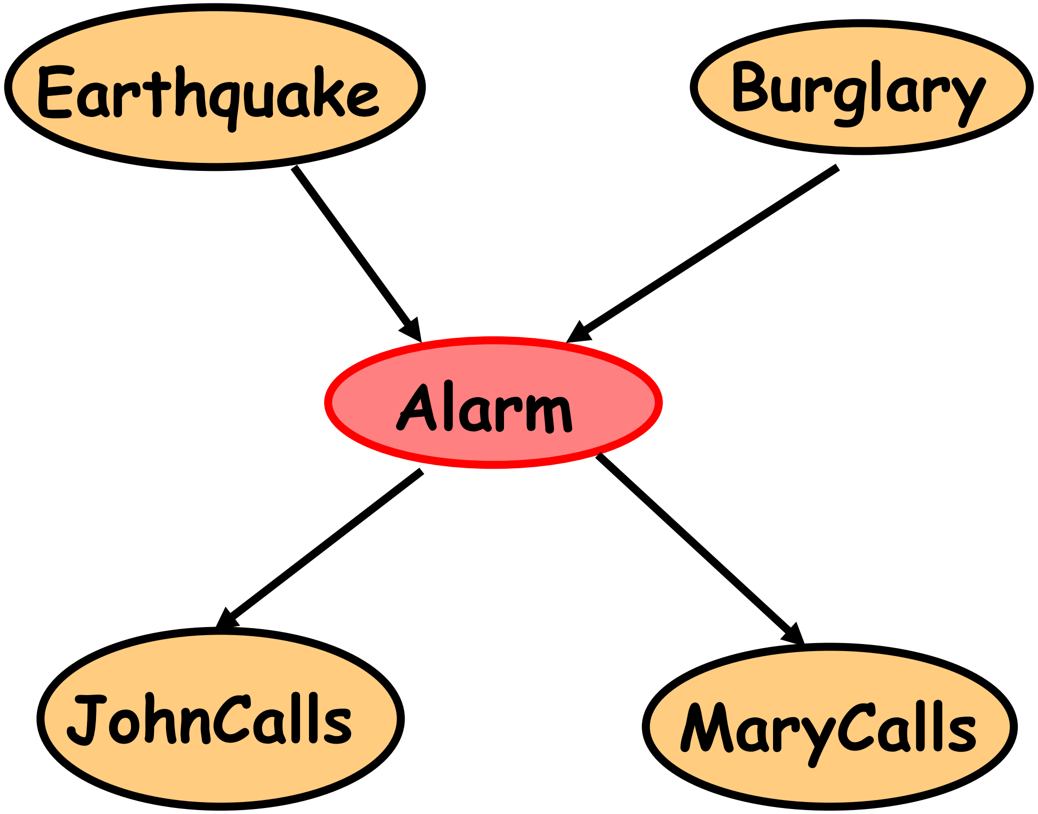
Causal Reasoning: P(j|e)



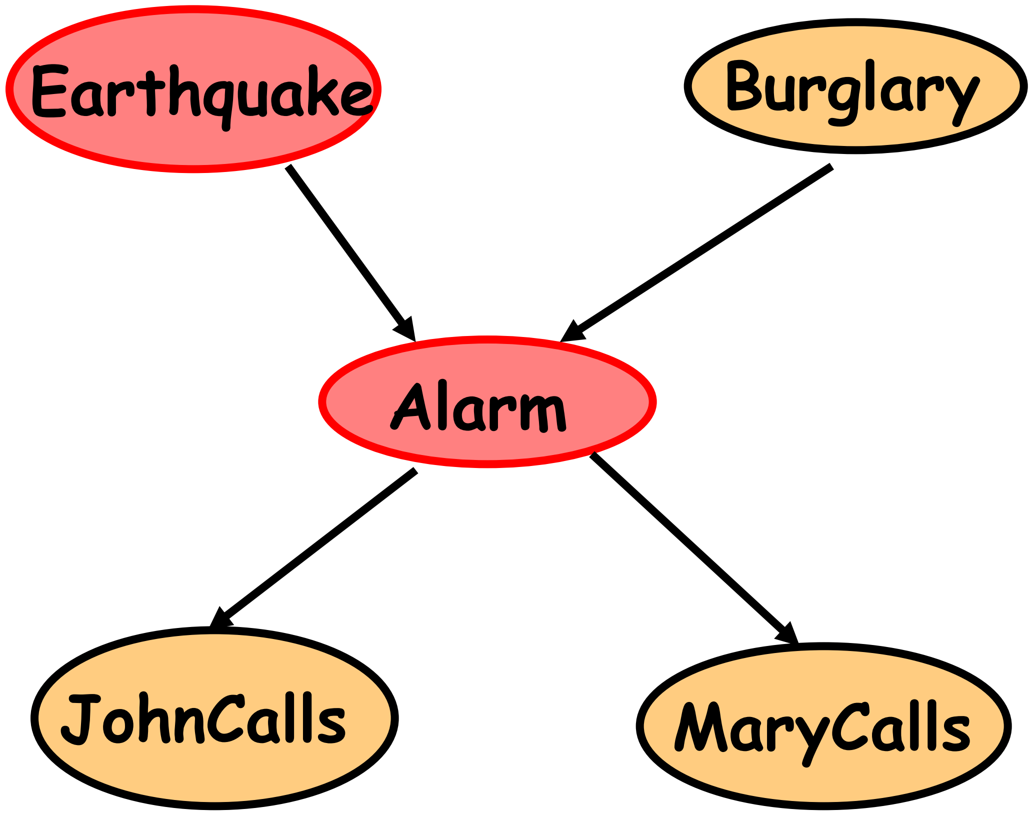
Evidential Reasoning: P(b|j)



