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 => Alarm

Earth-Quake => Alarm

Alarm => John-calls

Alarm => Mary-calls

If there is a burglary, will Mary call?

Check KB & E |= M

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

Check KB & ~M *doesn't entail* ~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 earth-quake prone. Alarm goes off when there is an earth-quake.
- 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 |= M

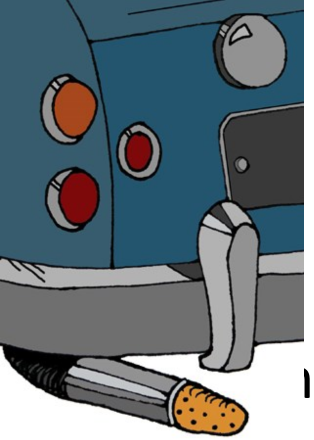
If Mary didn't call, is it possible that Burglary occurred?

Check KB & ~M *doesn't entail* ~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

- iscient & 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 => J

(also the exceptions may have interactions

c1&c5 => ~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?

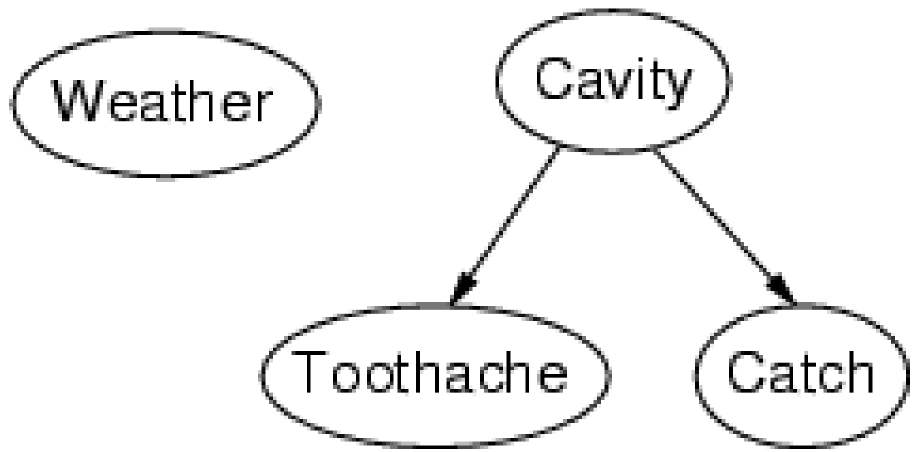
Qualification and Ramification problems make this an infeasible enterprise

Bayes Nets

- In general, joint distribution *P* over set of variables ($X_1 \times \ldots \times X_n$) 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:



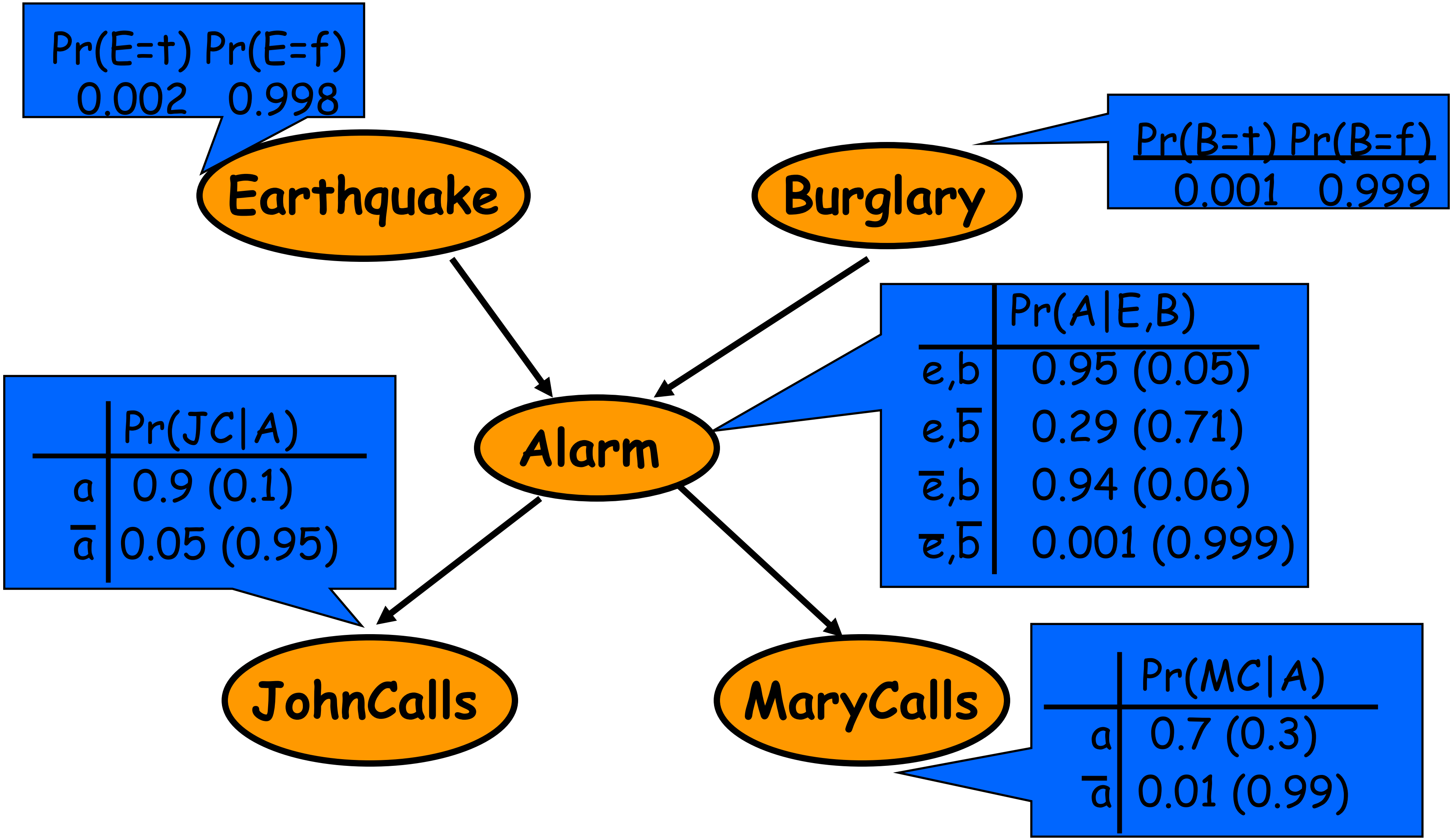
Weather is independent of the other variables

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



Pr(E=t)	Pr(E=f)
0.002	0.998

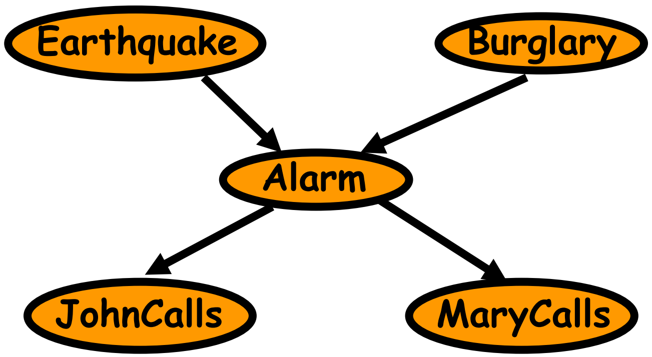
Pr(B=t)	Pr(B=f)
0.001	0.999

	Pr(A|E,B)
e,b	0.95 (0.05)
e,$\bar{b}$	0.29 (0.71)
$\bar{e}$,b	0.94 (0.06)
$\bar{e}$,$\bar{b}$	0.001 (0.999)

	Pr(JC|A)
a	0.9 (0.1)
$\bar{a}$	0.05 (0.95)

	Pr(MC|A)
a	0.7 (0.3)
$\bar{a}$	0.01 (0.99)

Earthquake Example (cont'd)



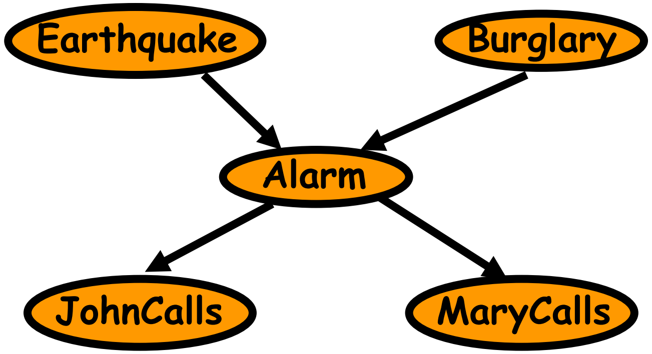
- **If we know *Alarm*, no other evidence influences our degree of belief in *JohnCalls***
  - $P(JC|MC,A,E,B) = P(JC|A)$
  - also: $P(MC|JC,A,E,B) = P(MC|A)$ and $P(E|B) = P(E)$
- **By the chain rule we have**

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

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

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

Earthquake Example (Global Semantics)



