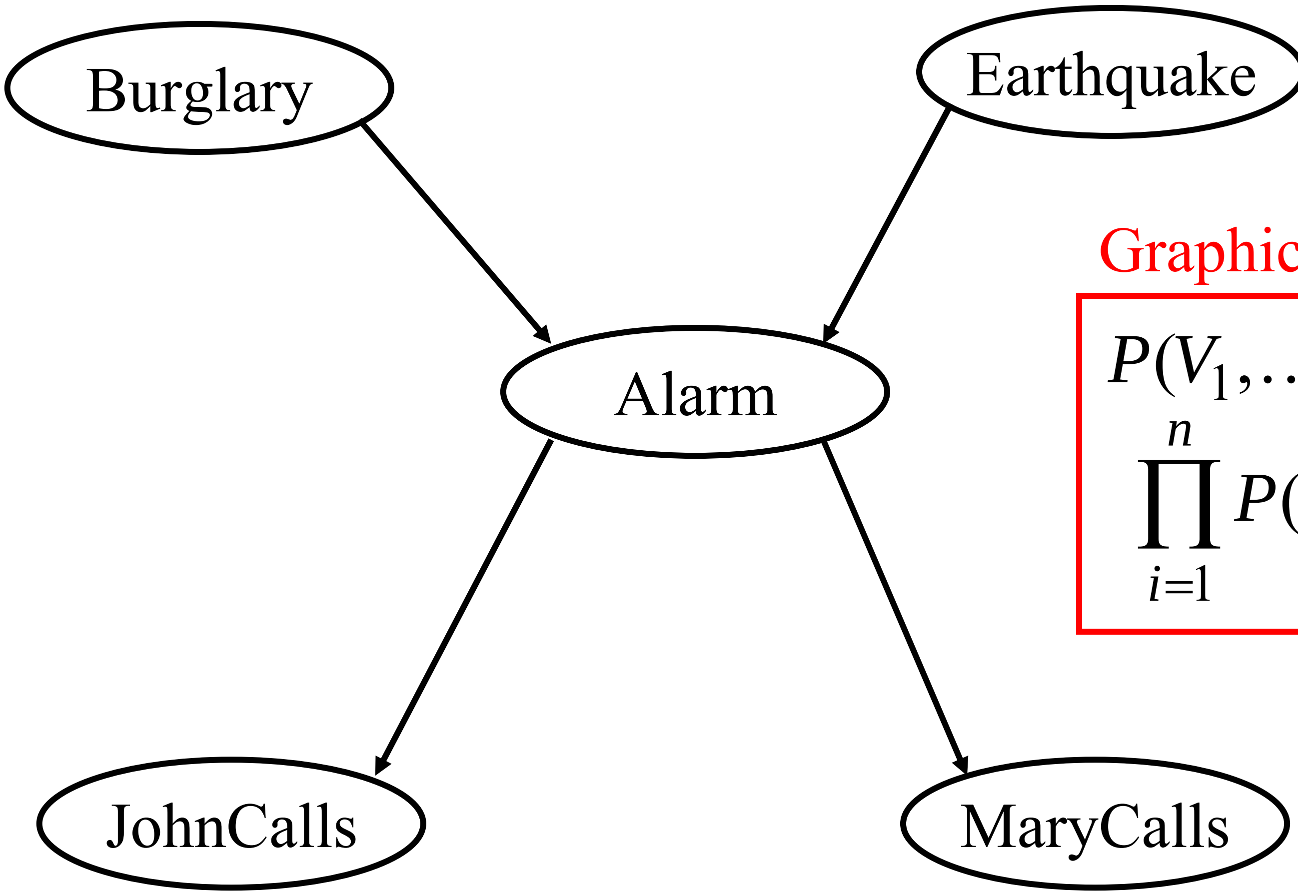
Compression in Bayes nets

- A Bayes net compresses the joint probability distribution over a set of variables in two ways:
  - Dependency structure
  - Parameterization
- Both kinds of compression derive from causal structure:
  - Causal locality
  - Independent causal mechanisms

Dependency structure

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

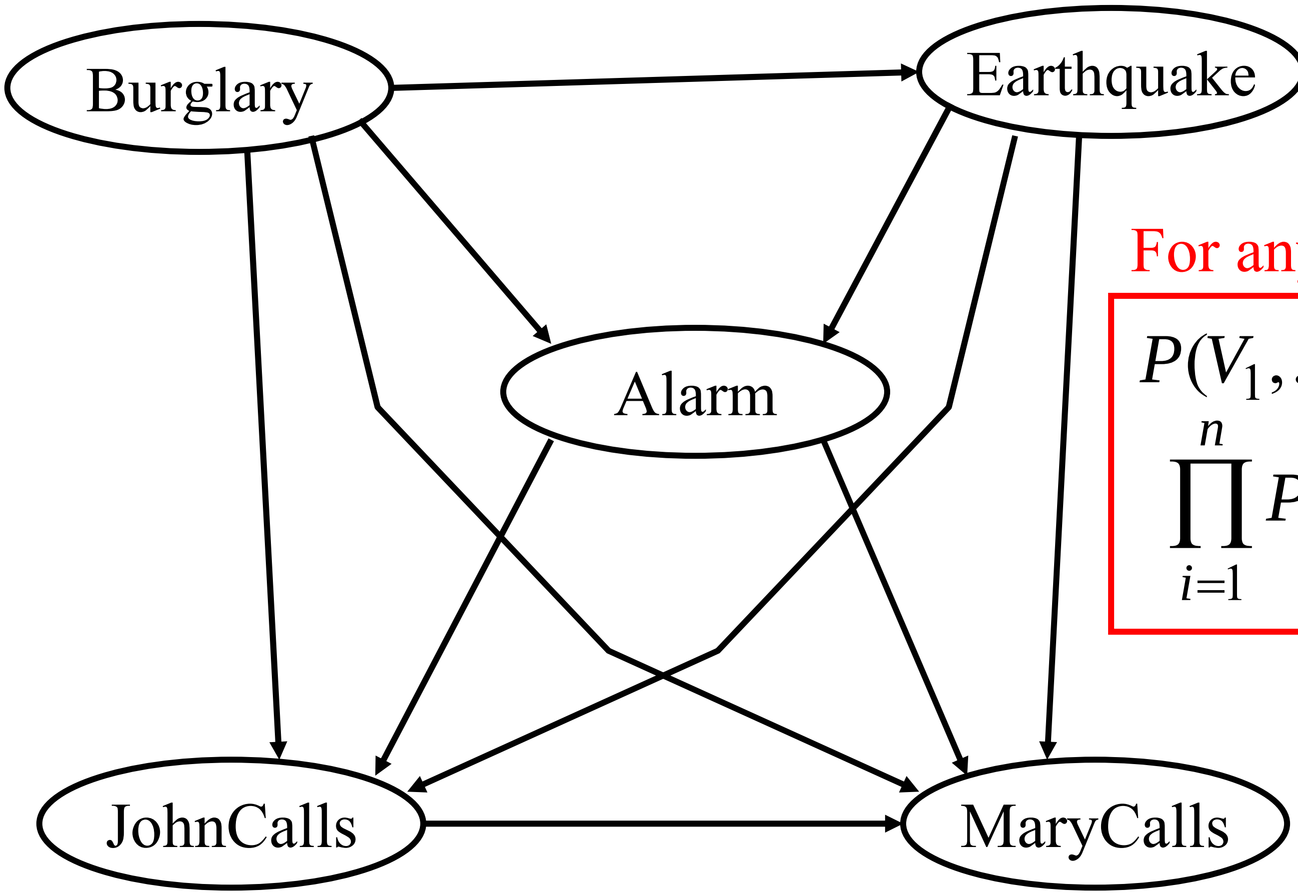


Graphical model asserts:

$$P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i \mid \text{parents}[V_i])$$