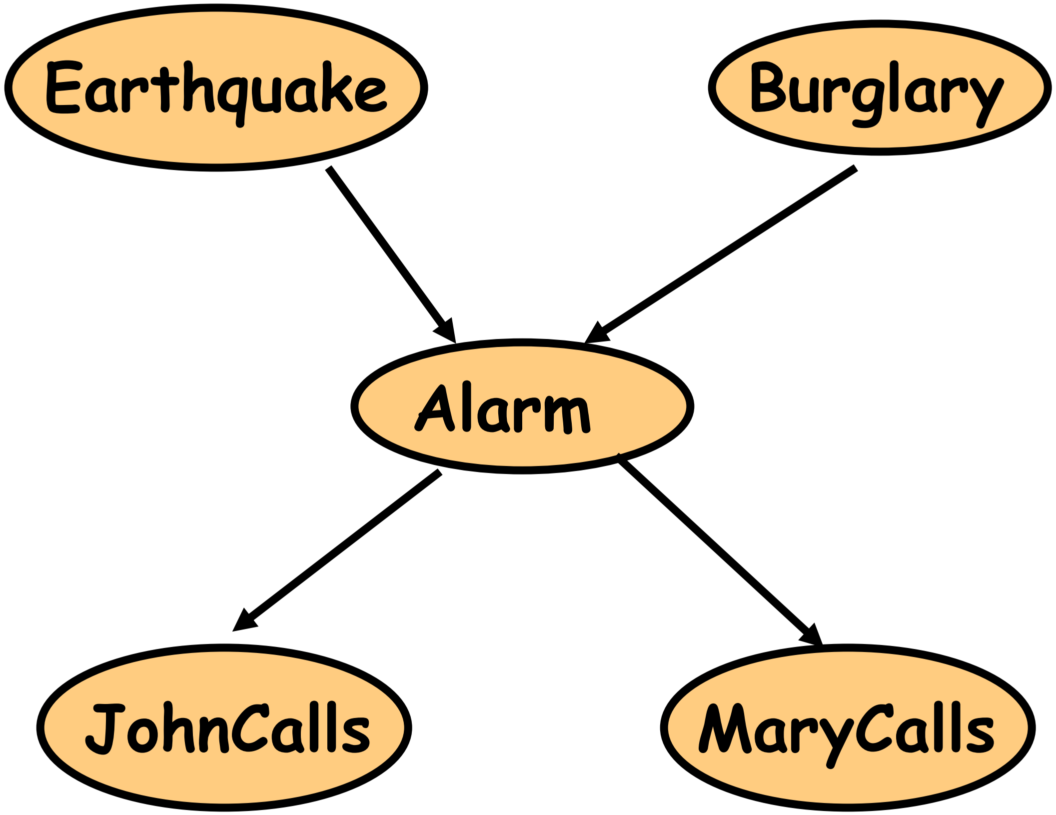
Intercausal Reasoning: P(e|a) vs P(b|a)



Intercausal Reasoning: P(b|a,e)

