Inference in Bayesian Networks

CMPSCI 383
Nov 1, 2011

Today's topics: exact and approximate inference

- Exact
  - Inference with joint probability distributions
  - Exact inference in Bayesian networks
  - Inference by enumeration
  - Complexity of exact inference
- Approximate
  - Inference by stochastic simulation
  - Simple sampling
  - Rejection sampling
  - Markov chain Monte Carlo (MCMC)

Inference terminology

- **Conditional probability table:** data structure that lists probabilities of a variable given one or more other variables.

- **Joint distribution:**
  distribution that is specified by a Bayesian network

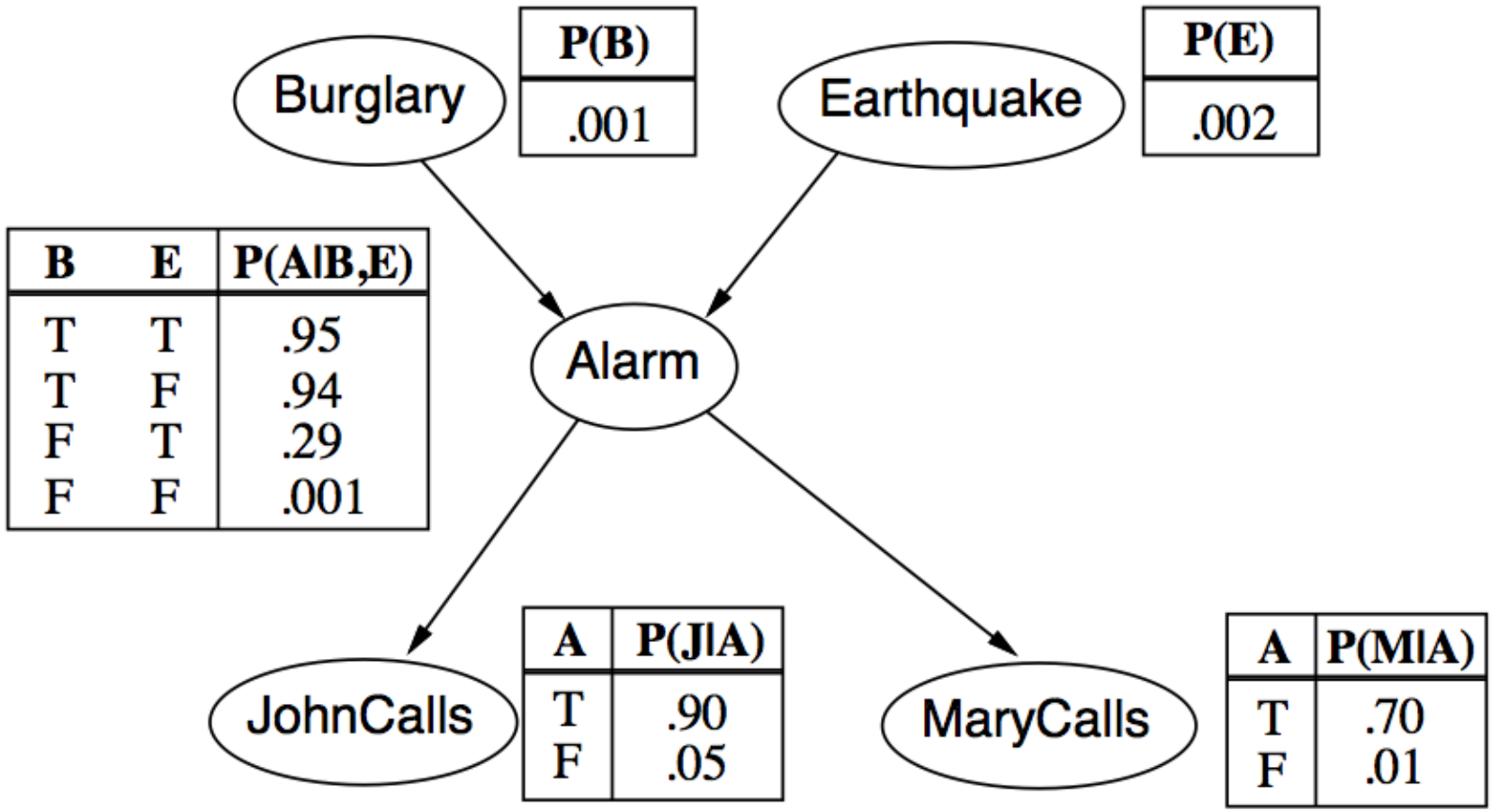
- **Inference**: produces the probability distribution of one or more variables given one or more other variables.