Dependency structure

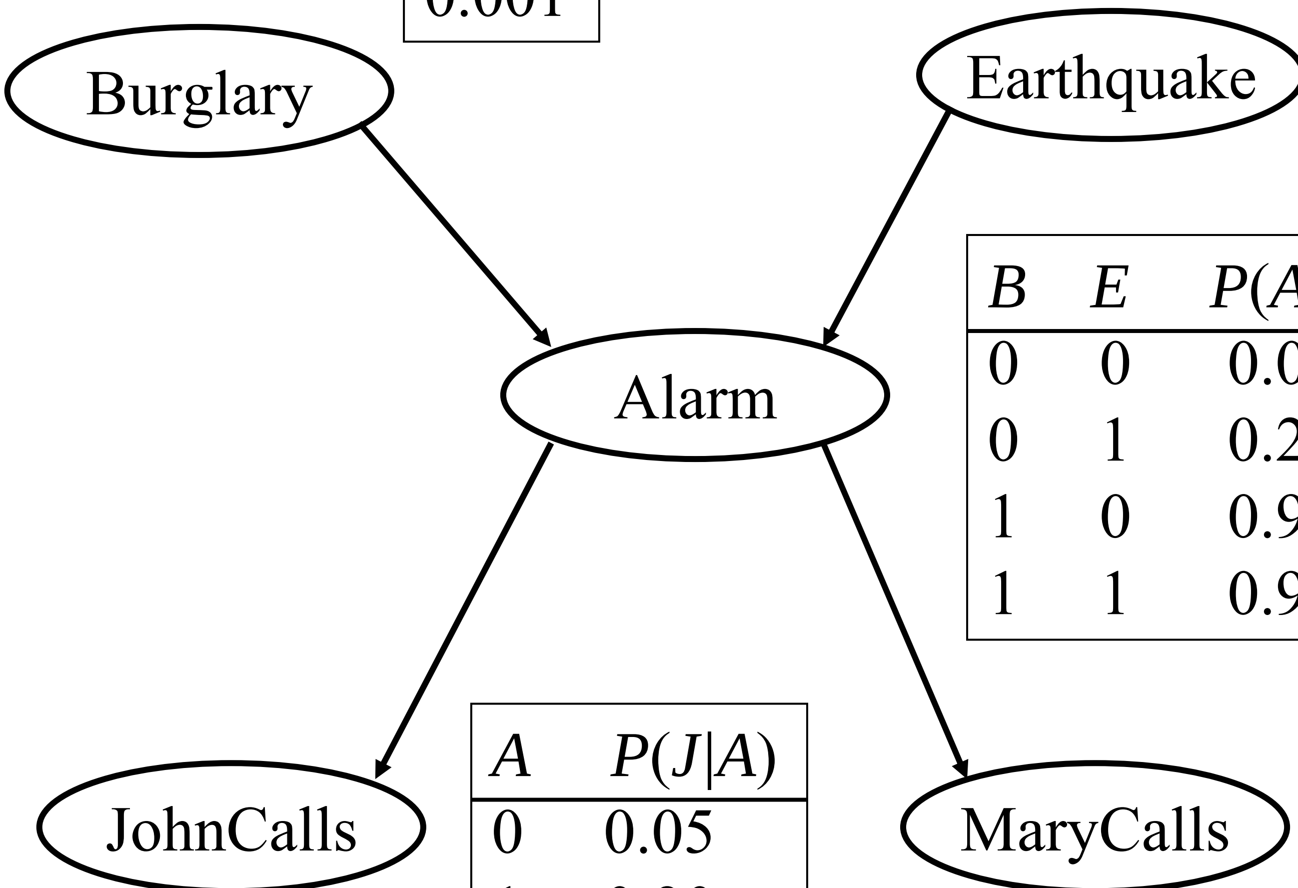
$$P(B,E,A,J,M) =$$
$$P(B)\,P(E \mid B)\,P(A \mid B,E)\,P(J \mid B,E,A)\,P(M \mid B,E,A,J)$$



For any distribution:

$$\boxed{P(V_1,\ldots,V_n) = \prod_{i=1}^{n} P(V_i \mid V_1,\ldots,V_{i-1})}$$

Parameterization



$P(B)$
0.001

$P(E)$
0.002

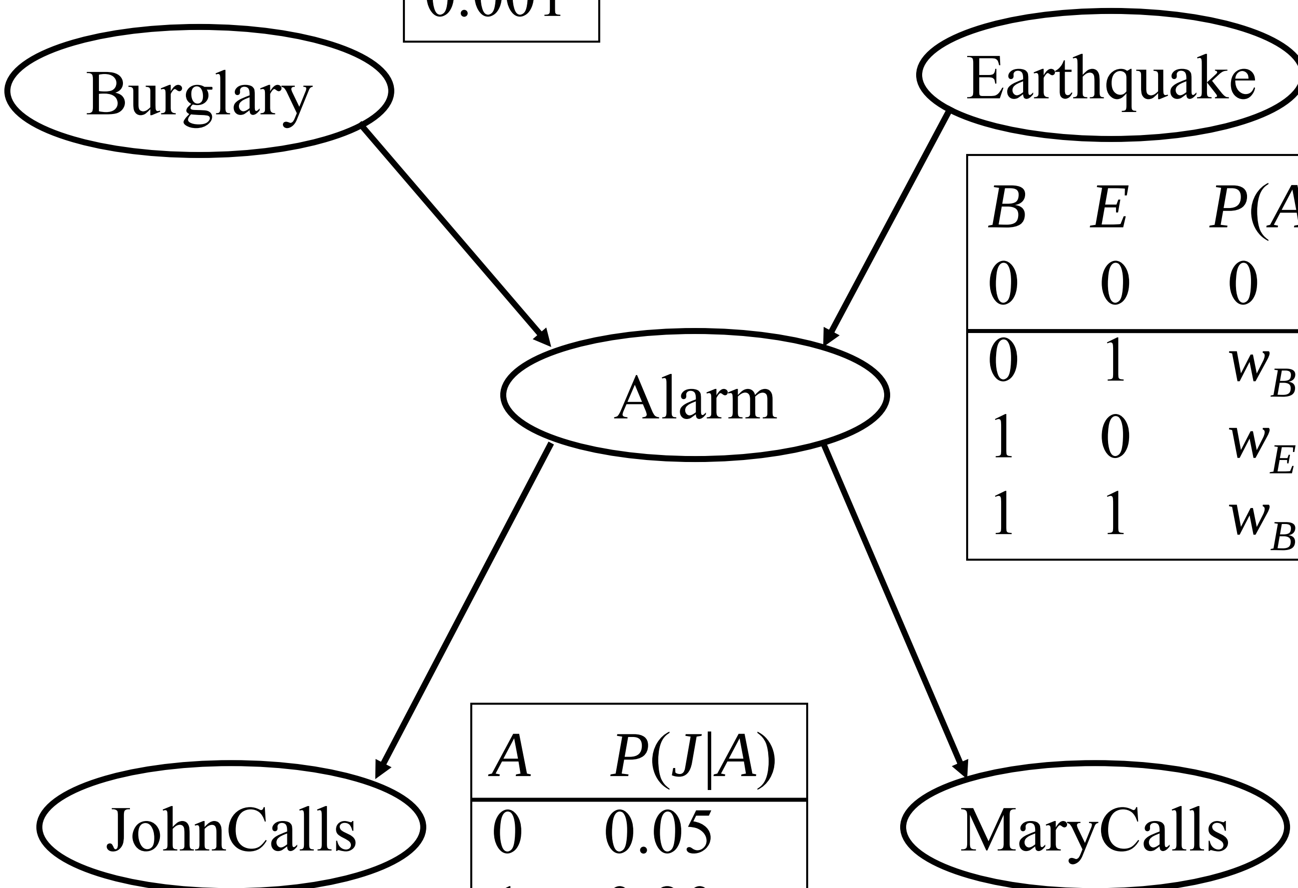
B	E	$P(A\|B,E)$
0	0	0.001
0	1	0.29
1	0	0.94
1	1	0.95

Full CPT

A	$P(J\|A)$
0	0.05
1	0.90

A	$P(M\|A)$
0	0.01
1	0.70

Parameterization



$P(B)$
0.001

$P(E)$
0.002

Noisy-OR

B	E	$P(A\|B,E)$
0	0	0
0	1	$w_B = 0.29$
1	0	$w_E = 0.94$
1	1	$w_B + (1-w_B)w_E$

FIX FOR NEXT YEAR Wb and We

A	$P(J\|A)$
0	0.05
1	0.90

A	$P(M\|A)$
0	0.01
1	0.70

Outline

- The semantics of Bayes nets
  - role of causality in structural compression
- Explaining away revisited
  - role of causality in probabilistic inference
- Sampling algorithms for approximate inference in graphical models

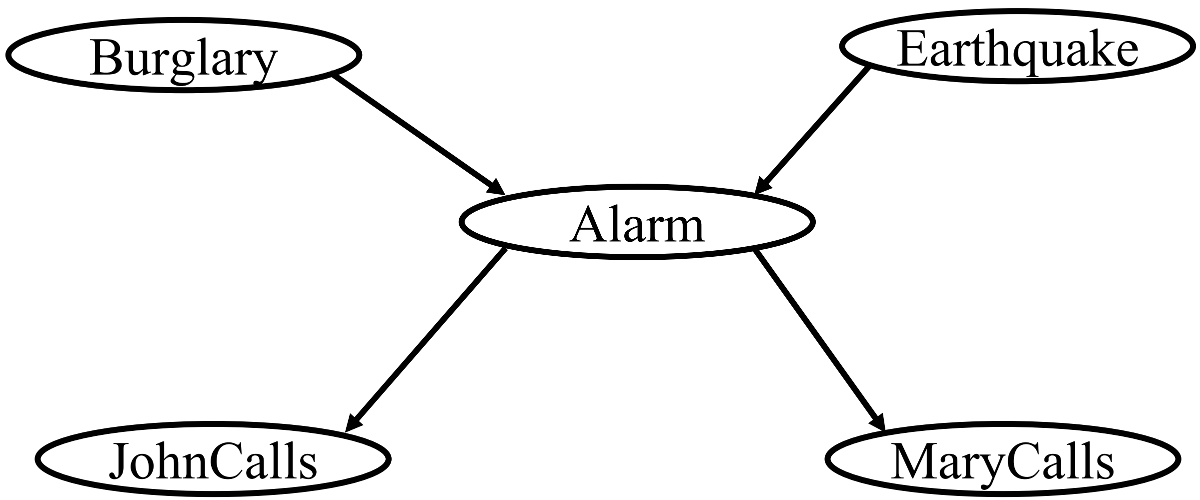
Outline

- The semantics of Bayes nets
  - role of causality in structural compression
- Explaining away revisited
  - role of causality in probabilistic inference
- Sampling algorithms for approximate inference in graphical models