- **We just proved**

  P(JC,MC,A,E,B) = ***P(JC|A) ·P(MC|A) ·P(A|B,E) ·P(E) ·P(B)***

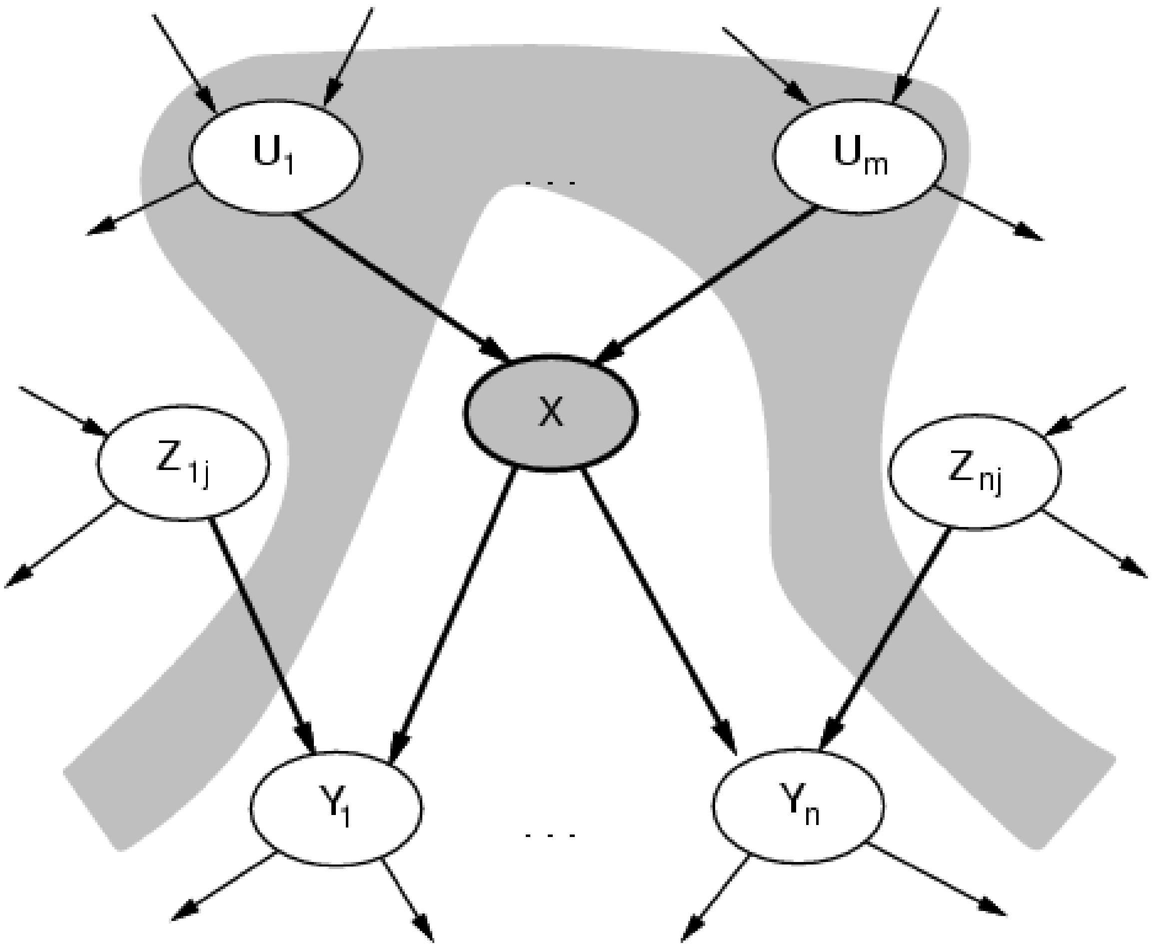
- **In general full joint distribution of a Bayes net is defined as**

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

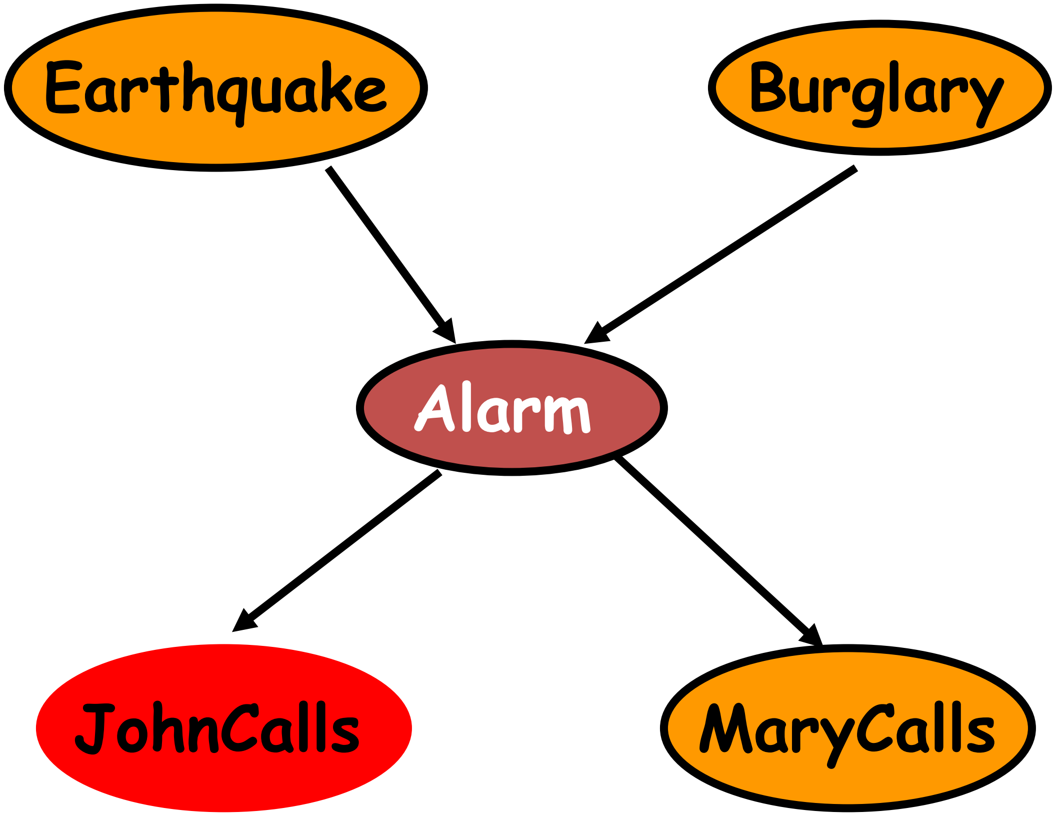
BNs: Qualitative Structure

- 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"

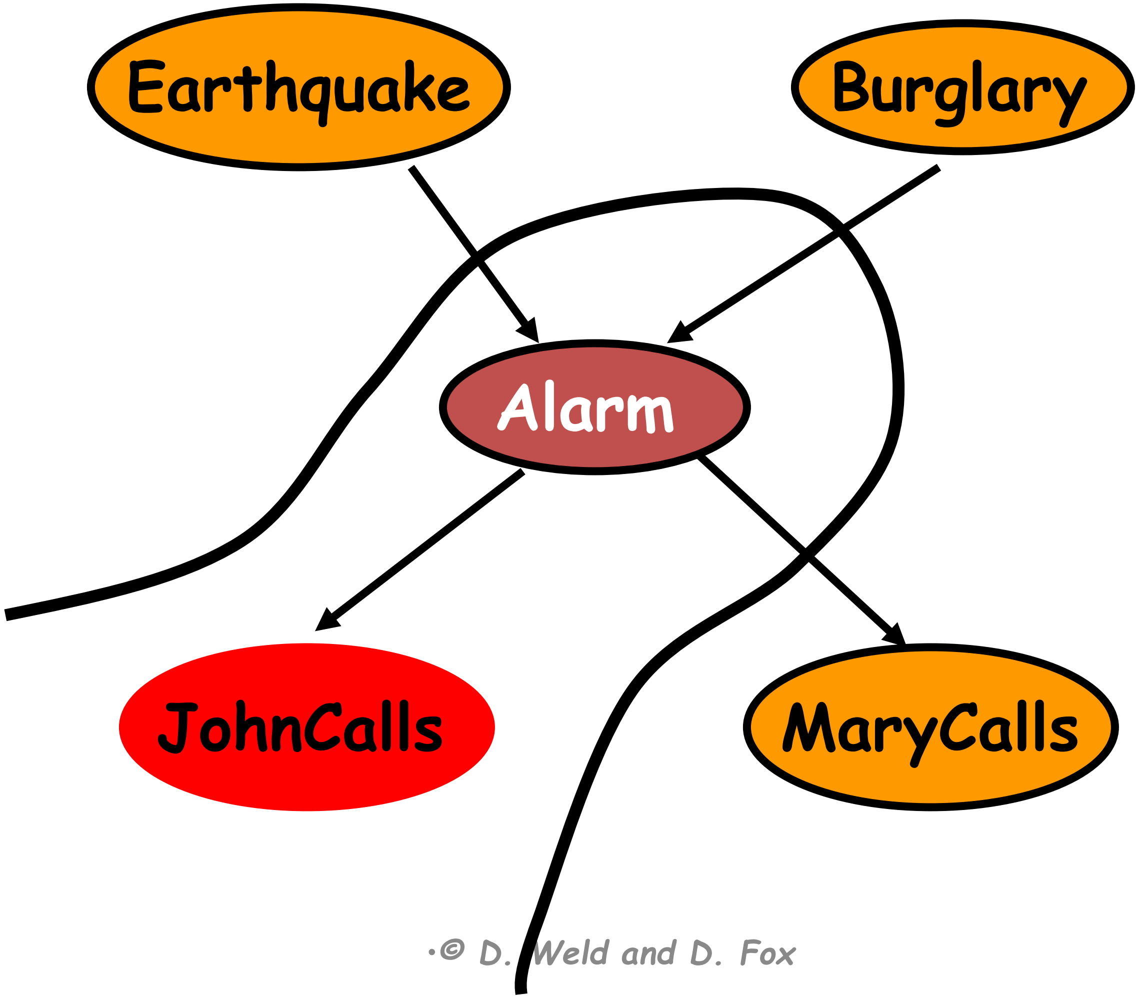
Given Parents, X is Independent of Non-Descendants