Example: Joint distribution

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

V = Cavity; T = Toothache; C = Catch

Example: Home security



Compact conditional distributions

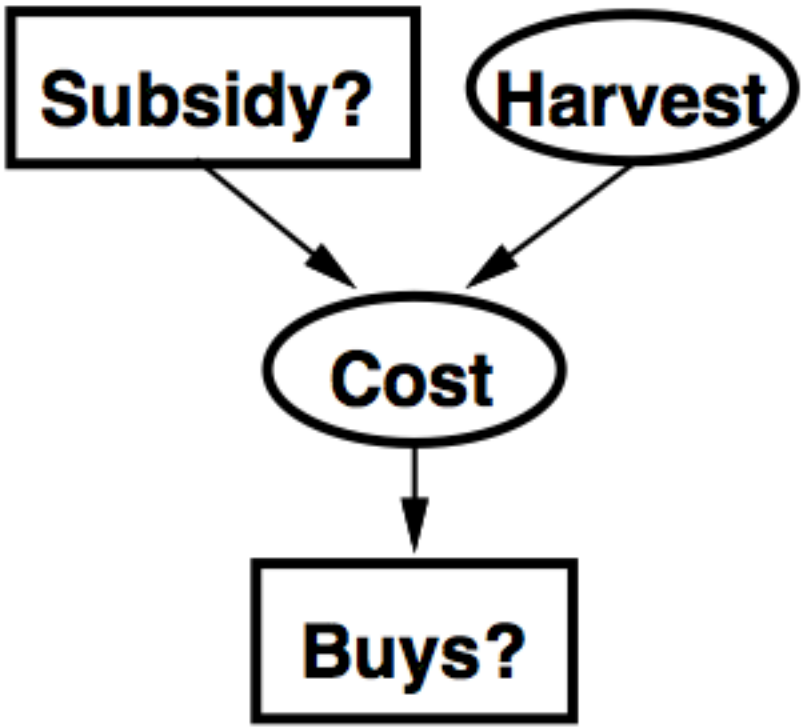
- Even conditional probability tables can be quite large
- Combining functions — that relate the value of the parents to the value of the child — is one way of reducing their size
- Example (for discrete variables): Noisy-OR

"inhibition probabilities"

$Cold$	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	$0.012 = 0.6 \times 0.2 \times 0.1$