Global semantics

Joint probability distribution factorizes into product of local conditional probabilities:

$$P(V_1,\ldots,V_n) = \prod_{i=1}^{n} P(V_i \mid \text{parents}[V_i])$$



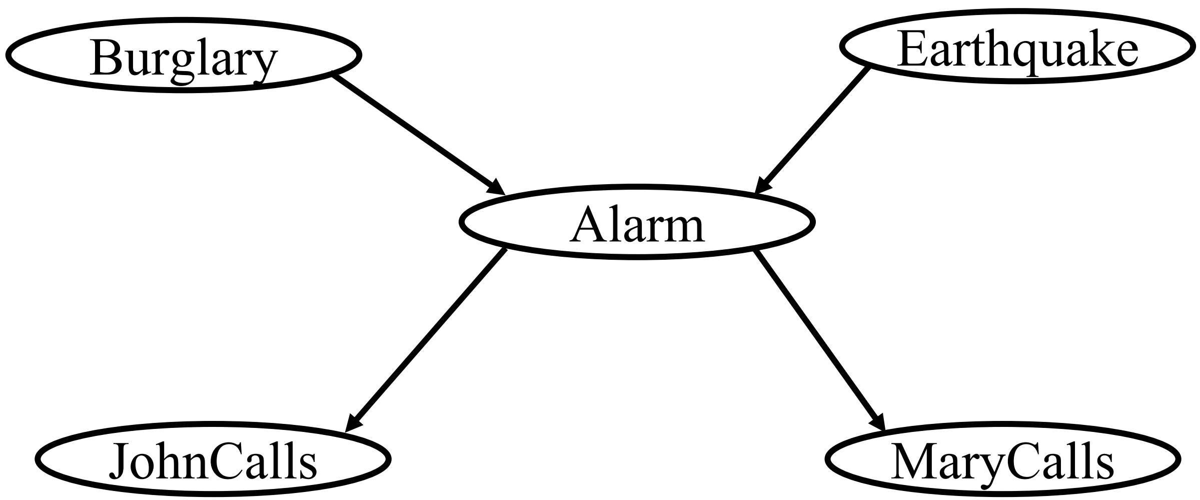
$$P(B, E, A, J, M) =$$

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

Global semantics

Joint probability distribution factorizes into product of local conditional probabilities:

$$P(V_1,\ldots,V_n) = \prod_{i=1}^{n} P(V_i \mid \text{parents}[V_i])$$



Necessary to assign a probability to any possible world, e.g.

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

Local semantics

Global factorization is equivalent to a set of constraints on pairwise relationships between variables.

"Markov property": Each node is conditionally independent of its non-descendants given its parents.

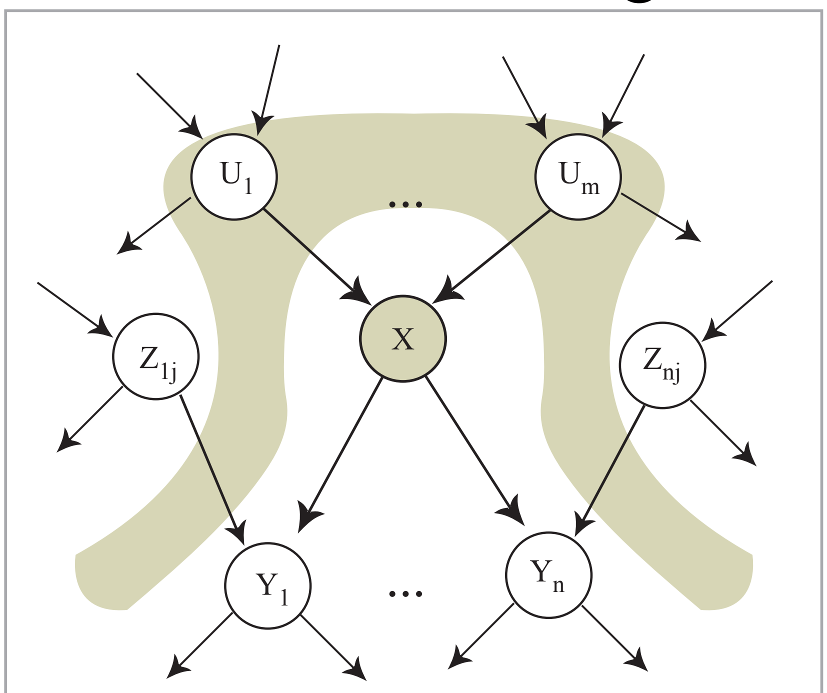


Figure by MIT OCW.

Local semantics

Global factorization is equivalent to a set of constraints on pairwise relationships between variables.

"Markov property": Each node is conditionally independent of its non-descendants given its parents.

Also: Each node is marginally (a priori) independent of any non-descendant unless they share a common ancestor.

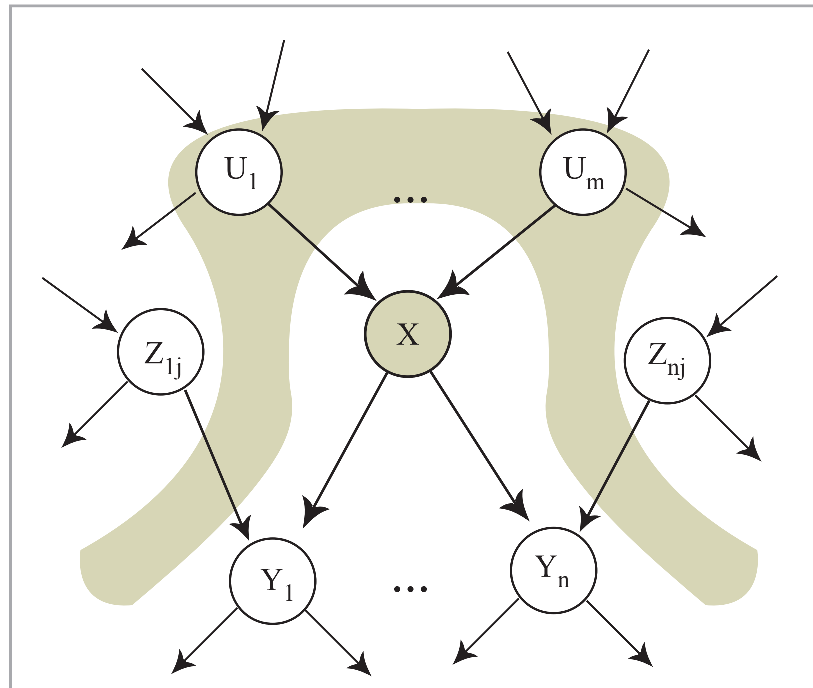


Figure by MIT OCW.

Local semantics

Global factorization is equivalent to a set of constraints on pairwise relationships between variables.

Each node is conditionally independent of all others given its "Markov blanket": parents, children, children's parents.