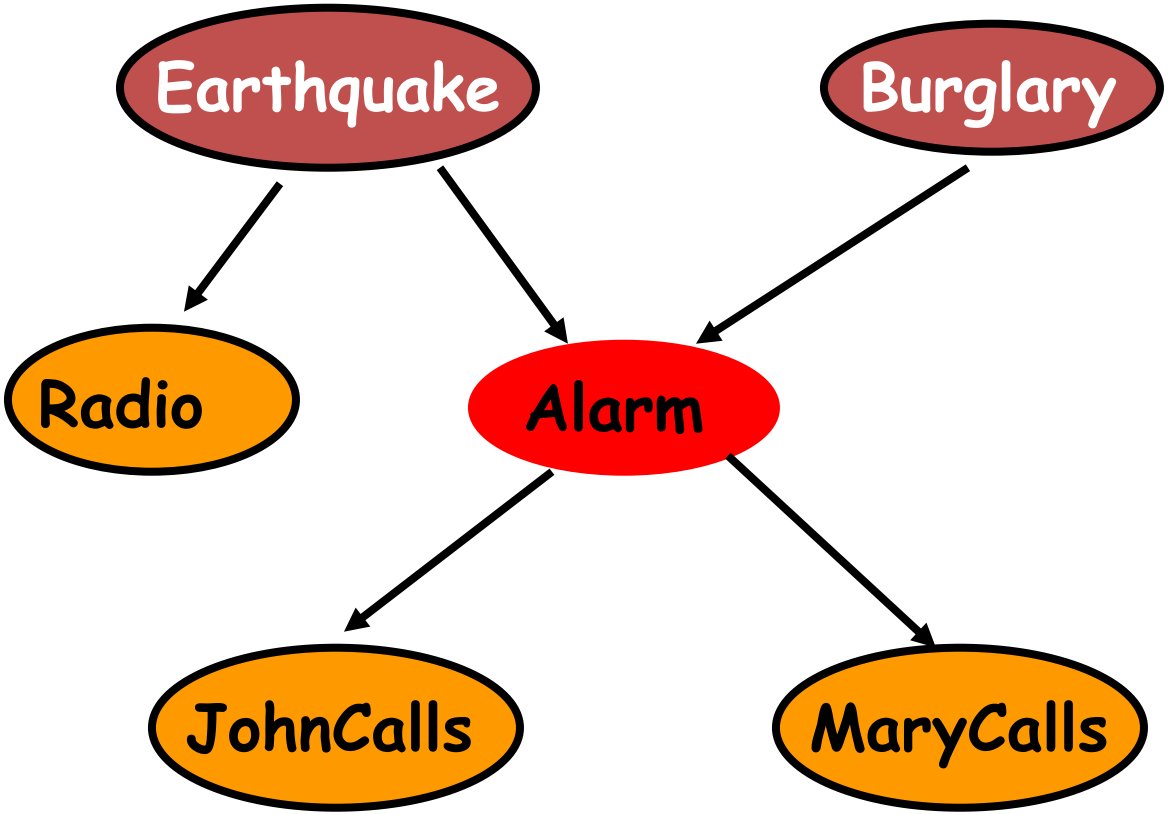
Examples



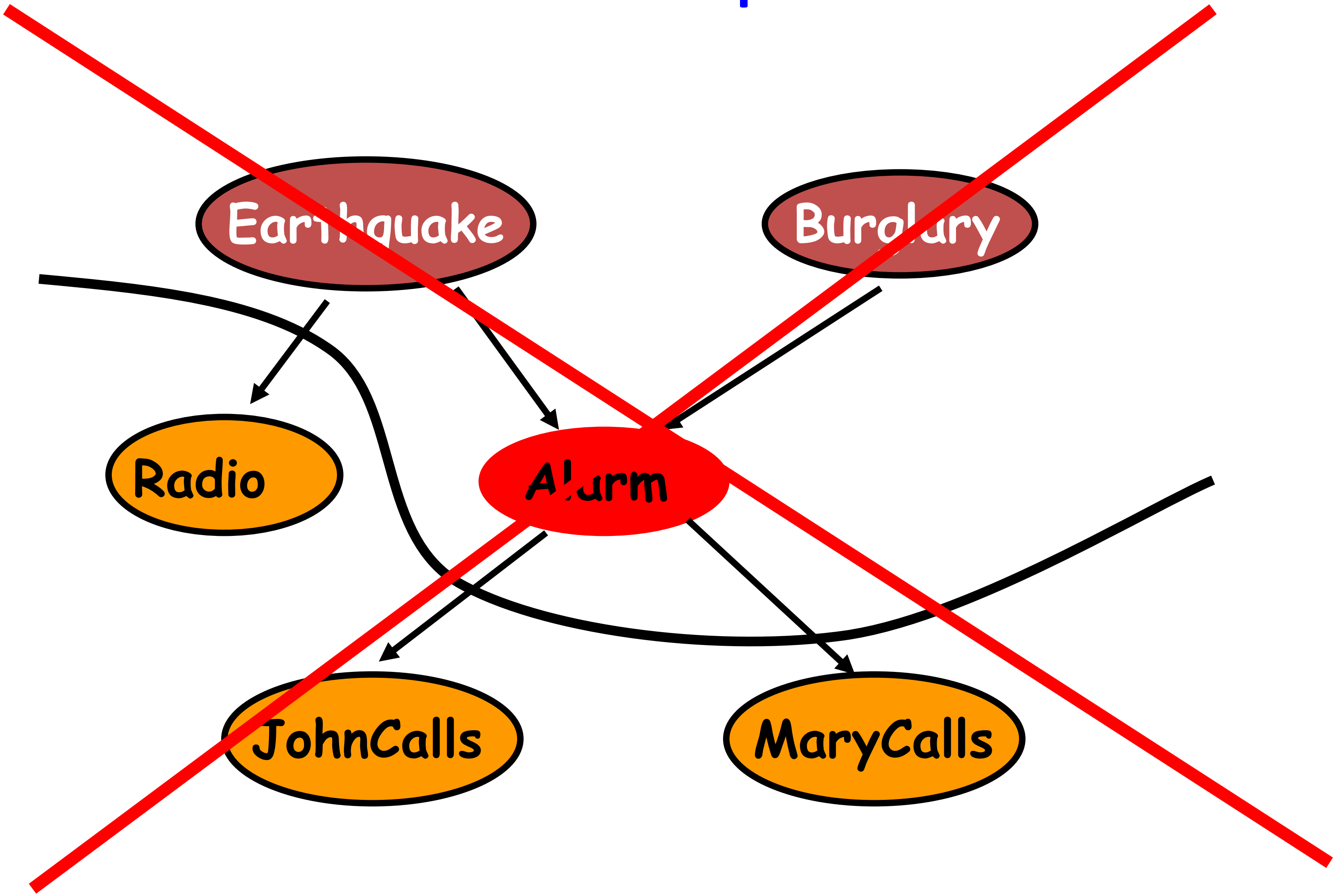
For Example



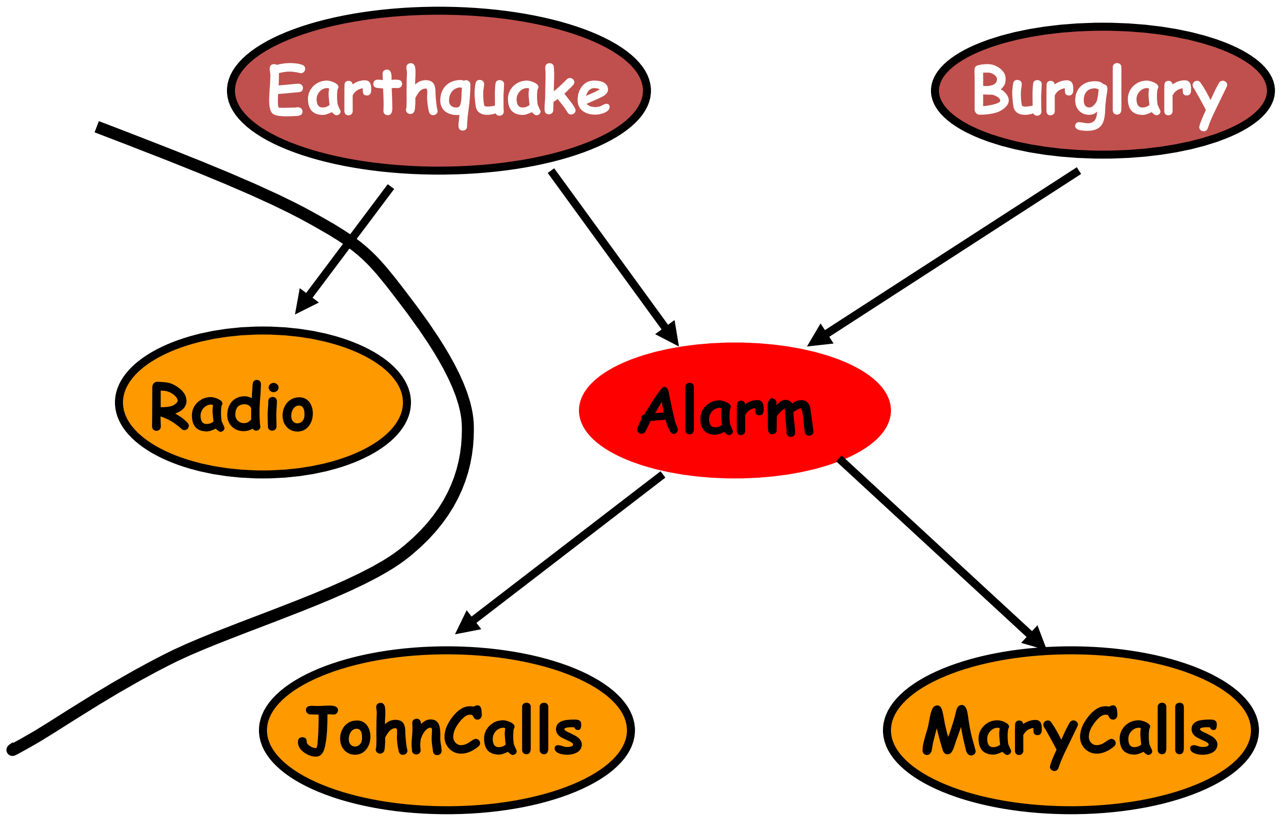
For Example



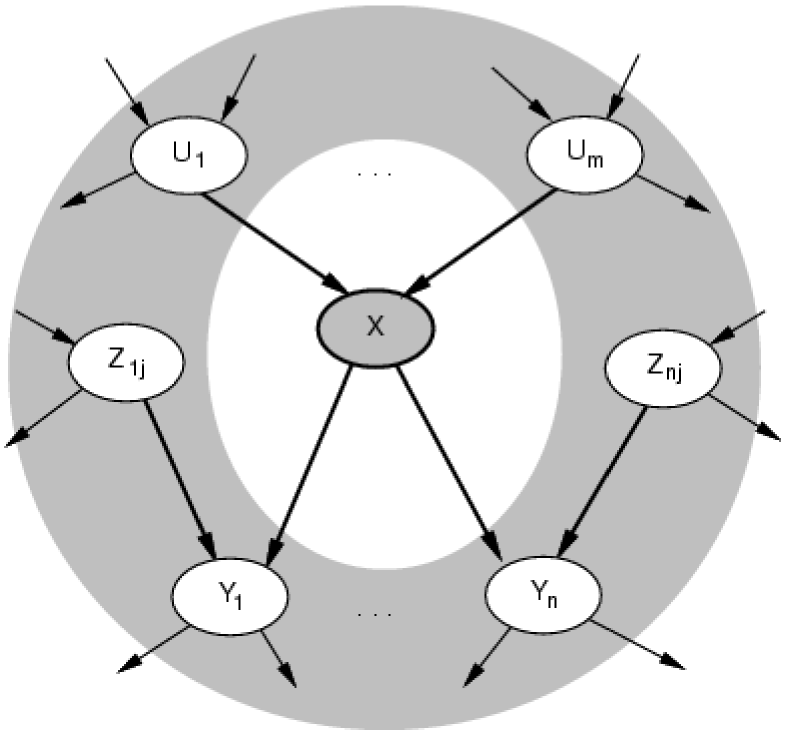
For Example



For Example

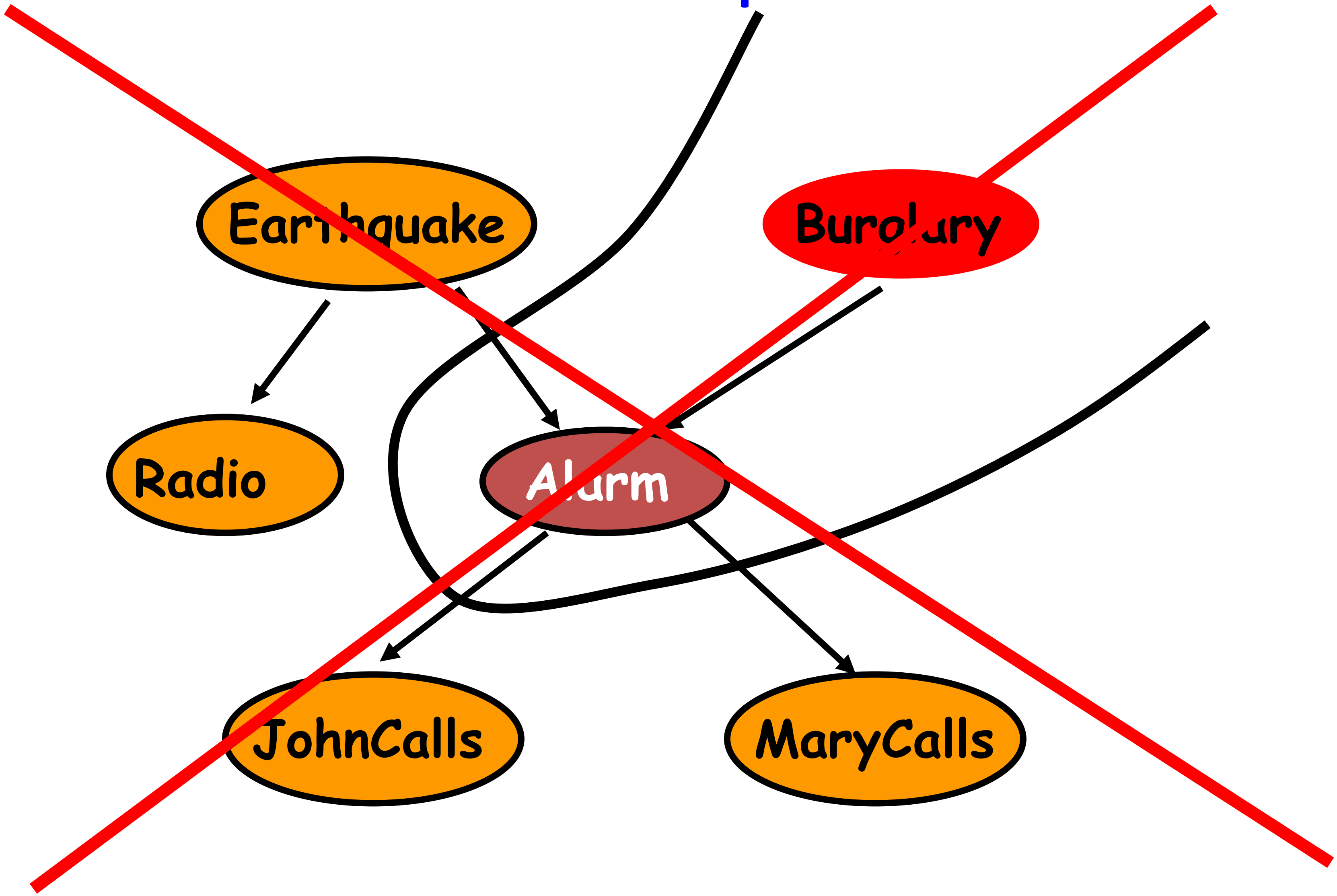


Given Markov Blanket, X is Independent of All Other Nodes

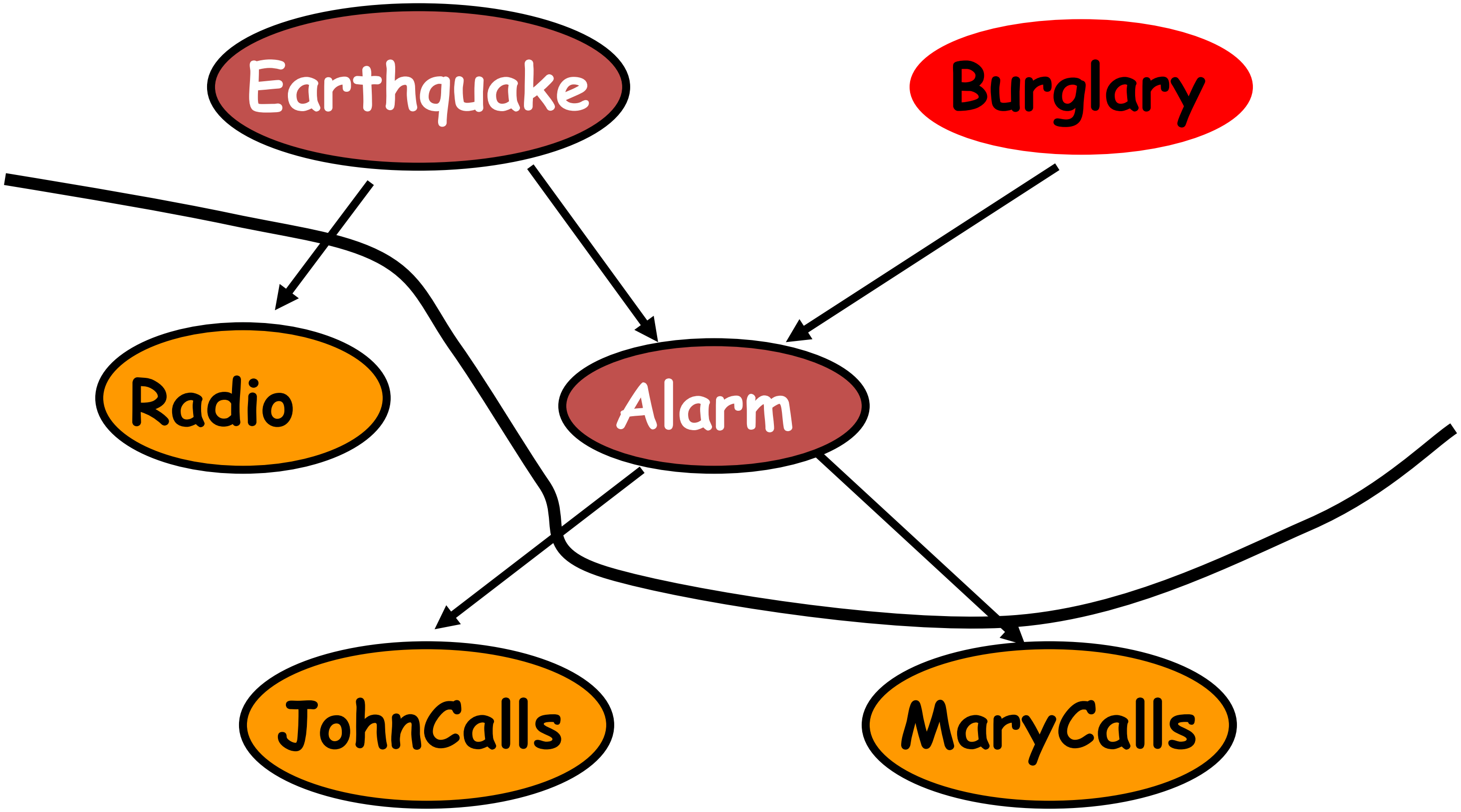


MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