Additional complexities: Mixed-mode nets

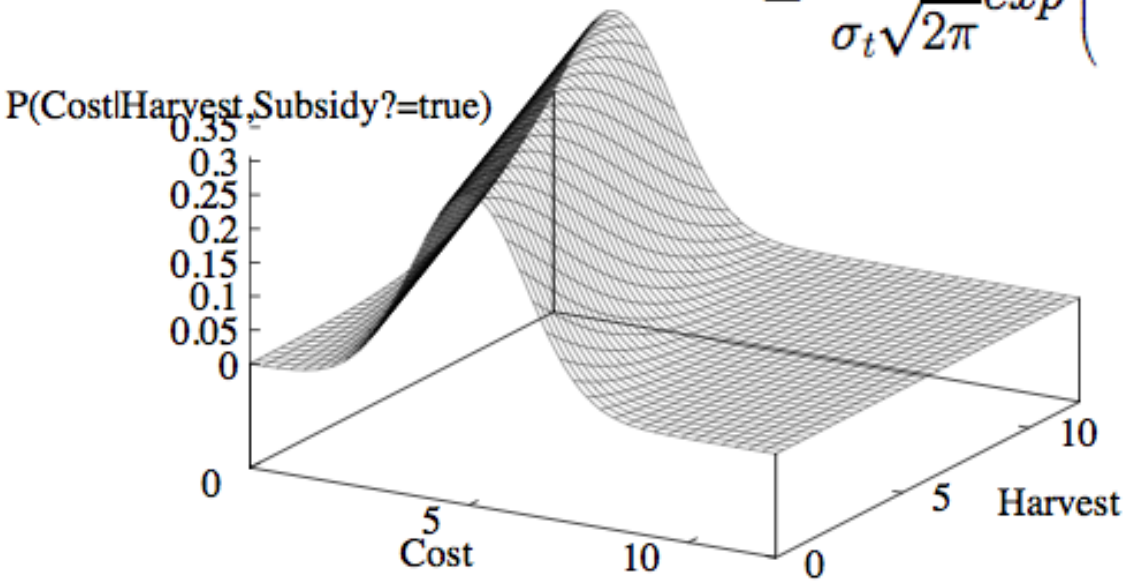
- We discussed how to handle discrete variables, but BNs can be used to represent and reason about a variety of variable types



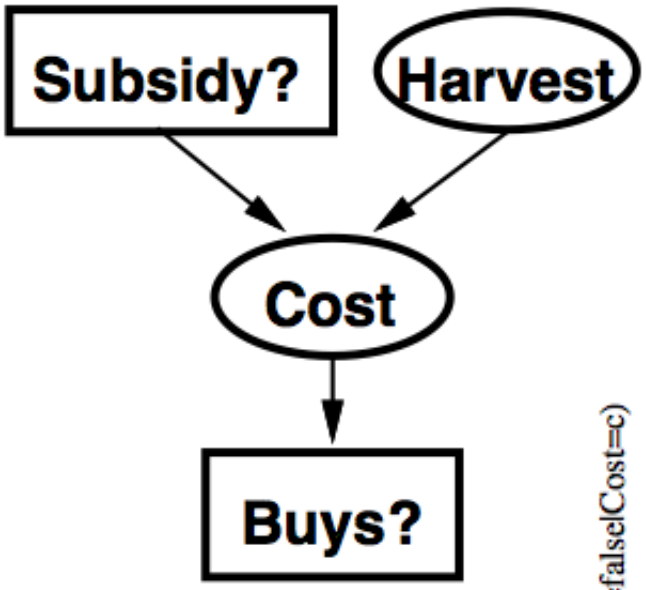
Compact conditional distributions

- For continuous variables, we can assume some linear functional dependence among the variables.
- For example, if Cost depends on Harvest and subsidy, for each value of subsidy...