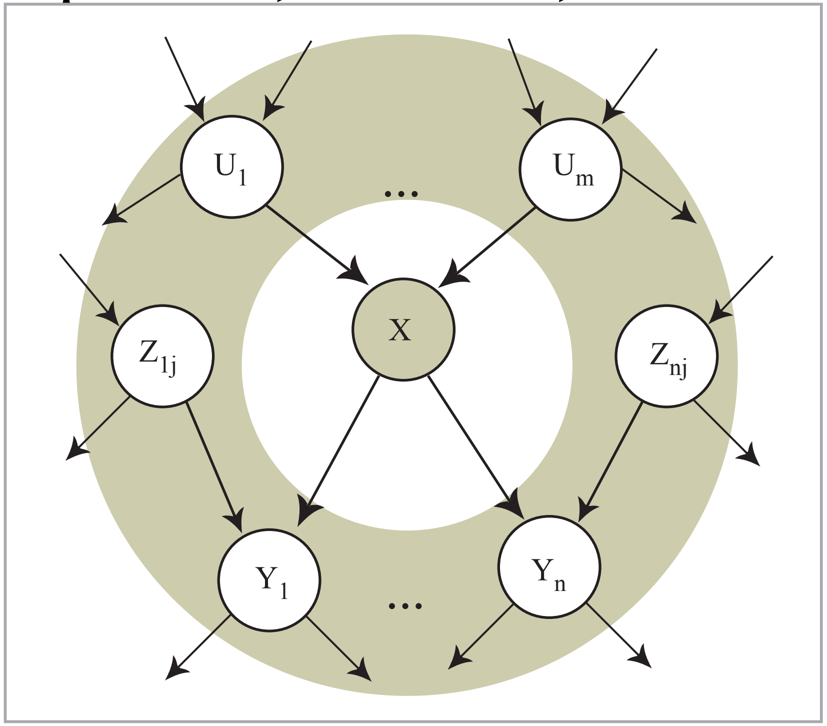
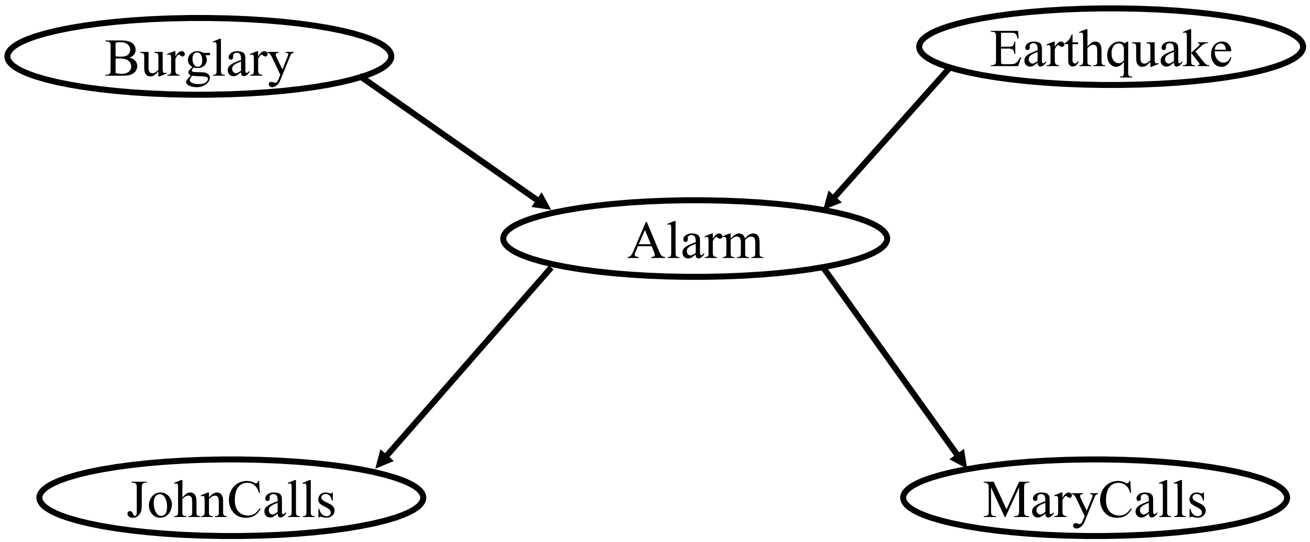


Figure by MIT OCW.

Example



JohnCalls and *MaryCalls* are marginally (a priori) dependent, but conditionally independent given *Alarm*. ["Common cause"]

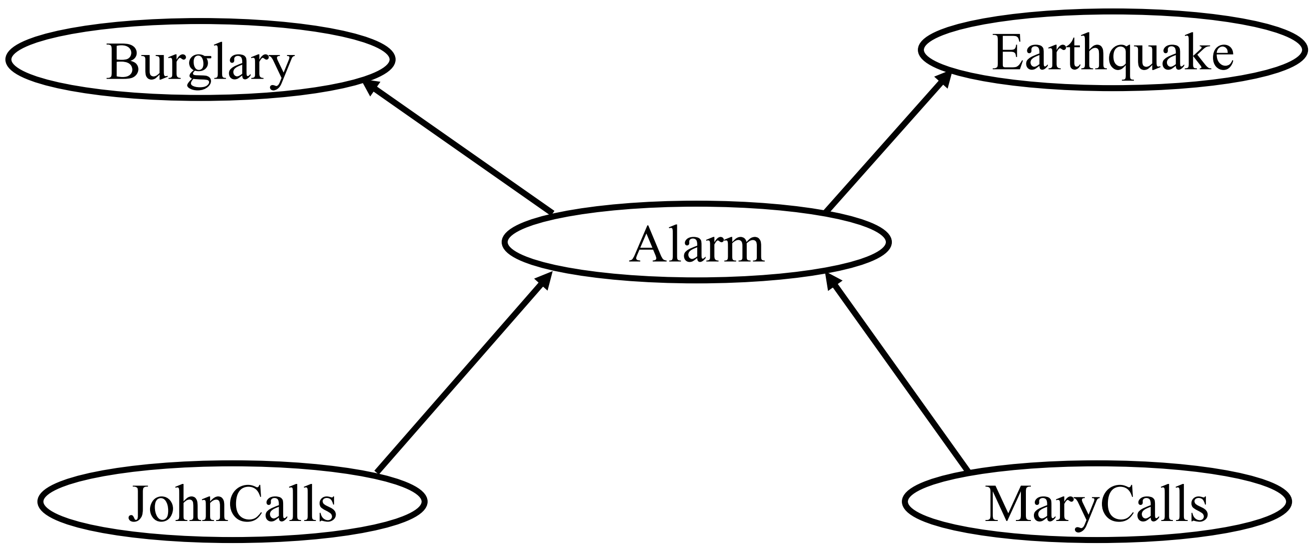
Burglary and *Earthquake* are marginally (a priori) independent, but conditionally dependent given *Alarm*. ["Common effect"]

Constructing a Bayes net

- Model reduces all pairwise dependence and independence relations down to a basic set of pairwise dependencies: graph edges.
- An analogy to learning kinship relations
  - Many possible bases, some better than others
  - A basis corresponding to direct causal mechanisms seems to compress best.

An alternative basis

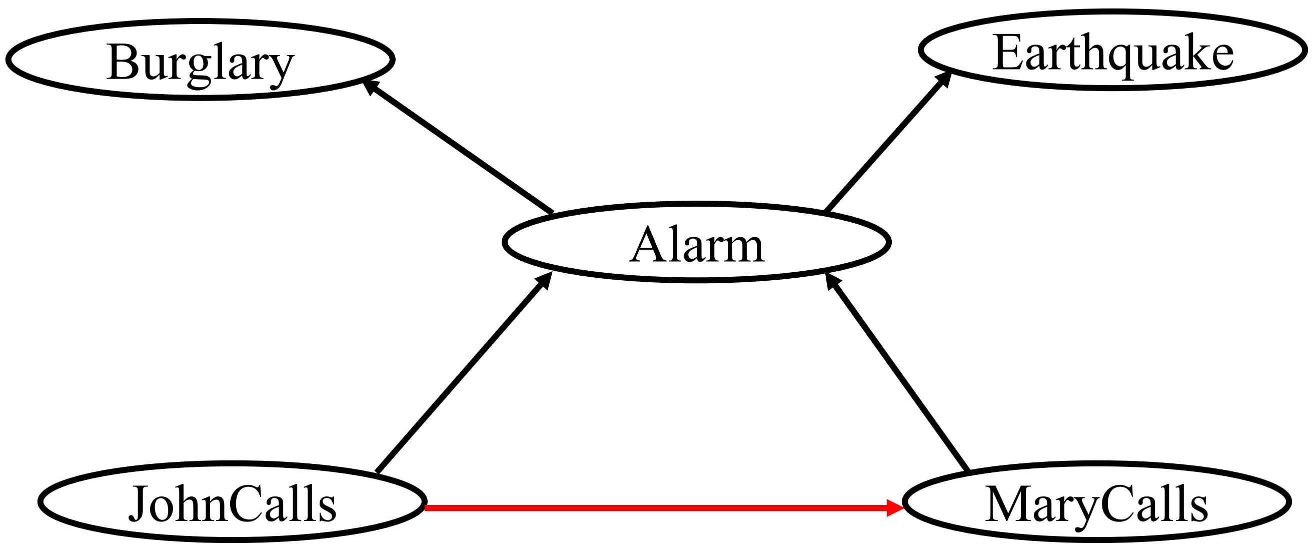
Suppose we get the direction of causality wrong...



- Does not capture the dependence between callers: falsely believes $P(JohnCalls, MaryCalls) = P(JohnCalls)\,P(MaryCalls)$.