For Example



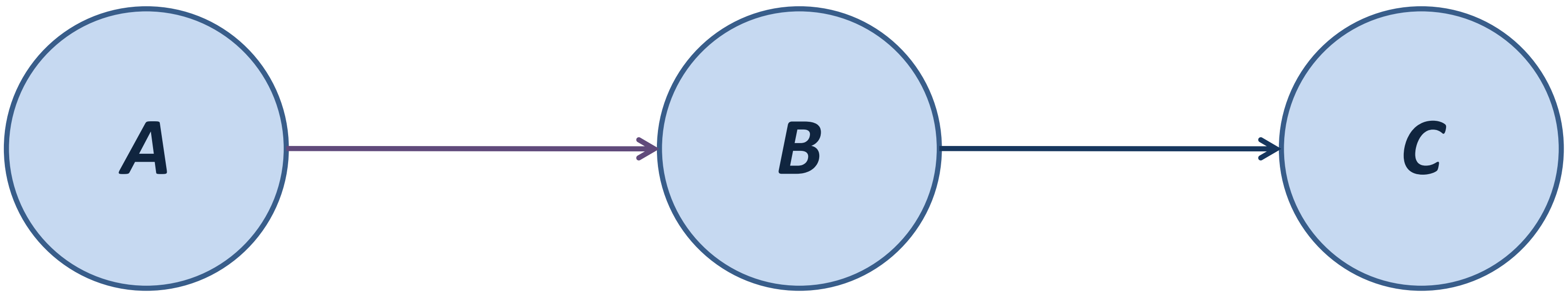
For Example



d-Separation

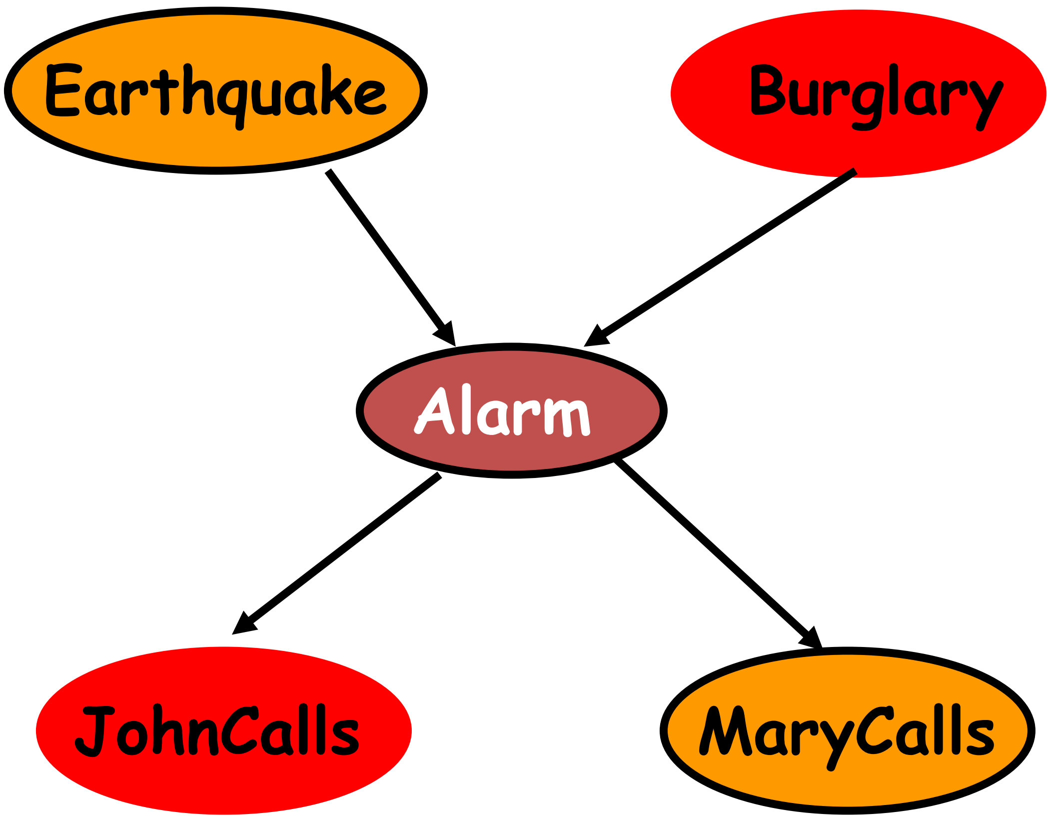
- An undirected path between two nodes is “cut off” if information cannot flow across one of the nodes in the path
- Two nodes are d-separated if every undirected path between them is cut off
- Two sets of nodes are d-separated if every pair of nodes, one from each set, is d-separated