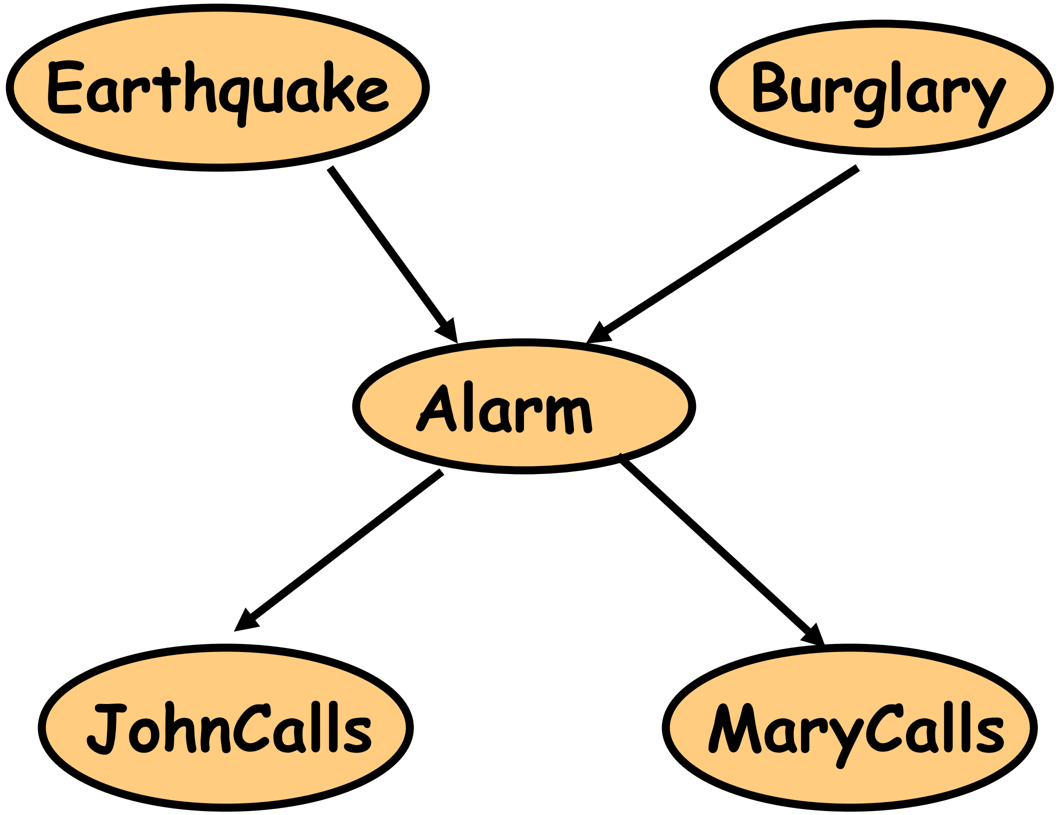


Inference Example: P(b|j,m)



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

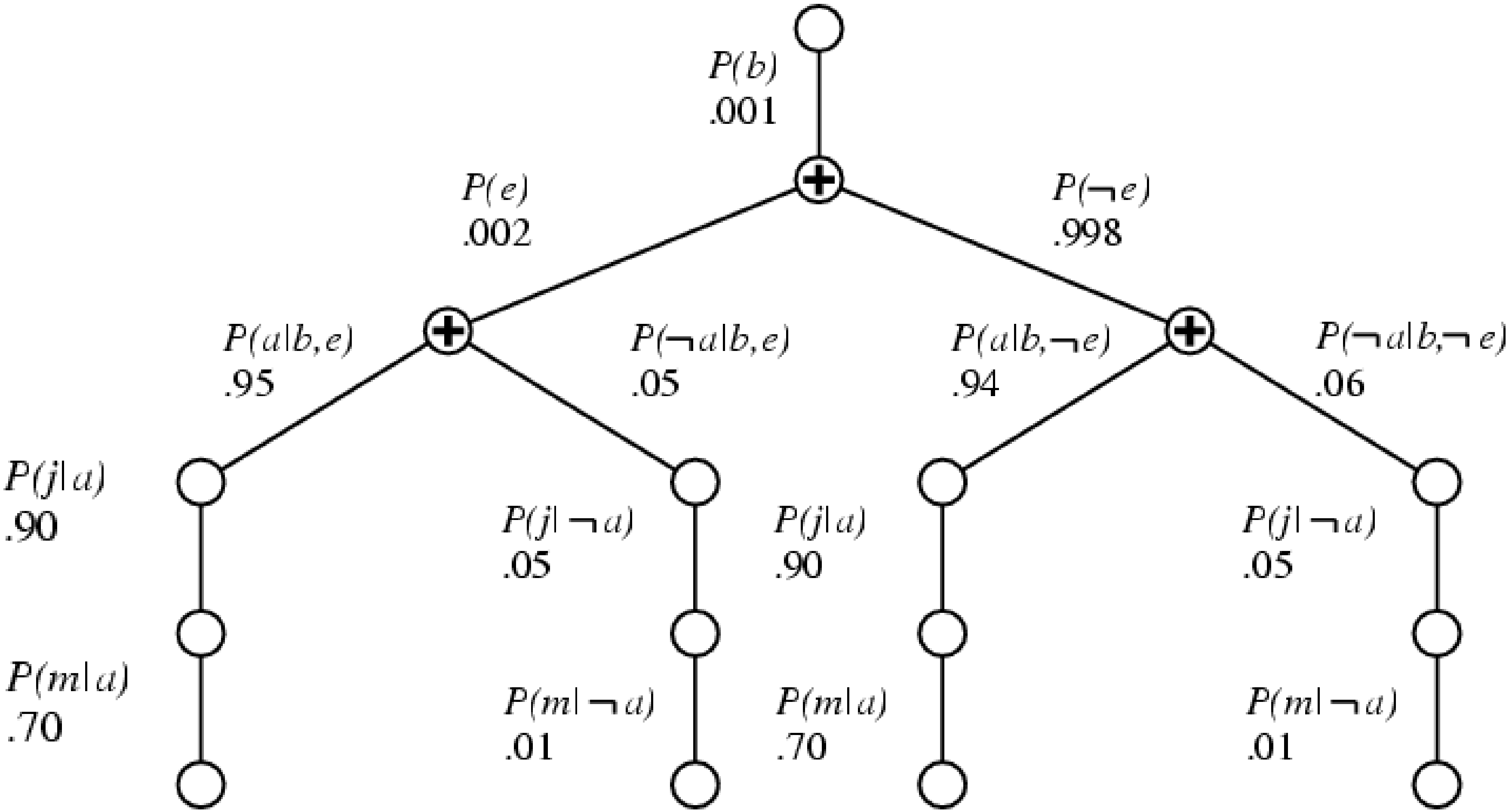
P(B | J=true, M=true)