An alternative basis

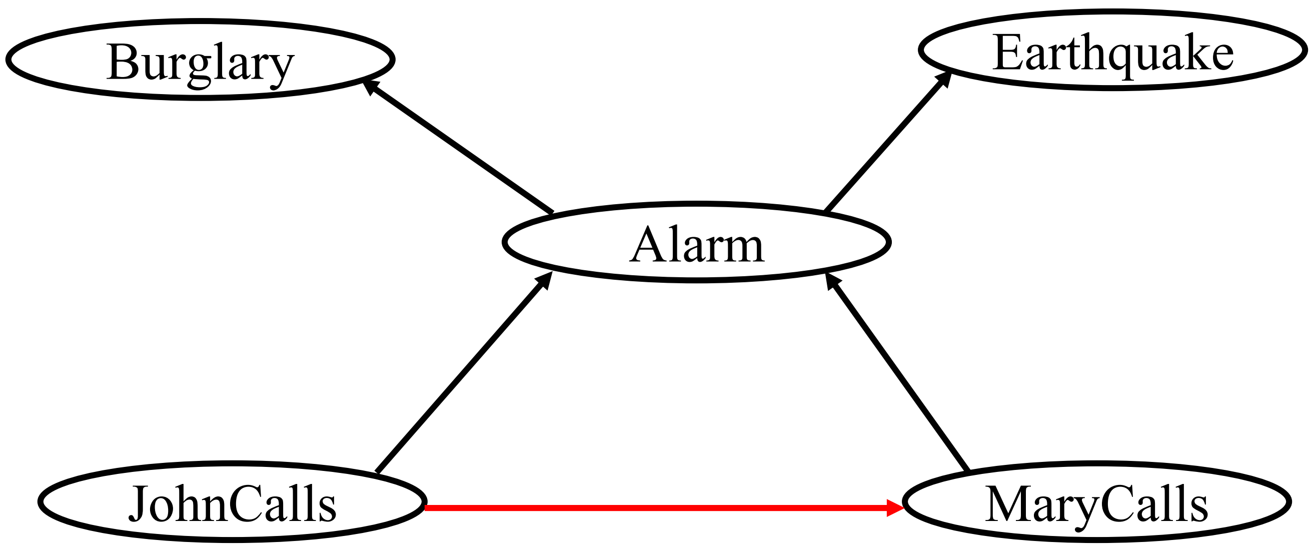
Suppose we get the direction of causality wrong...



- Inserting a new arrow captures this correlation.
- This model is too complex: does not believe that $P(JohnCalls, MaryCalls|Alarm) =$
  $P(JohnCalls|Alarm)\, P(MaryCalls|Alarm)$

An alternative basis

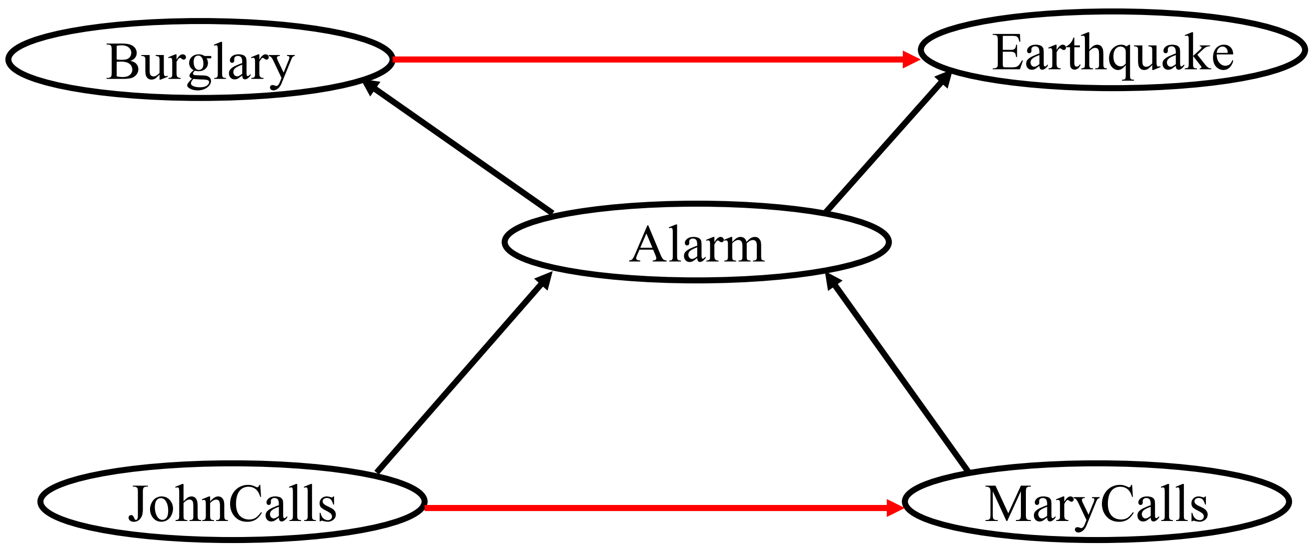
Suppose we get the direction of causality wrong...



- Does not capture conditional dependence of causes ("explaining away"): falsely believes that $P(Burglary, Earthquake|Alarm) =$
  $P(Burglary|Alarm)\ P(Earthquake|Alarm)$

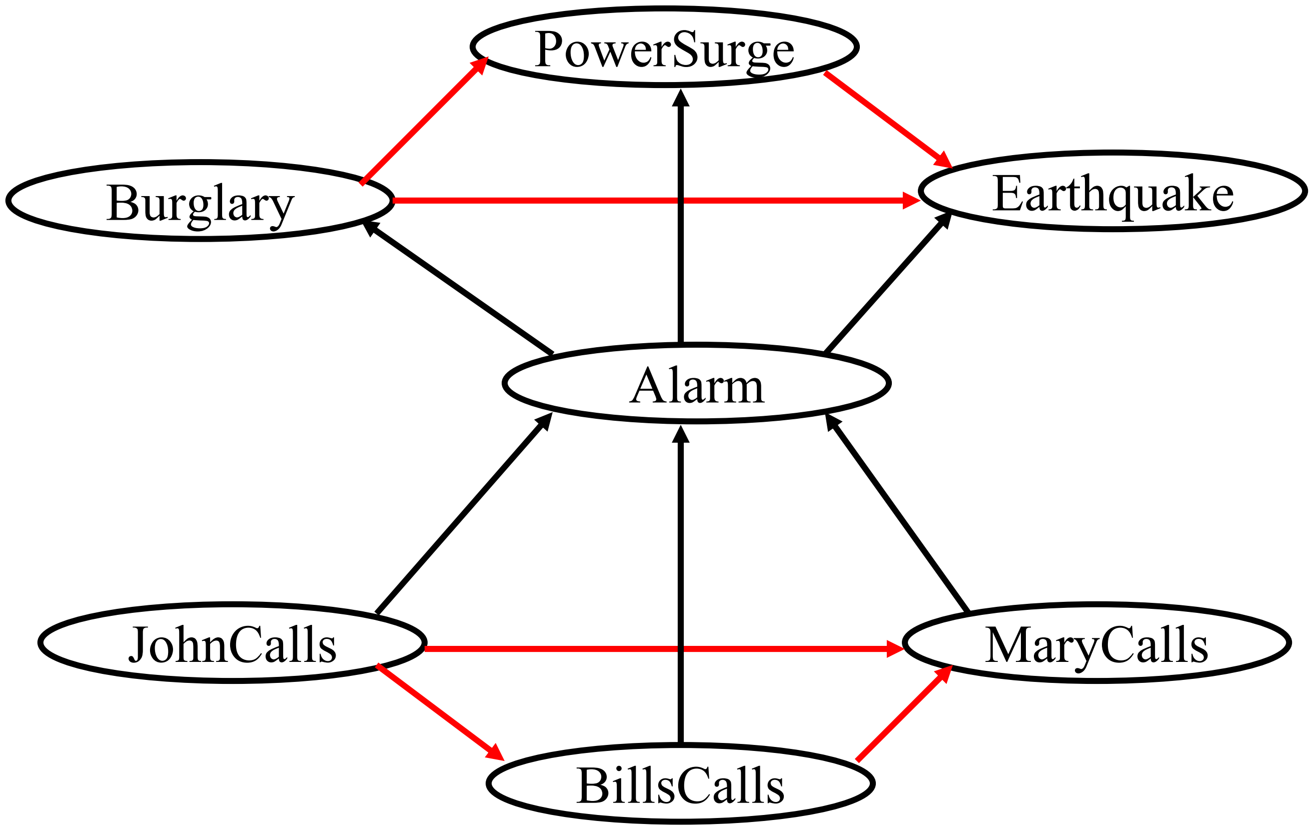
An alternative basis

Suppose we get the direction of causality wrong...



- Another new arrow captures this dependence.
- But again too complex: does not believe that $P(Burglary, Earthquake) = P(Burglary)P(Earthquake)$

Suppose we get the direction of causality wrong...



- Adding more causes or effects requires a combinatorial proliferation of extra arrows. Too general, not modular, too many parameters….

Constructing a Bayes net

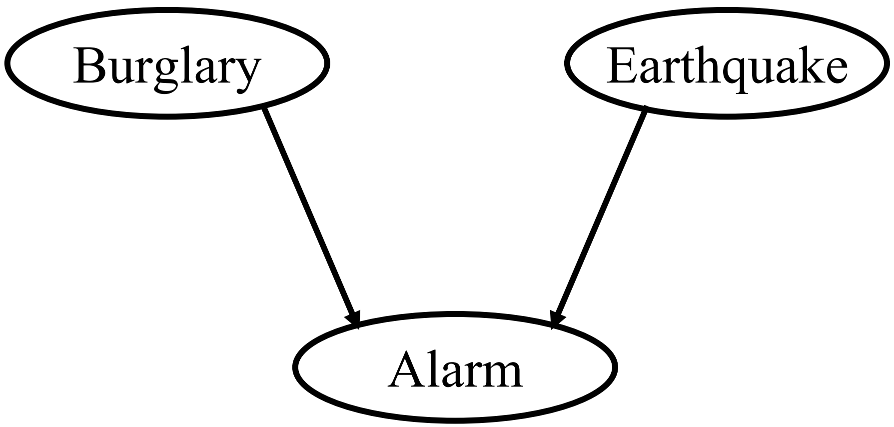
- Model reduces all pairwise dependence and independence relations down to a basic set of pairwise dependencies: graph edges.
- An analogy to learning kinship relations
  - Many possible bases, some better than others
  - A basis corresponding to direct causal mechanisms seems to compress best.
- Finding the minimal dependence structure suggests a basis for learning causal models.