d-Separation

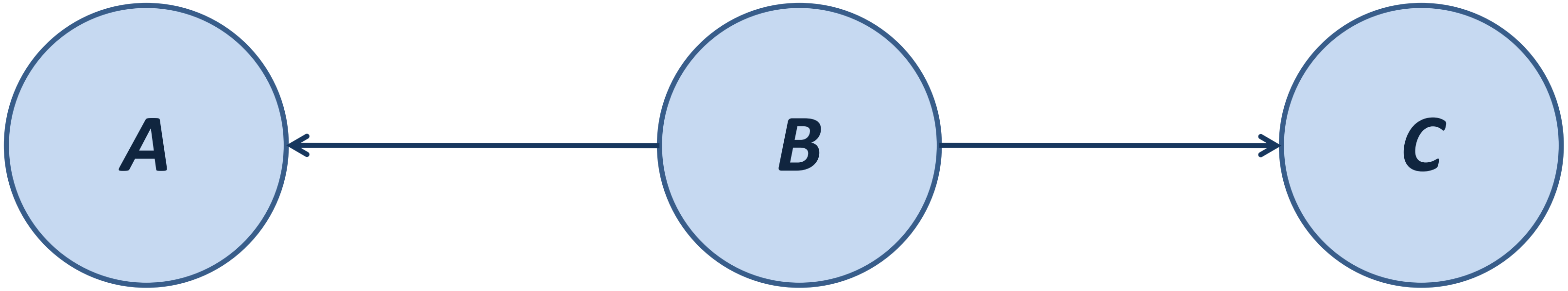


Linear connection: Information can flow between A and C if and only if we do not have evidence at B

For Example

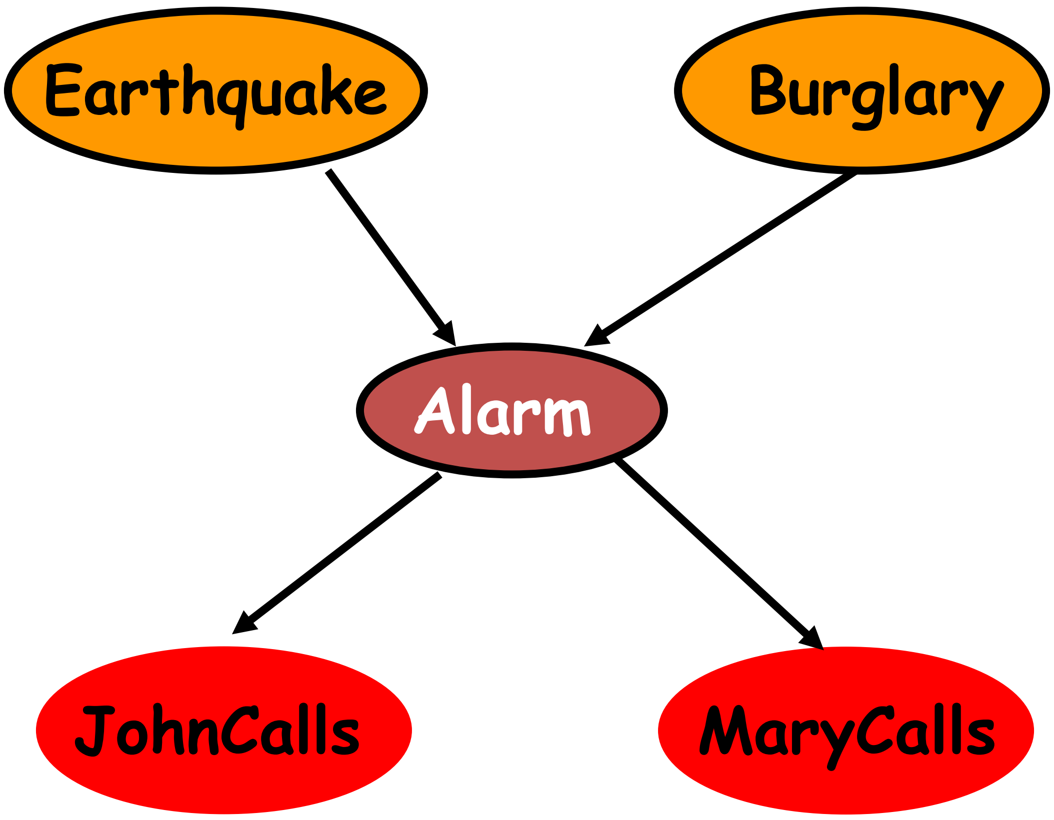


d-Separation (continued)

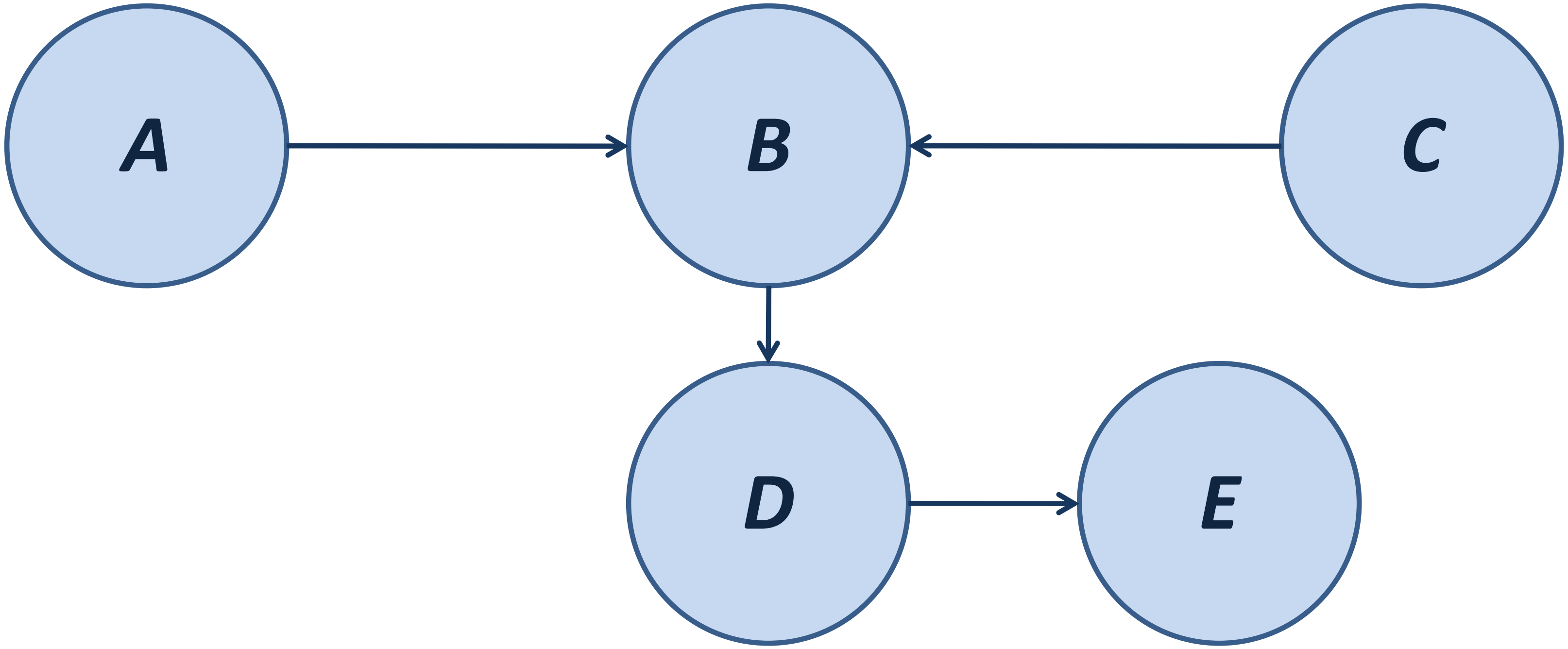


Diverging connection: Information can flow between A and C if and only if we **do not** *have evidence at B*