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

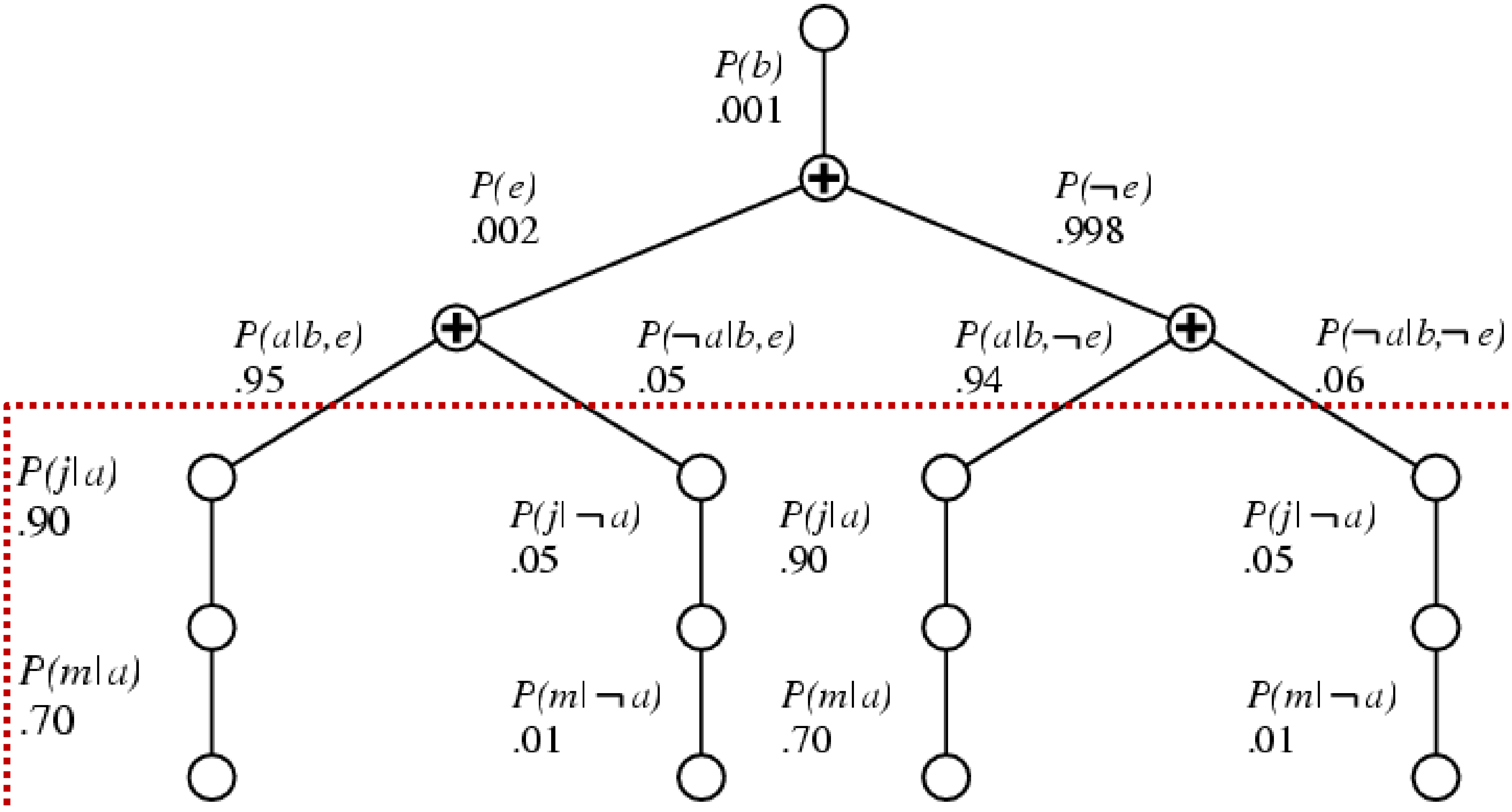
Variable Elimination

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



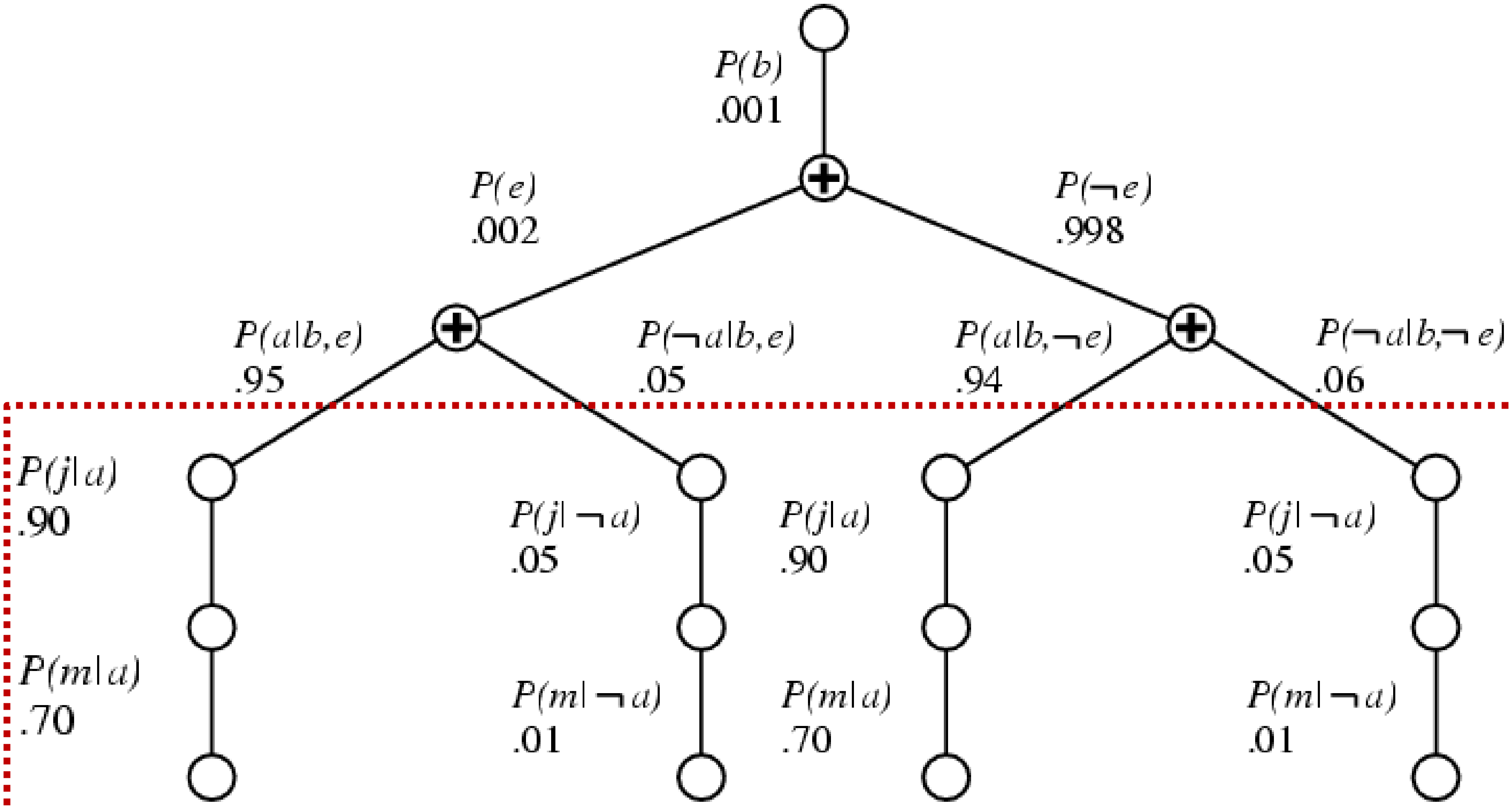
Variable Elimination

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



Variable Elimination

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