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

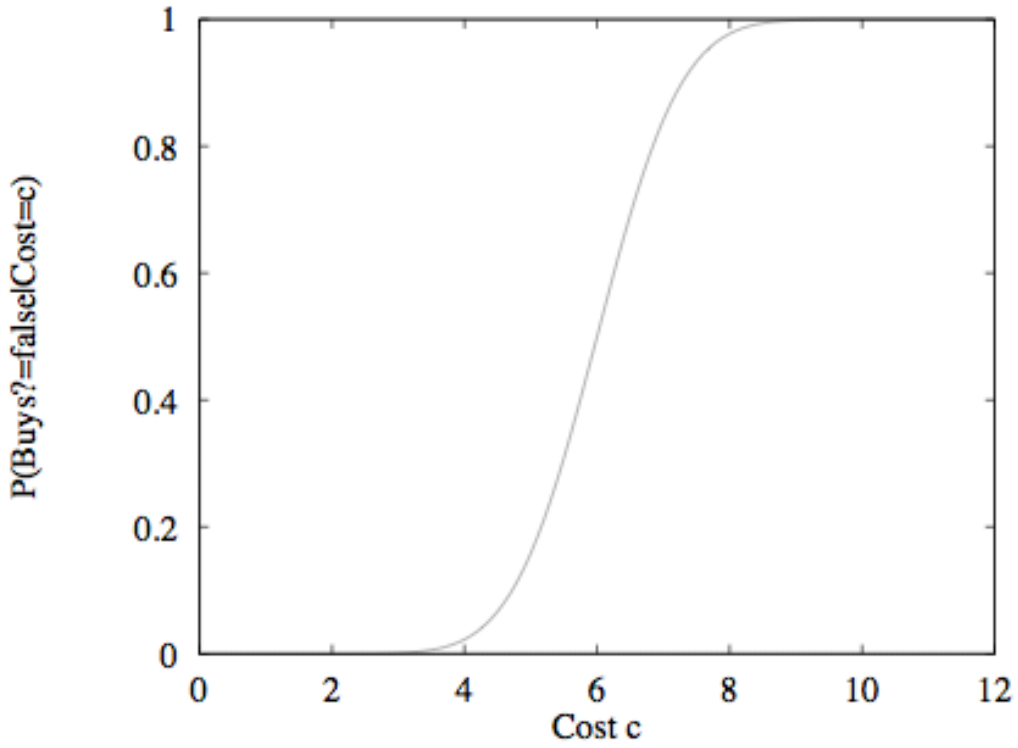


Compact conditional distributions



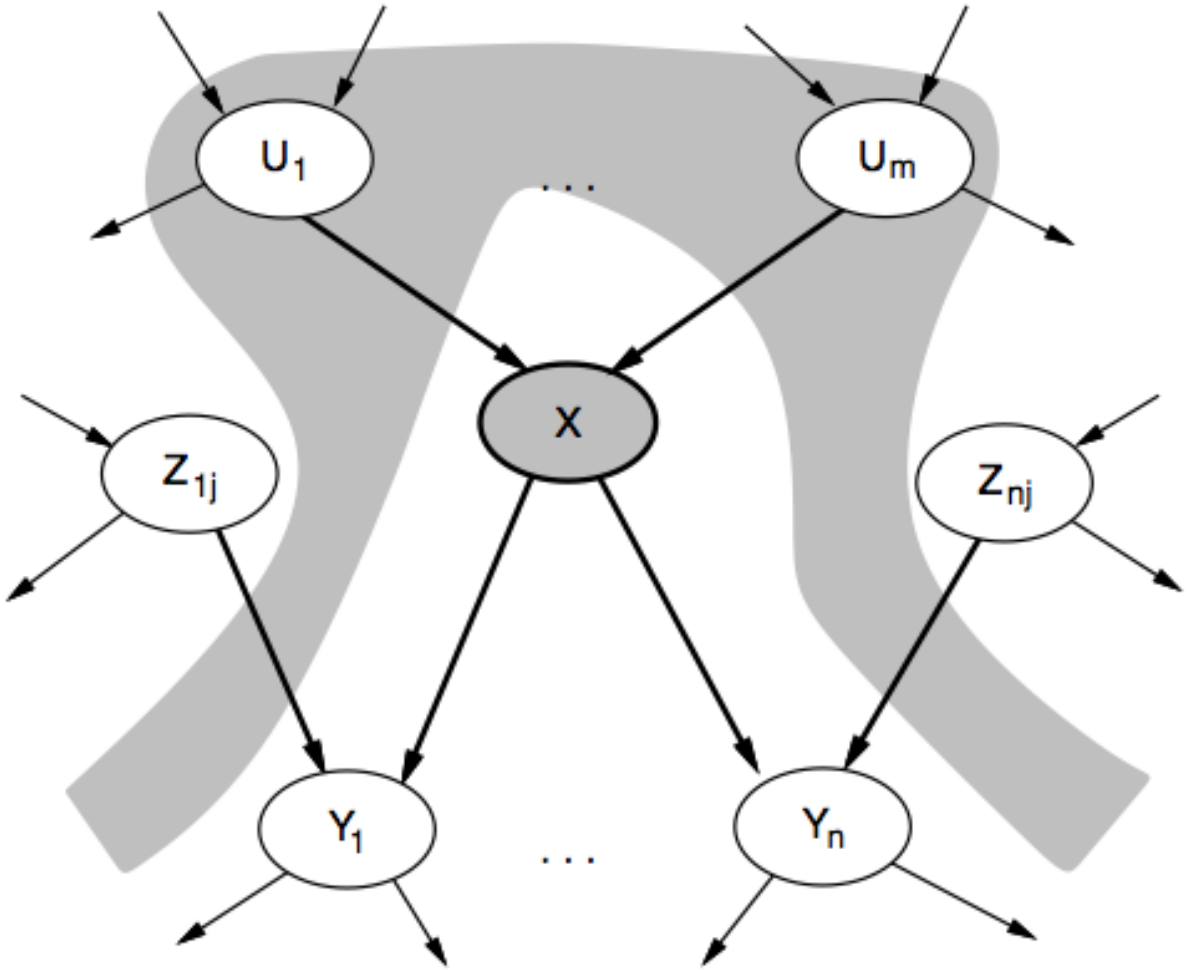
$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$