For Example

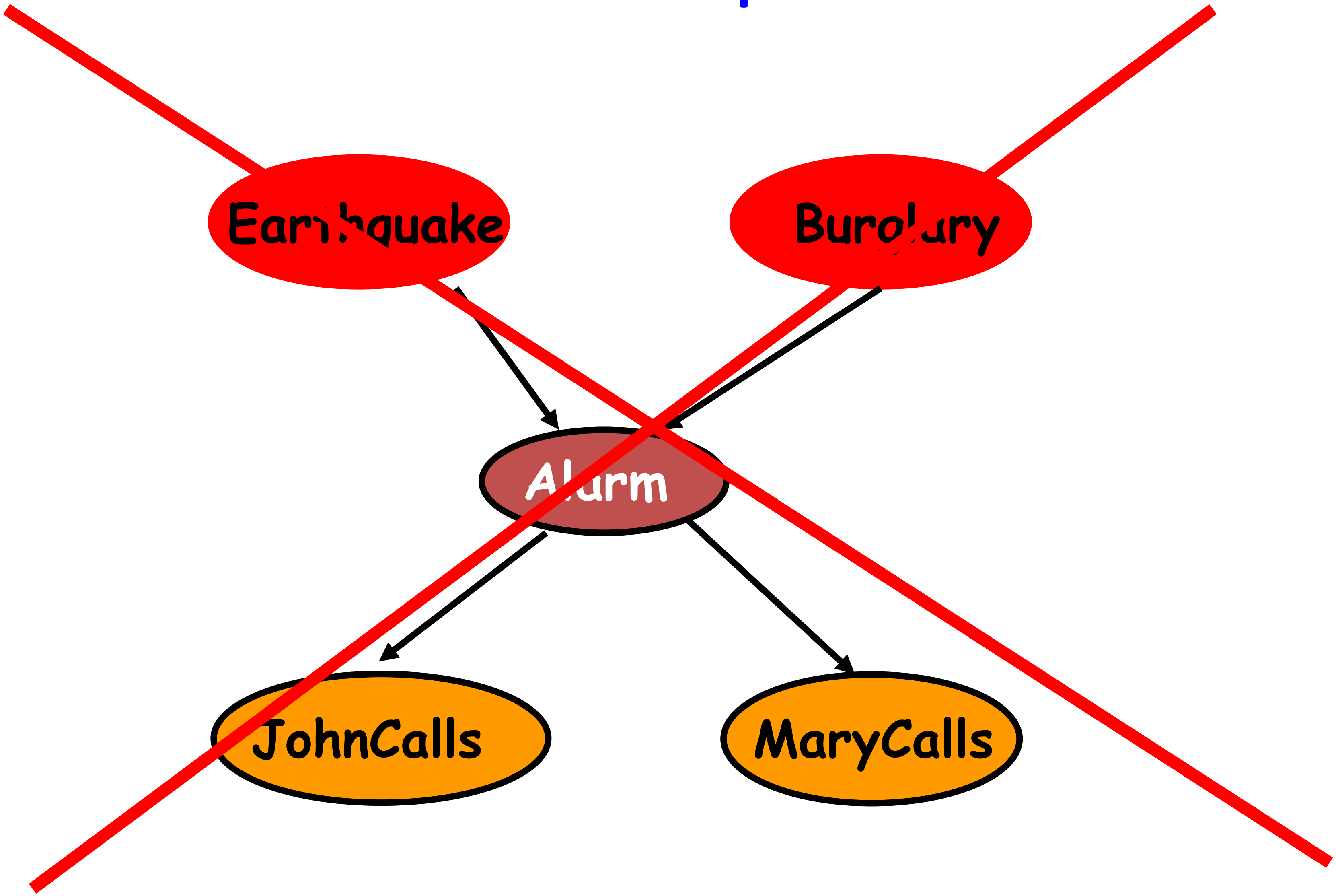


d-Separation (continued)



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

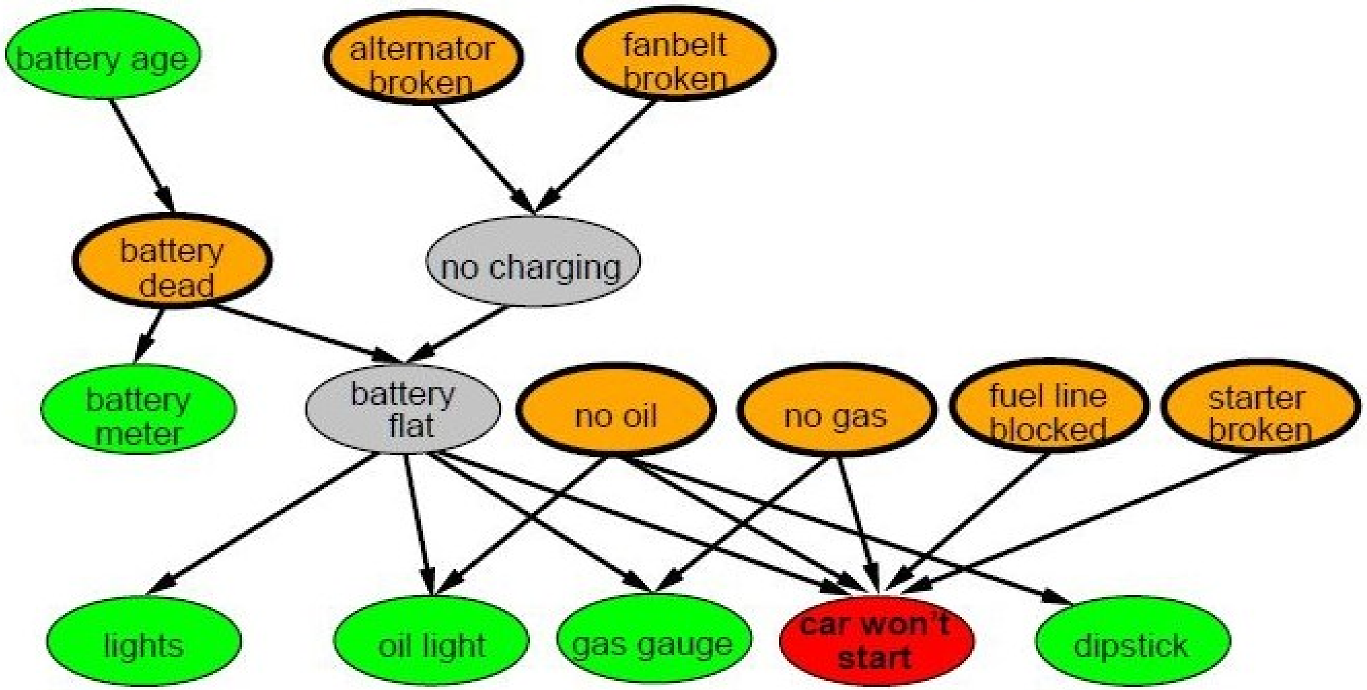


d-Separation

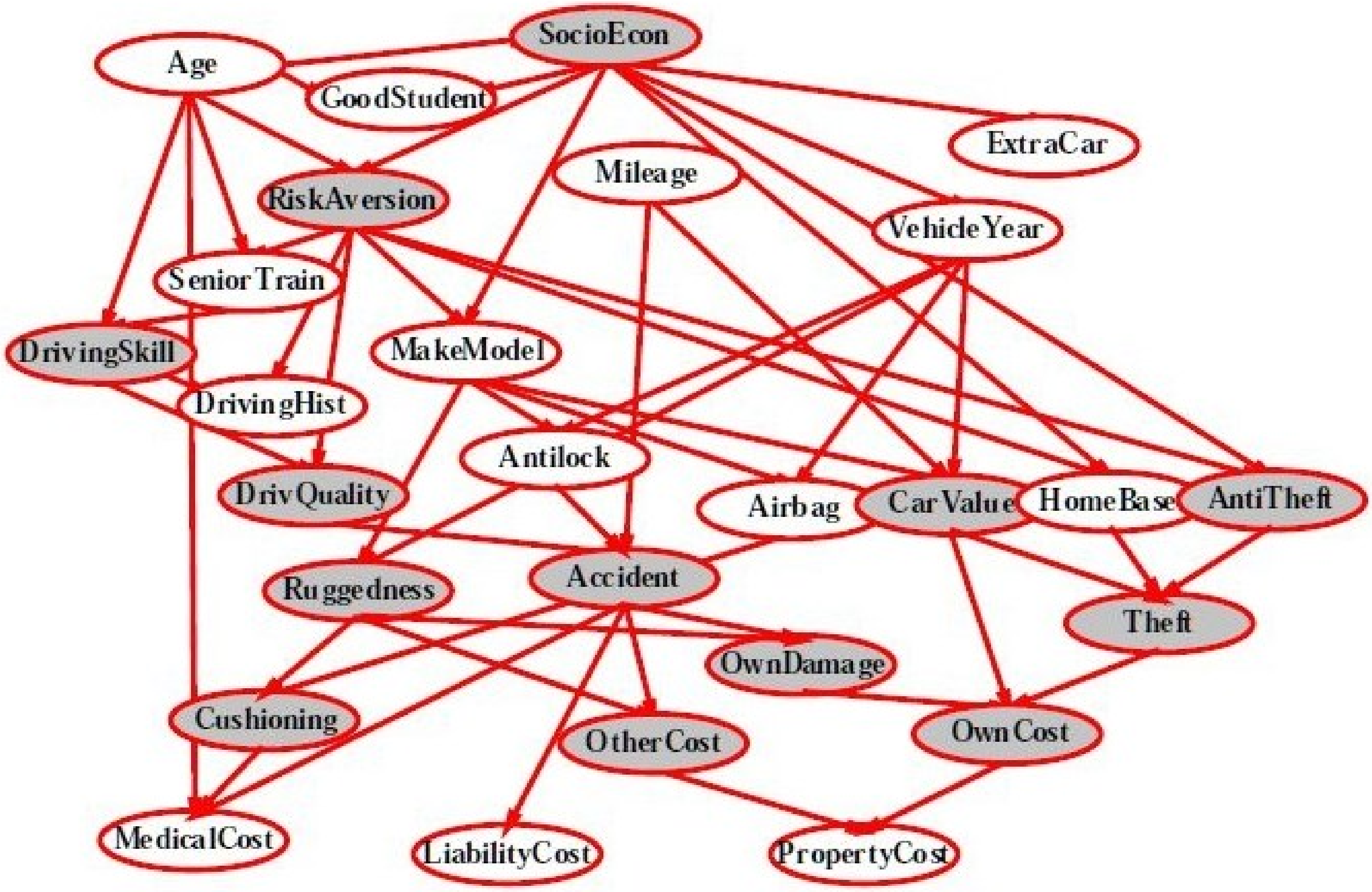
- An undirected path between two nodes is “cut off” if information cannot flow across one of the nodes in the path
- Two nodes are d-separated if every undirected path between them is cut off
- Two sets of nodes are d-separated if every pair of nodes, one from each set, is d-separated