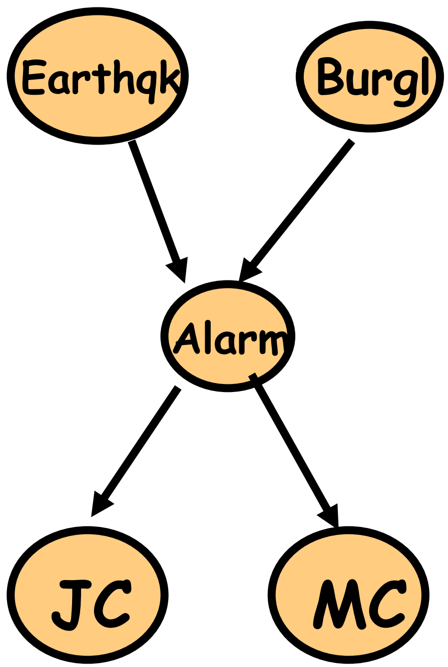
Repeated computations ➔ Dynamic Programming

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)

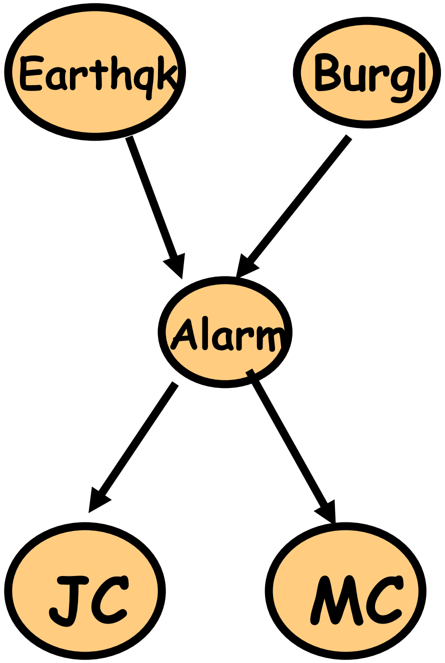
P(J)



Example of VE: P(JC)

P(J)

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

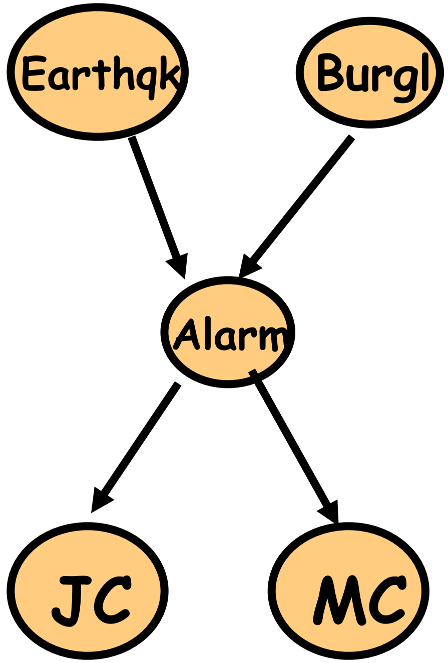


Example of VE: P(JC)

P(J)

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

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



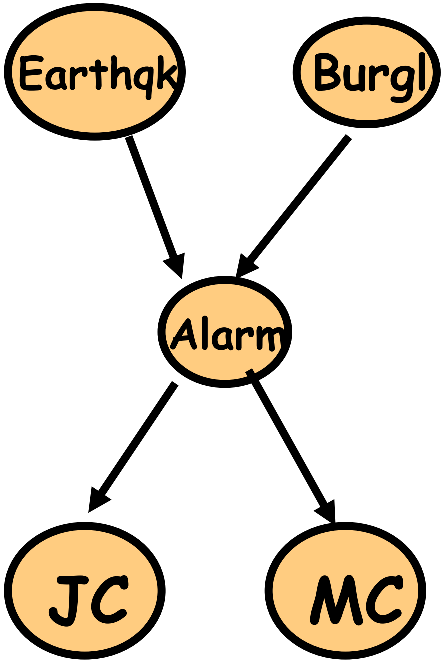
Example of VE: P(JC)

P(J)

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

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

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$



Example of VE: P(JC)