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

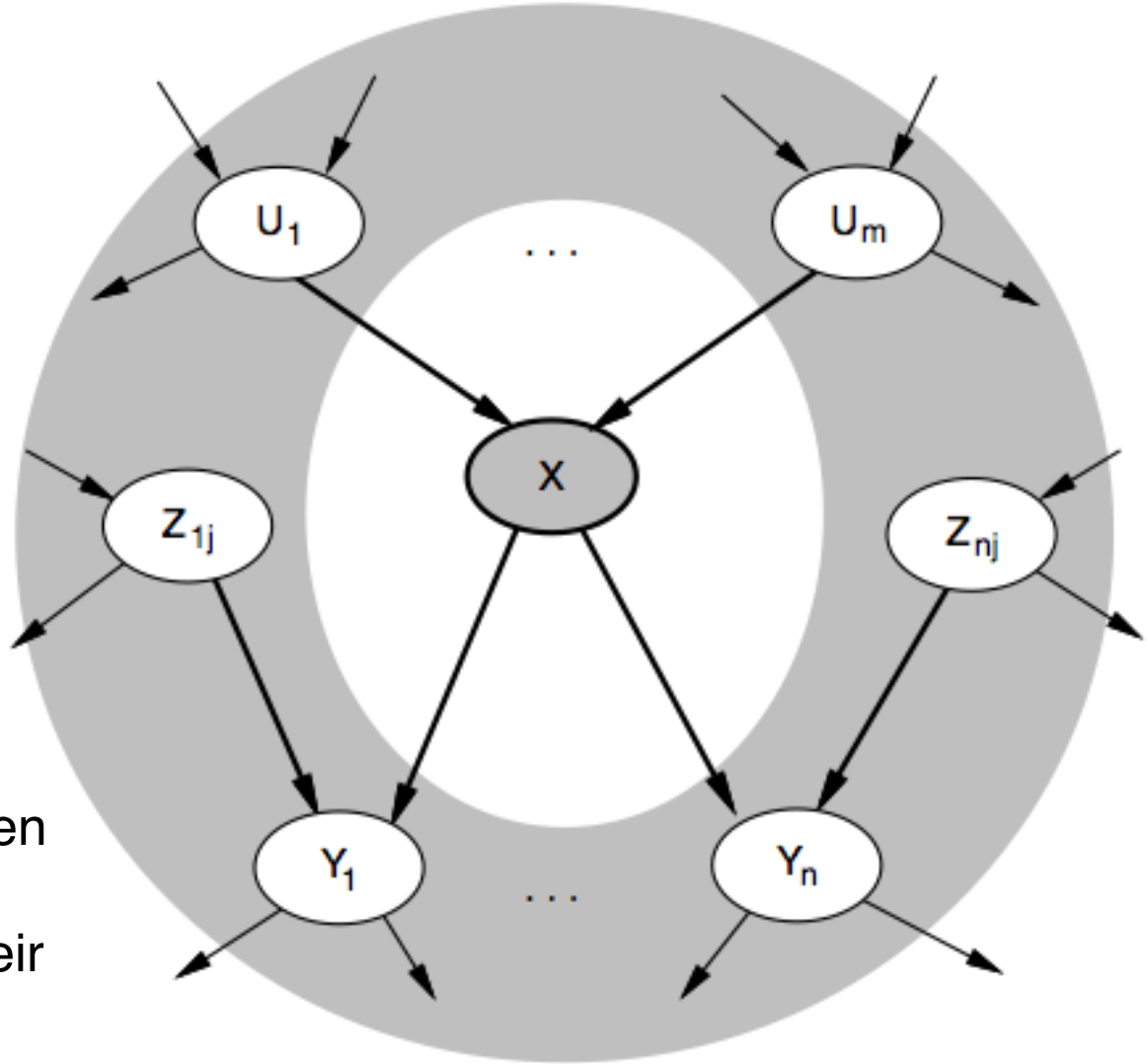


Conditional Independence

Node X is conditionally independent of its non-descendants given its parents.

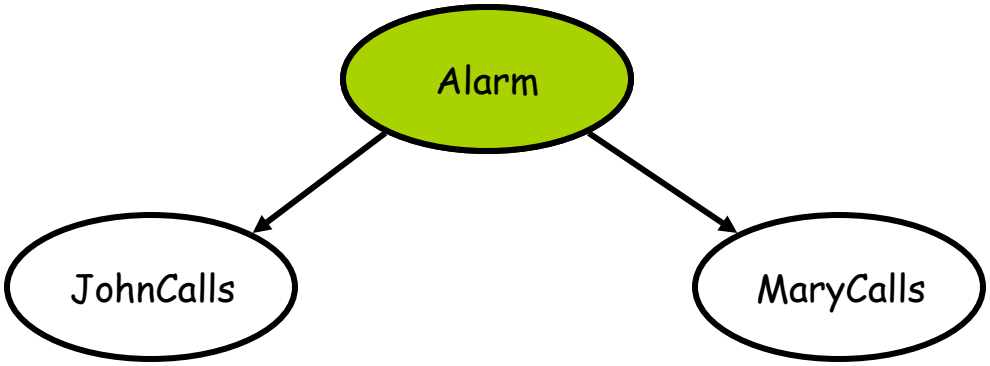