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



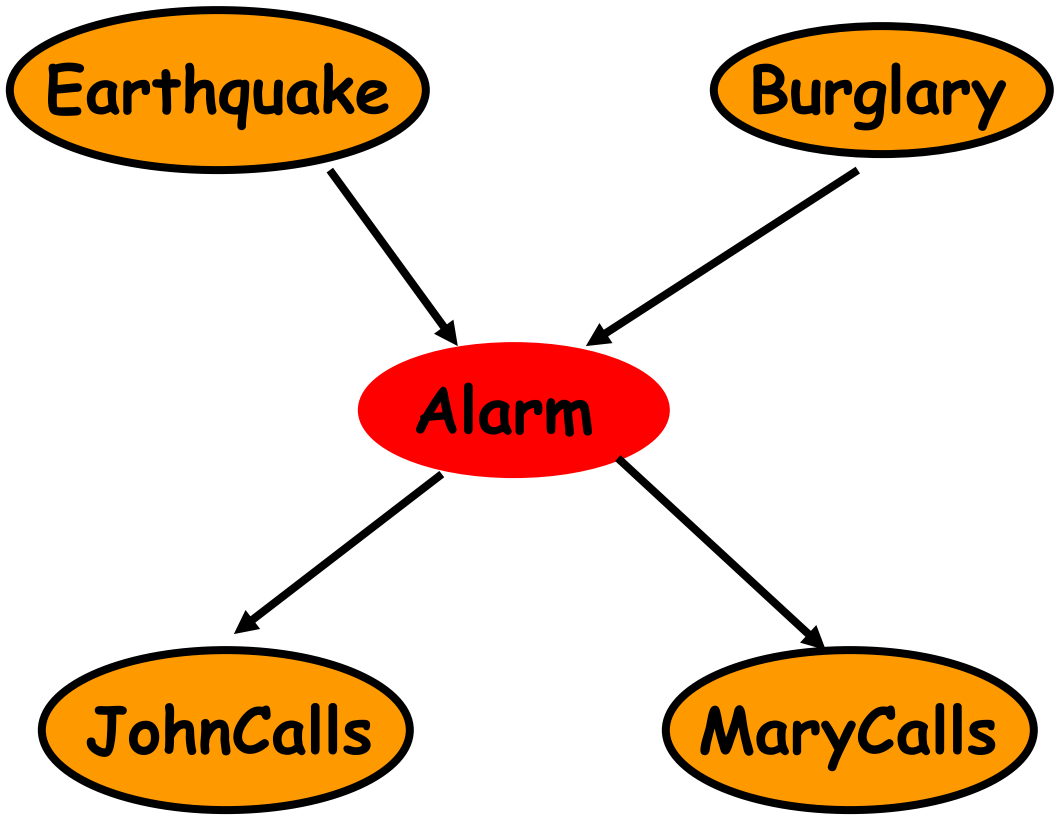
Other Applications

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

Inference in BNs

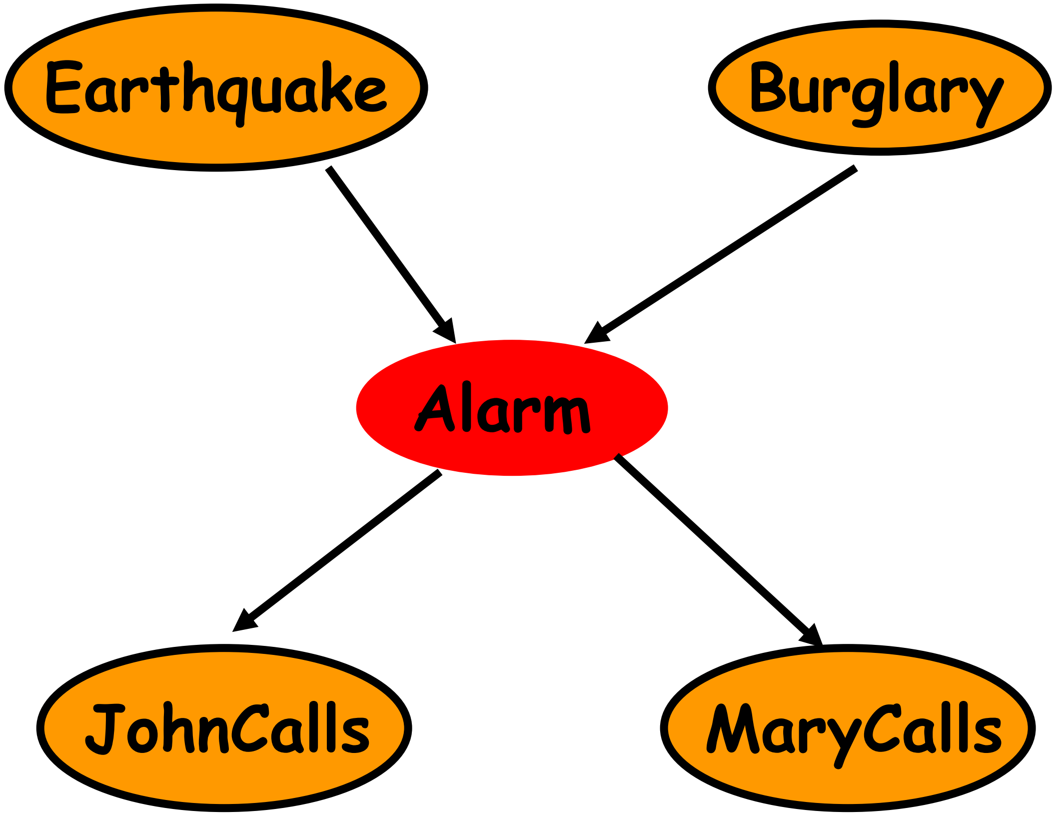
- The graphical independence representation
  - yields efficient inference schemes
- We generally want to compute
  - Marginal probability: *Pr(Z),*
  - *Pr(Z|**E**)* where ***E*** is (conjunctive) evidence
    - Z: query variable(s),
    - E: evidence variable(s)
    - everything else: hidden variable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

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



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

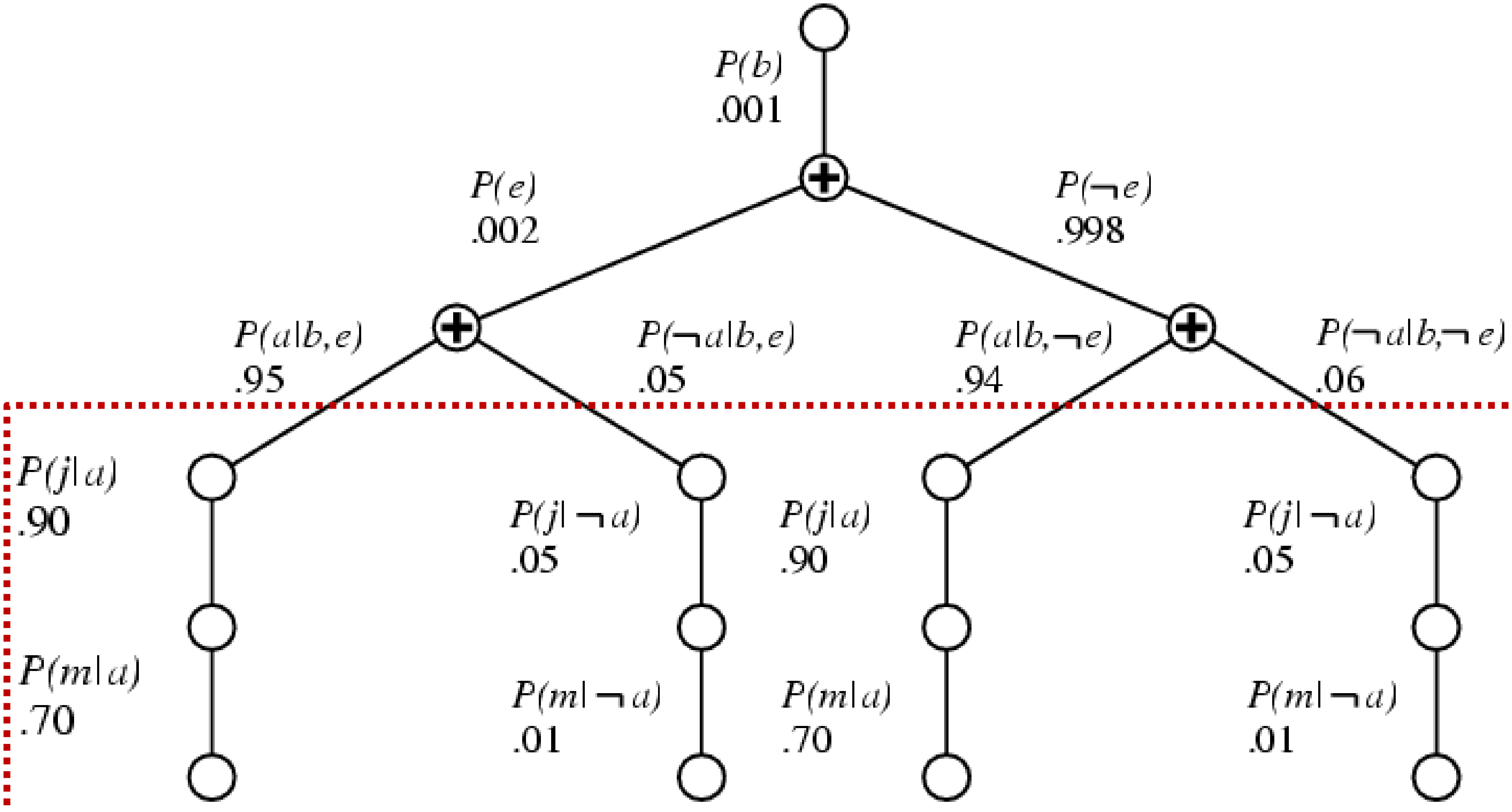
P(B | J=true, M=true)



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

Variable Elimination

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



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)

$P(J)$

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

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

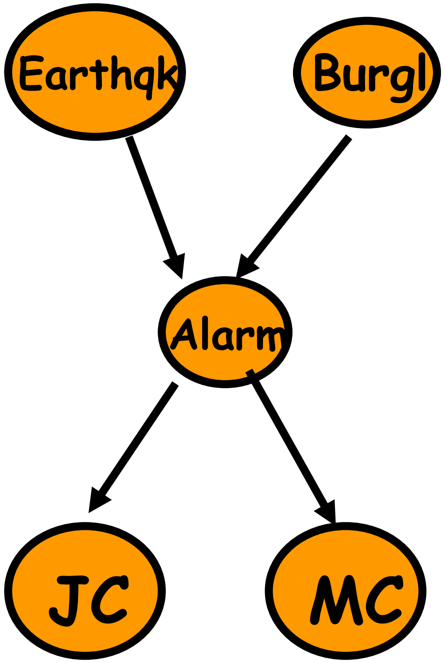
$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