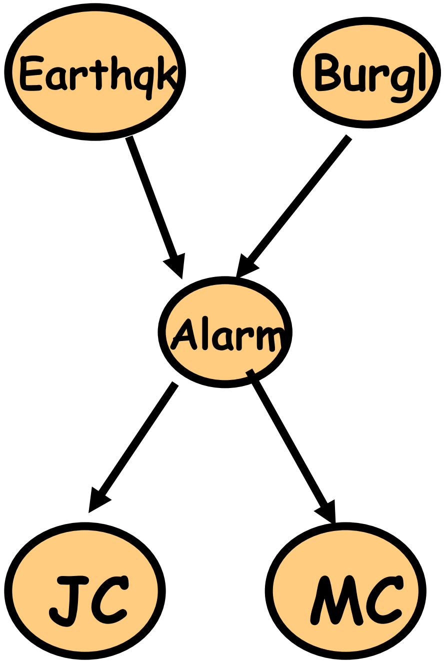
$P(J)$

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

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

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ f1(A,B)$



Example of VE: P(JC)

$P(J)$

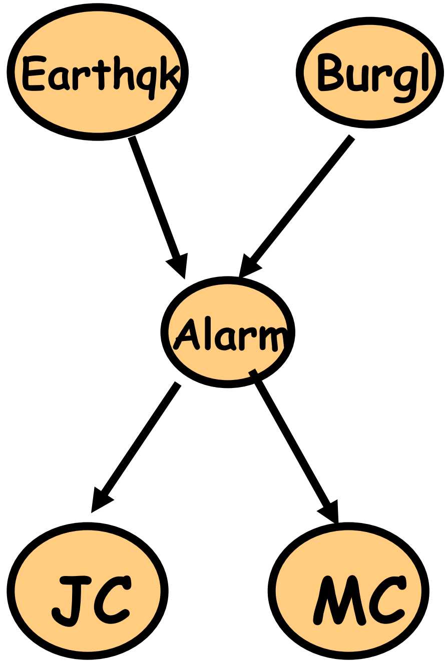
$= \Sigma_{M,A,B,E}\ P(J,M,A,B,E)$

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

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

$= \Sigma_{A}P(J|A)\ \Sigma_{M}P(M|A)\ \Sigma_{B}P(B)\ f1(A,B)$

$= \Sigma_{A}P(J|A)\ \Sigma_{M}P(M|A)\ f2(A)$



Example of VE: P(JC)

$P(J)$

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

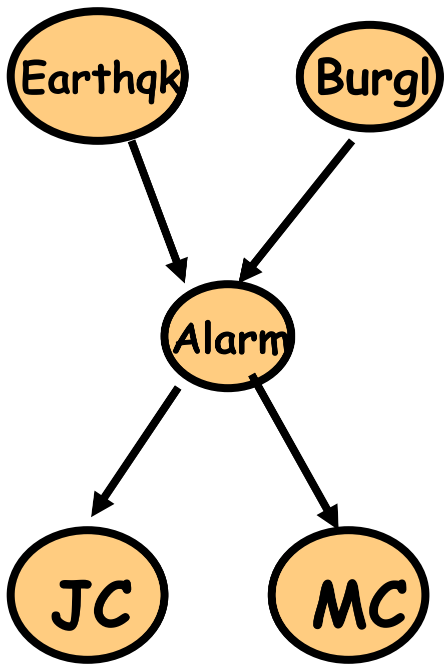
$= \Sigma_{M,A,B,E}\ P(J|A)P(M|A)\ P(B)P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ f1(A,B)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ f2(A)$

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



Example of VE: P(JC)

$P(J)$

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

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

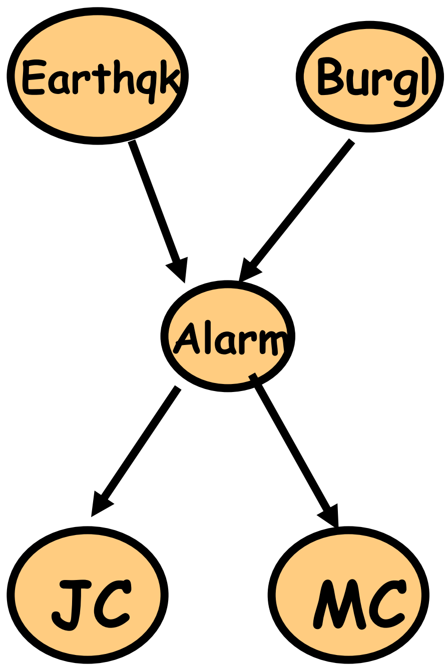
$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ f1(A,B)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ f2(A)$