Conditional Independence



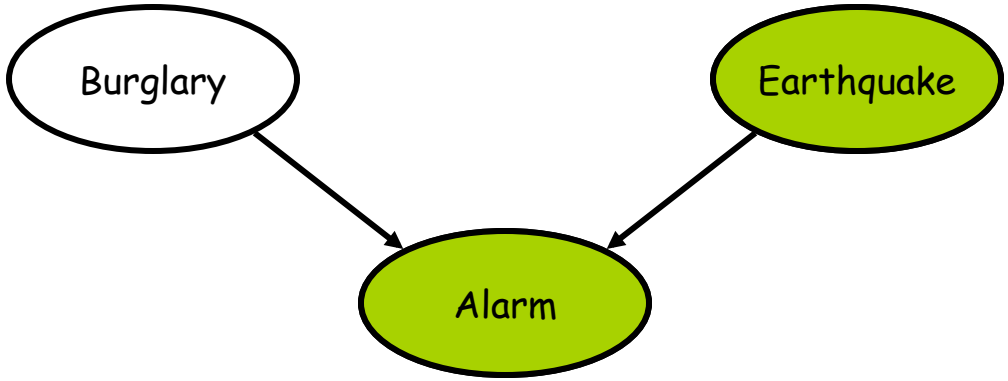
Node X is conditionally independent of all other nodes in the network given its “Markov blanket” (its parents, children, and their parents).