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

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

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

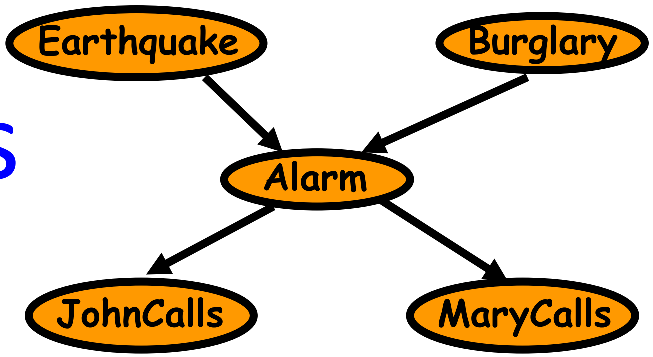
$= f4(J)$



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



$P(J)$

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

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

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

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

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

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

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

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

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

Learning in Bayes Nets

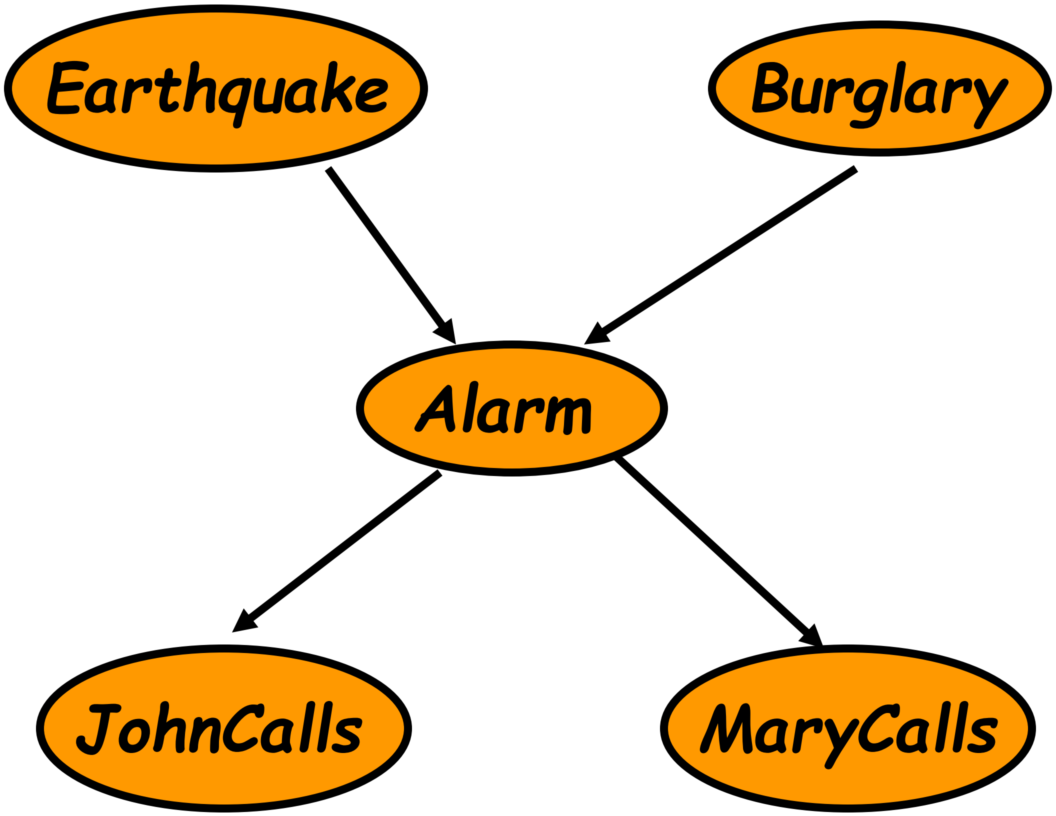
Mausam

(Based on slides by Stuart Russell, Marie desJardins, Subbarao Kambhampati, Dan Weld)

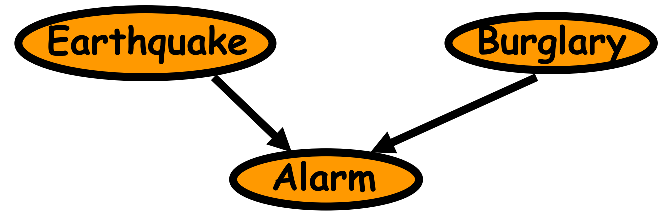
Parameter Estimation

- Learn all the CPTs in a Bayesian Net
- Data → Model → Queries
- Key idea: counting!

Burglars and Earthquakes



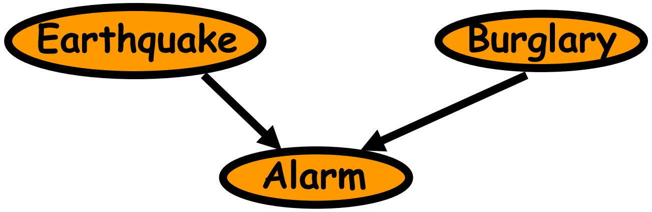
Counting



E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e,b	
$e,\bar{b}$	
$\bar{e},b$	
$\bar{e},\bar{b}$	

Counting

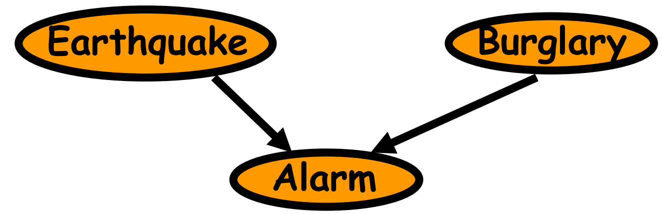


E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e, b	
$e, \bar{b}$	
$\bar{e}, b$	
$\bar{e}, \bar{b}$	

$P(\bar{a}|e, b) = ?$

Counting

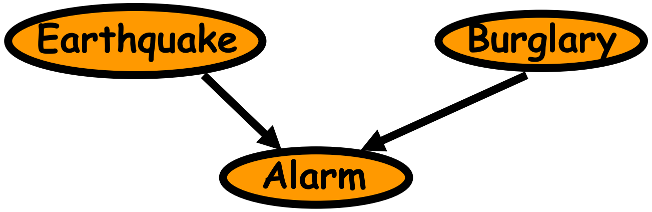


E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e,b	
$e,\bar{b}$	
$\bar{e},b$	
$\bar{e},\bar{b}$	~0.01

$$P(\bar{a}|e,\ b) = ?$$

Counting

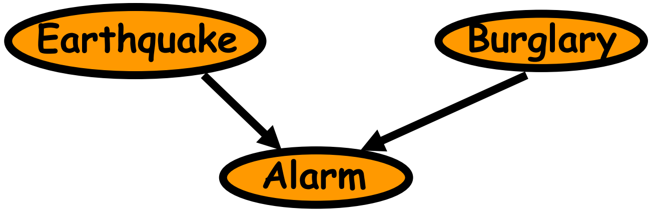


E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e,b	
$e,\bar{b}$	
$\bar{e}$,b	0.83
$\bar{e},\bar{b}$	~0.01

$P(a\,\overline{|}\,e,\ b) = ?$

Counting

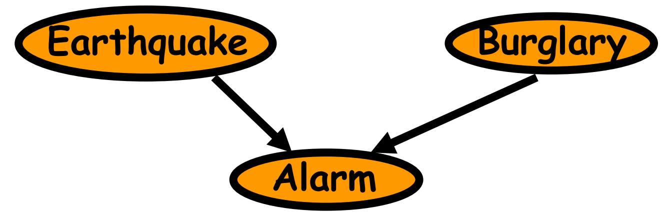


E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e,b	
$e,\bar{b}$	0.2
$\bar{e},b$	0.83
$\bar{e},\bar{b}$	~0.01

P(a|e, b) = ?

Counting



E	B	A	#
0	0	0	1000
0	0	1	10
0	1	0	20
0	1	1	100
1	0	0	200
1	0	1	50
1	1	0	0
1	1	1	5

	Pr(A|E,B)
e,b	*1*
$e,\bar{b}$	0.2
$\bar{e},b$	0.83
$\bar{e},\bar{b}$	~0.01

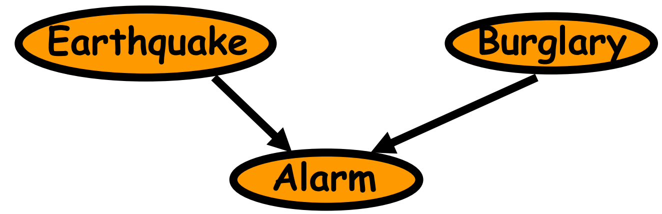
Bad idea to have prob as 0 or

- ***stumps Gibbs sampling***
- ***low prob states become impos***

Solution: Smoothing

- Why?
  - To deal with events observed zero times.
  - "event": a particular ngram
- How?
  - To shave a little bit of probability mass from the higher counts, and pile it instead on the zero counts
- Laplace Smoothing/Add-one smoothing
  - assume each event was observed at least once.
  - add 1 to all frequency counts
- Add m instead of 1 (m could be $>$ or $<$ 1)

Counting w/ Smoothing



E	B	A	#
0	0	0	1000+1
0	0	1	10+1
0	1	0	20+1
0	1	1	100+1
1	0	0	200+1
1	0	1	50+1
1	1	0	0+1
1	1	1	5+1

	Pr(A|E,B)
e,b	***0.86***
$e,\bar{b}$	~0.2
$\bar{e},b$	~0.83
$\bar{e},\bar{b}$	~0.01

ML vs. MAP Learning

- ML: maximum likelihood (what we just did)
  - find parameters that maximize the prob of seeing the data D
  - $\text{argmax}_\theta P(D|\theta)$
  - easy to compute (for example, just counting)
  - assumes uniform prior
- Prior: your belief before seeing any data
  - Uniform prior: all parameters equally likely
- MAP: maximum a posteriori estimate
  - maximize prob of parameters after seeing data D
  - $\text{argmax}_\theta P(\theta|D) = \text{argmax}_\theta P(D|\theta)P(\theta)$
  - allows user to input additional domain knowledge
  - better parameters when data is sparse…
  - reduces to ML when infinite data

Other Graphical Models