Outline

- The semantics of Bayes nets
  - role of causality in structural compression
- Explaining away revisited
  - role of causality in probabilistic inference
- Sampling algorithms for approximate inference in graphical models

Explaining away

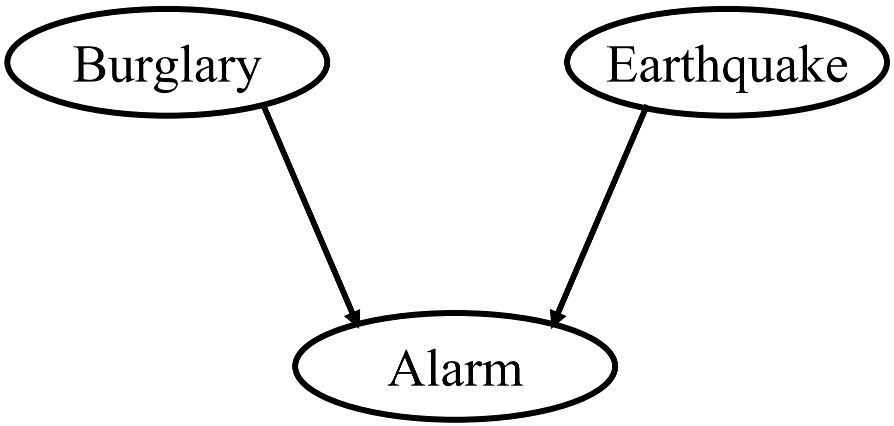
- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

Explaining away

- Logical OR: Independent deterministic causes



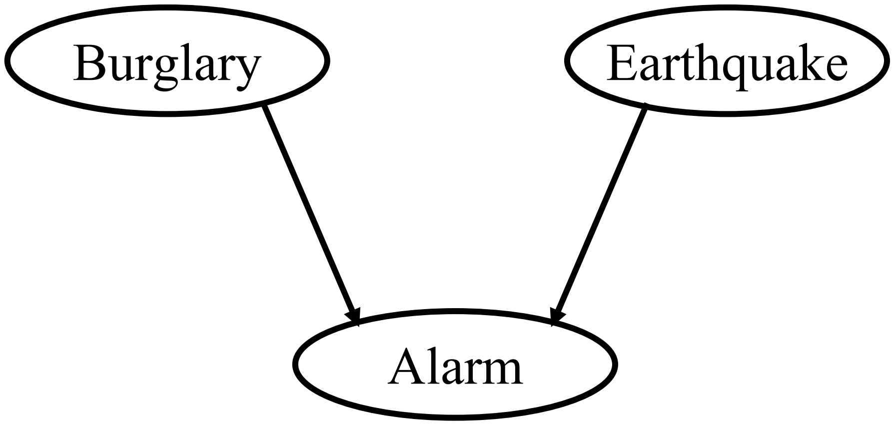
B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

A priori, no correlation between *B* and *E*:

$$P(b,e) = P(b)\ P(e)$$

Explaining away

- Logical OR: Independent deterministic causes



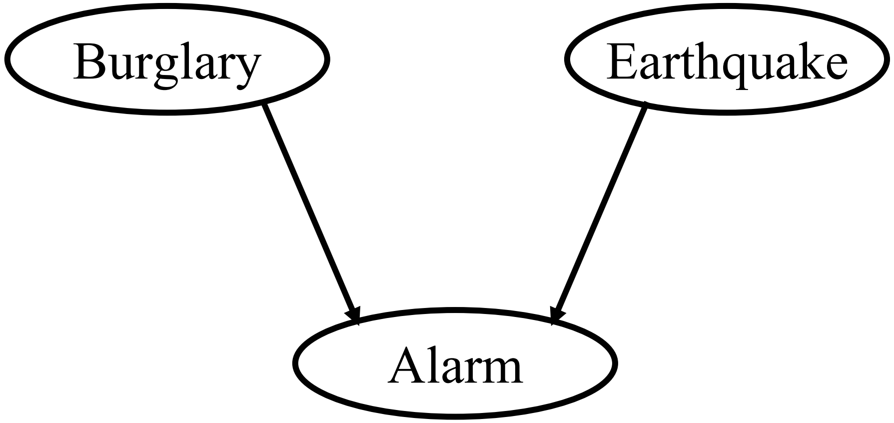
B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing $A = a$ …

$$P(b \mid a) = \frac{\overbrace{P(a \mid b)}^{=1} P(b)}{P(a)}$$

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A \mid B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

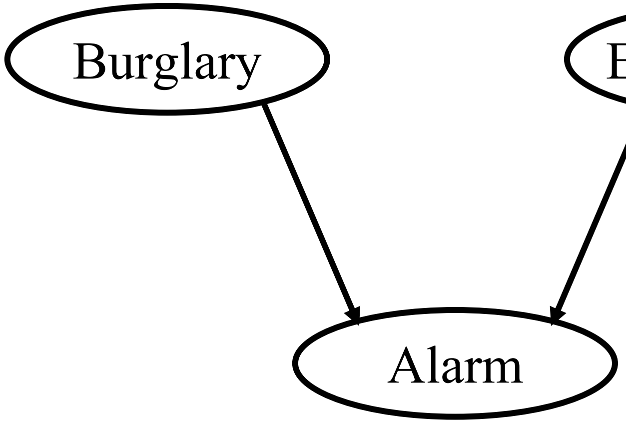
After observing $A = a$ …

$$P(b \mid a) = \frac{P(b)}{P(a)} \quad > P(b)$$

May be a big increase if $P(a)$ is small.

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

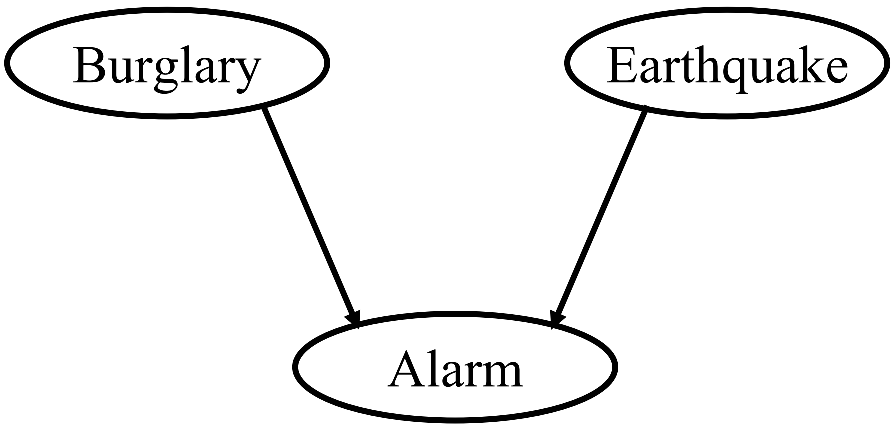
After observing $A = a$ …

$$P(b \mid a) = \frac{P(b)}{P(b) + P(e) - P(b)P(e)} \quad > P(b)$$

May be a big increase if $P(b)$, $P(e)$ are small.