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

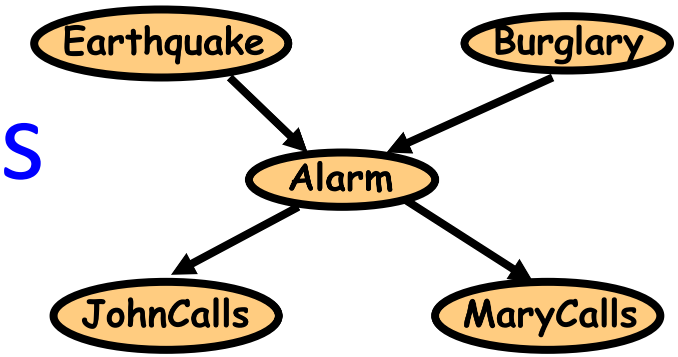
$= f4(J)$



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



$P(J)$

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

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

$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)\ \boxed{\Sigma_M P(M|A)}$

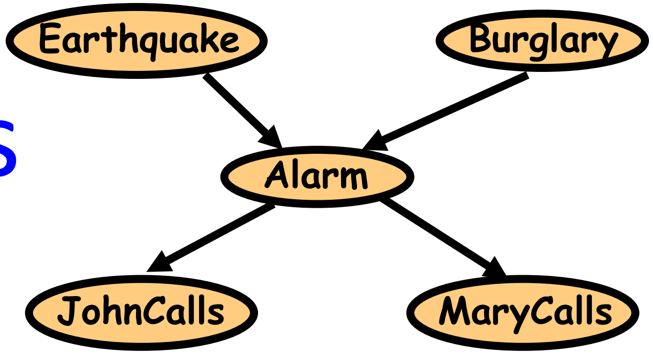
$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ f1(A,B)$

$= \Sigma_A P(J|A)\ f2(A)$

$= f3(J)$

Irrelevant variables



$P(J)$

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

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

$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)\ \boxed{\Sigma_M P(M|A)}$

$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_B P(B)\ f1(A,B)$

$= \Sigma_A P(J|A)\ f2(A)$

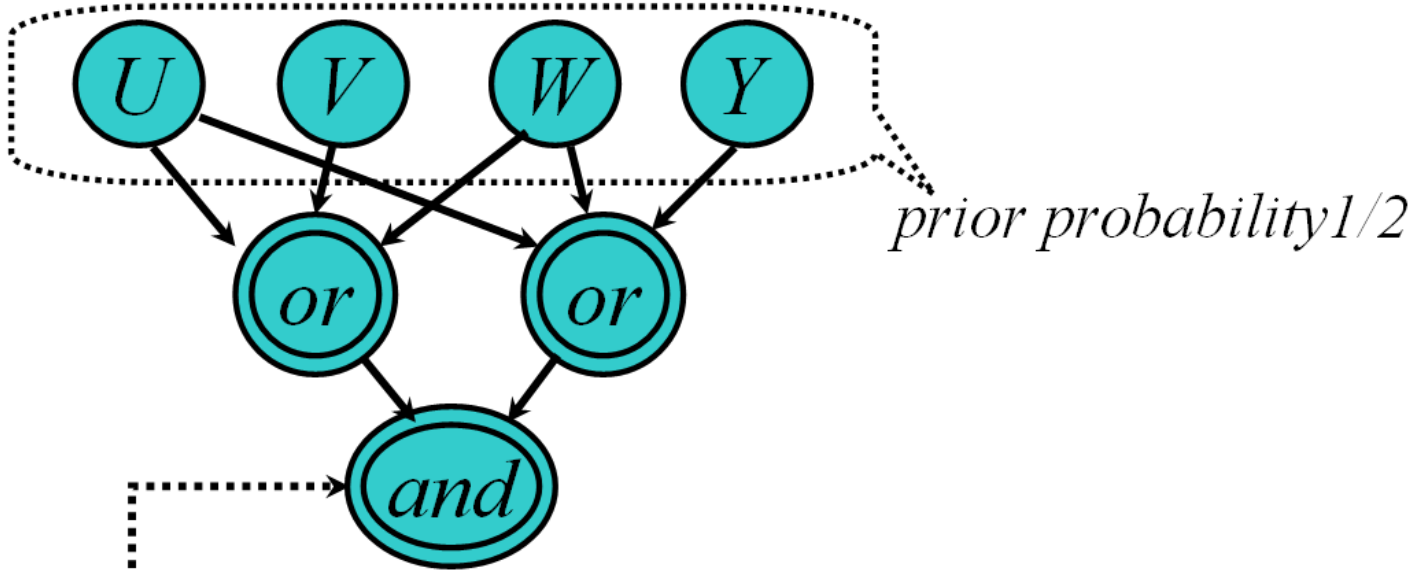
$= f3(J)$ M is irrelevant to the computation

Thm: Y is irrelevant unless Y $\in$ Ancestors(Z U 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* ϕ

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