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

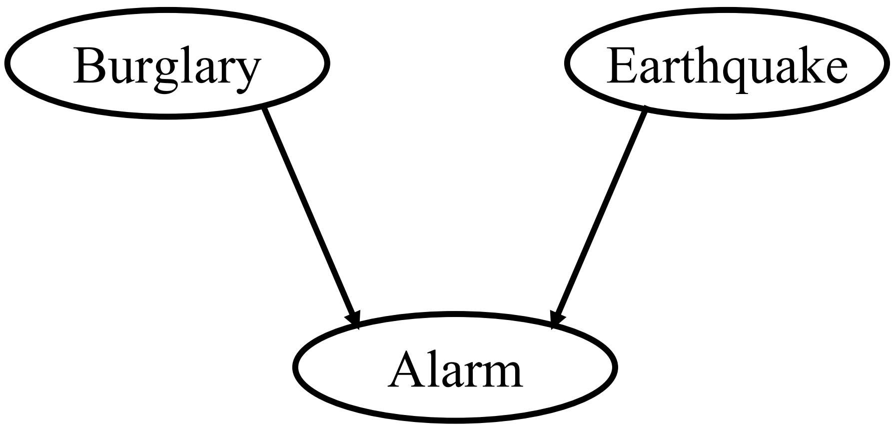
After observing $A = a$, $E = e$, …

$$P(b \mid a, e) = \frac{P(a \mid b, e)\, P(b \mid e)}{P(a \mid e)}$$

Both terms = 1

Explaining away

- Logical OR: Independent deterministic causes



B	E	$P(A\|B,E)$
0	0	0
0	1	1
1	0	1
1	1	1

After observing $A = a$, $E = e$, …

$$P(b \mid a, e) = \frac{P(a \mid b, e)\, P(b \mid e)}{P(a \mid e)}$$

$$= P(b \mid e) = P(b)$$

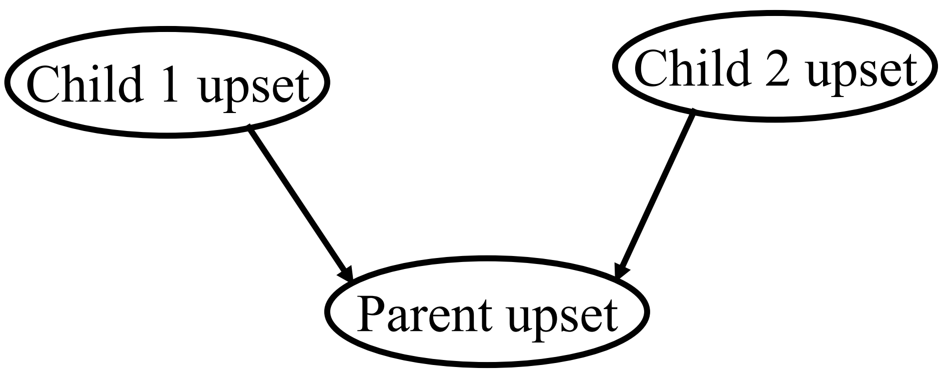
"Explaining away" or
"Causal discounting"

Explaining away

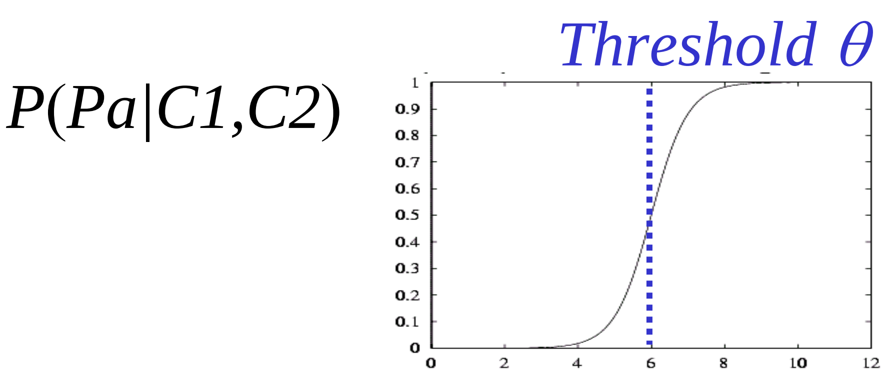
- Depends on the functional form (the parameterization) of the CPT
  - OR or Noisy-OR: Discounting
  - AND: No Discounting
  - Logistic: Discounting from parents with positive weight; augmenting from parents with negative weight.
  - Generic CPT: Parents become dependent when conditioning on a common child.

Parameterizing the CPT

- Logistic: Independent probabilistic causes with varying strengths w_i and a threshold θ



$C1$	$C2$	$P(Pa \mid C1,C2)$
0	0	$1/[1+\exp(\theta)]$
0	1	$1/[1+\exp(\theta - w_1)]$
1	0	$1/[1+\exp(\theta - w_2)]$
1	1	$1/[1+\exp(\theta - w_1 - w_2)]$



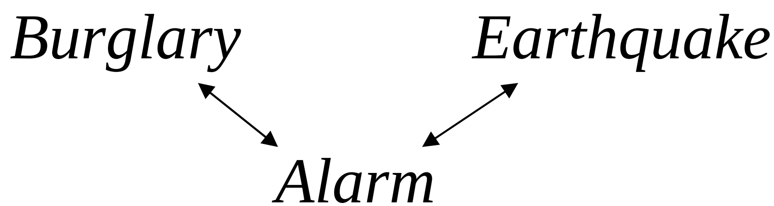
$Annoyance = C1 * w_1 + C2 * w_2$

Contrast w/ conditional reasoning

$$\begin{tikzcd}
\textit{Rain} \arrow[dr] & & \textit{Sprinkler} \arrow[dl] \\
& \textit{Grass Wet} &
\end{tikzcd}$$

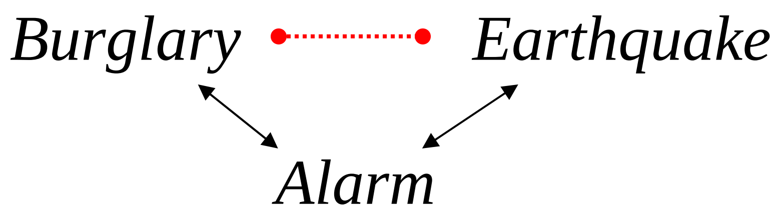
- Formulate IF-THEN rules:
  - IF *Rain* THEN *Wet*
  - ~~IF *Wet* THEN *Rain*~~ IF *Wet* AND NOT *Sprinkler* THEN *Rain*
- Rules do not distinguish directions of inference
- Requires combinatorial explosion of rules

Spreading activation or recurrent neural networks



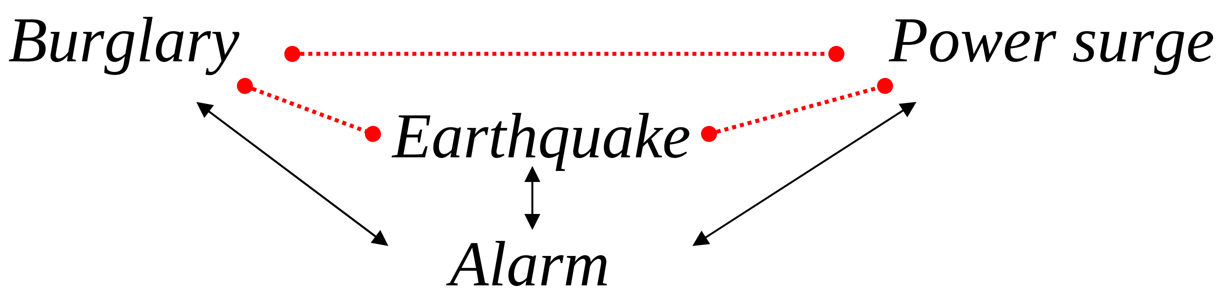
- Excitatory links: *Burglary* $\longleftrightarrow$ *Alarm*, *Earthquake* $\longleftrightarrow$ *Alarm*
- Observing earthquake, *Alarm* becomes more active.
- Observing alarm, *Burglary* and *Earthquake* become more active.
- Observing alarm and earthquake, *Burglary* cannot become less active. No explaining away!

Spreading activation or recurrent neural networks



- Excitatory links: *Burglary* $\longleftrightarrow$ *Alarm*, *Earthquake* $\longleftrightarrow$ *Alarm*
- Inhibitory link: *Burglar* *Earthquake*
- Observing alarm, *Burglary* and *Earthquake* become more active.
- Observing alarm and earthquake, *Burglary* becomes less active: explaining away.

Spreading activation or recurrent neural networks



- Each new variable requires more inhibitory connections.
- Interactions between variables are not causal.
- Not modular.
  – Whether a connection exists depends on what other connections exist, in non-transparent ways.
  – Combinatorial explosion of connections

The relation between PDP and Bayes nets

- To what extent does Bayes net inference capture insights of the PDP approach?
- To what extent do PDP networks capture or approximate Bayes nets?

Summary

Bayes nets, or directed graphical models, offer a powerful representation for large probability distributions:

- Ensure tractable storage, inference, and learning
- Capture causal structure in the world and canonical patterns of causal reasoning.
- This combination is not a coincidence.

Still to come

- Applications to models of categorization
- More on the relation between causality and probability:

$$\text{Causal structure} \downarrow \text{Statistical dependencies}$$

- Learning causal graph structures.
- Learning causal abstractions ("diseases cause symptoms")
- What's missing from graphical models

Outline

- The semantics of Bayes nets
  - role of causality in structural compression
- Explaining away revisited
  - role of causality in probabilistic inference
- Sampling algorithms for approximate inference in graphical models

Motivation

- What is the problem of inference?
  - Reasoning from observed variables to unobserved variables
    - Effects to causes (diagnosis):

      $$P(Burglary = 1 | JohnCalls = 1, MaryCalls = 0)$$

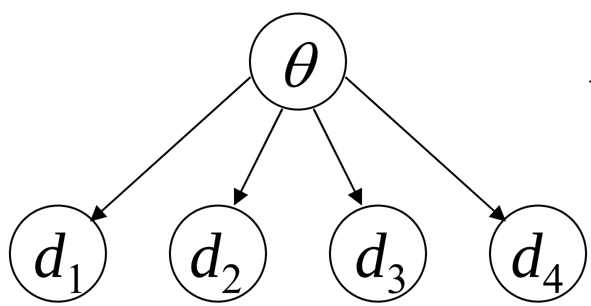
    - Causes to effects (prediction):

      $$P(JohnCalls = 1 | Burglary = 1)$$

      $$P(JohnCalls = 0, MaryCalls = 0 | Burglary = 1)$$

Motivation

- What is the problem of inference?
  - Reasoning from observed variables to unobserved variables.
  - Learning, where hypotheses are represented by unobserved variables.
    - e.g., Parameter estimation in coin flipping:

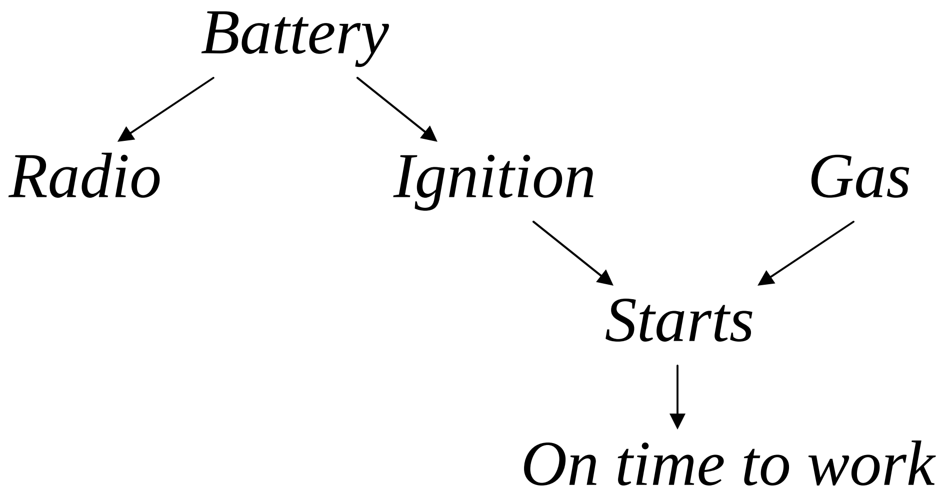


$P(\mathrm{H}) = \theta$

Motivation

- What is the problem of inference?
  - Reasoning from observed variables to unobserved variables.
  - Learning, where hypotheses are represented by unobserved variables.
- Why is it hard?
  - In principle, must consider all possible states of all variables connecting input and output variables.

A more complex system



- Joint distribution sufficient for any inference:

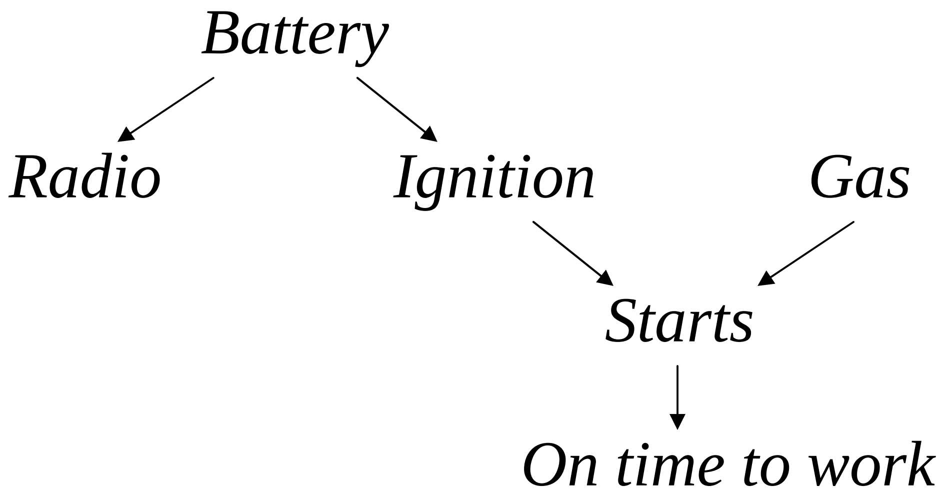
$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \frac{\sum_{B,R,I,S} P(B,R,I,G,S,O)}{P(G)}$$

$$\boxed{P(A) = \sum_{B} P(A,B)}$$

"marginalization"

A more complex system

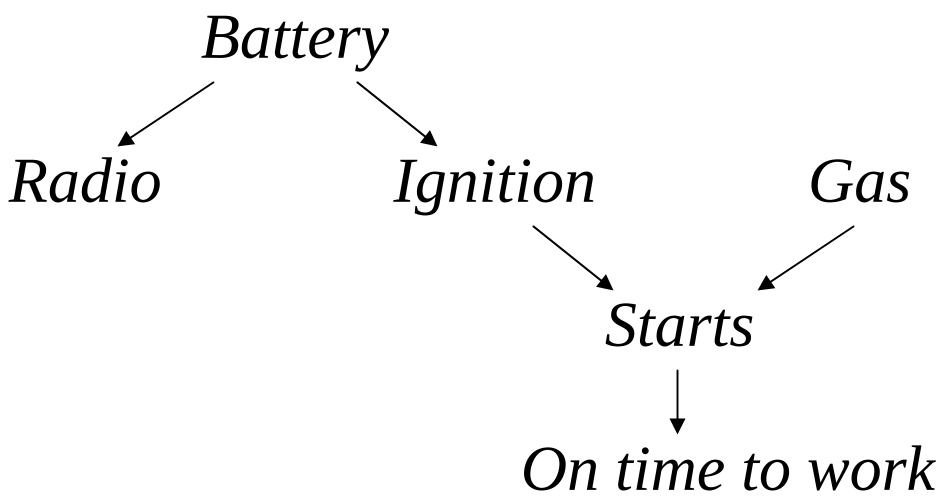


- Joint distribution sufficient for any inference:

$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \frac{\sum_{B,R,I,S} P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)}{P(G)}$$

A more complex system

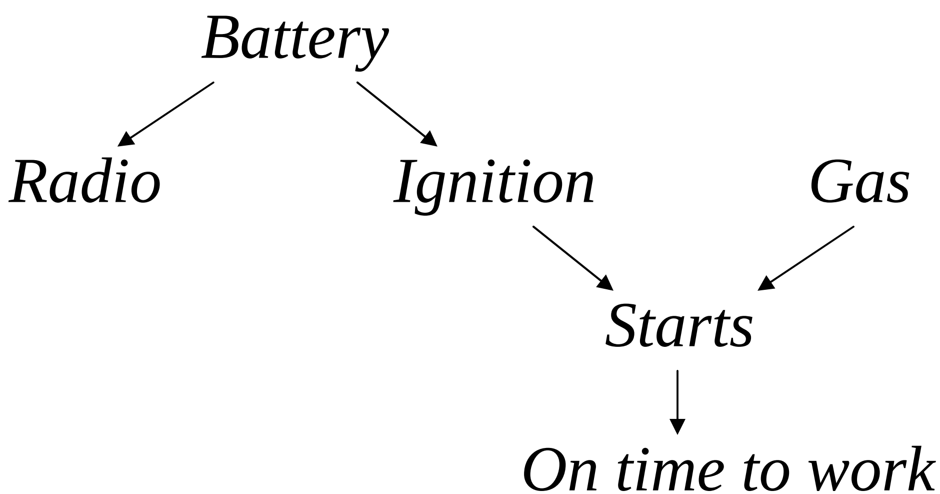


- Joint distribution sufficient for any inference:

$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

$$P(O \mid G) = \frac{P(O,G)}{P(G)} = \sum_{S} \left(\sum_{B,I} P(B)P(I \mid B)P(S \mid I,G) \right) P(O \mid S)$$

A more complex system



- Joint distribution sufficient for any inference:

$$P(B,R,I,G,S,O) = P(B)P(R \mid B)P(I \mid B)P(G)P(S \mid I,G)P(O \mid S)$$

- Exact inference algorithms via local computations
  – for graphs without loops: belief propagation
  – in general: variable elimination or junction tree, but these will still take exponential time for complex graphs.

Sampling possible worlds

$< cloudy, \neg sprinkler, rain, wet >$

$< \neg cloudy, sprinkler, \neg rain, wet >$

$< \neg cloudy, sprinkler, \neg rain, wet >$

$< \neg cloudy, sprinkler, \neg rain, wet >$

$< cloudy, \neg sprinkler, \neg rain, \neg wet >$

$< cloudy, \neg sprinkler, rain, wet >$

$< \neg cloudy, \neg sprinkler, \neg rain, \neg wet >$

. . .

As the sample gets larger, the frequency of each possible world approaches its true prior probability under the model.

How do we use these samples for inference?

Summary

- Exact inference methods do not scale well to large, complex networks
- Sampling-based approximation algorithms can solve inference and learning problems in arbitrary networks, and may have some cognitive reality.
  - Rejection sampling, Likelihood weighting
    - Cognitive correlate: imagining possible worlds
  - Gibbs sampling
    - Neural correlate: Parallel local message-passing dynamical system
    - Cognitive correlate: “Two steps forward, one step back” model of cognitive development