Conditional independence (revisited)



- Are *JohnCalls* and *MaryCalls* independent?
  - No, they are not completely independent
  - Whether they are independent is *conditional* on the value of *Alarm*
- If the value of *Alarm* is known, are *JohnCalls* and *MaryCalls* independent?
  - Yes, for each known value of *A, J* and *M* are independent

Conditional independence (revisited)

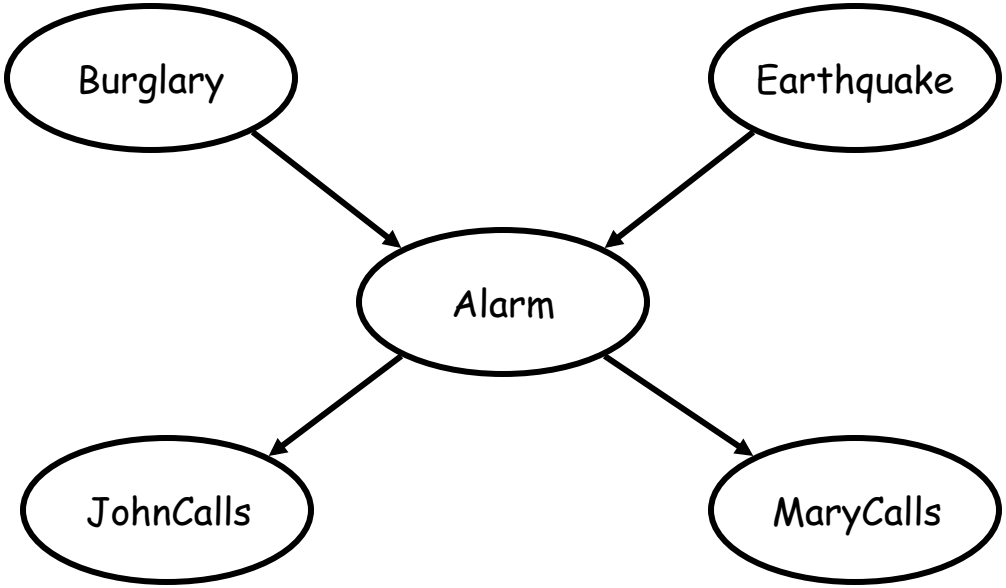


- Are *Burglary* and *Earthquake* cond. independent?
  - Yes, nodes are conditionally independent of their non-descendents given their parents
- Are they completely independent?
  - No, one can 'explain away' the other if *Alarm* is known.

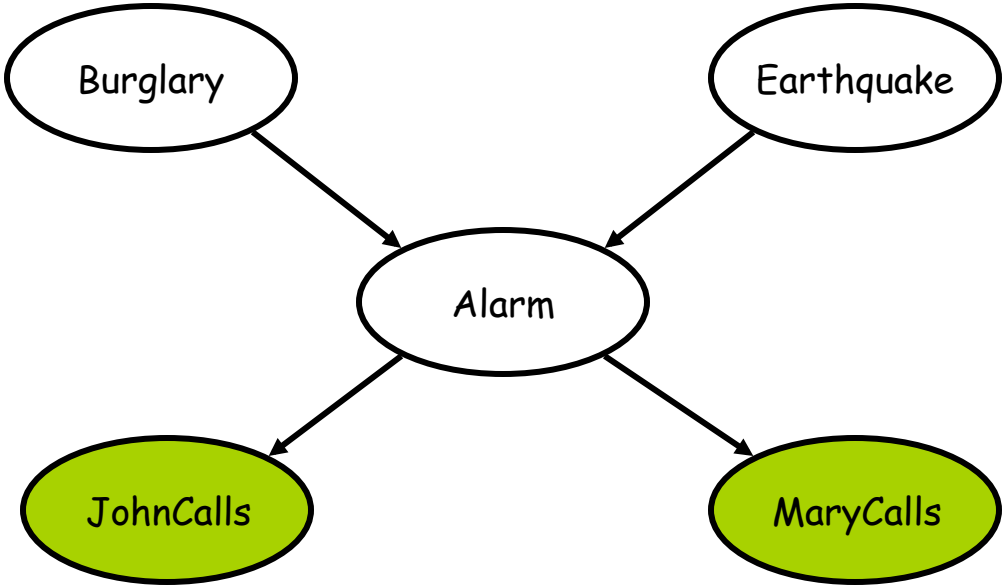
Inference tasks

- Simple queries
  - Compute posterior marginal $P(X_i|E=e)$
  - $P(NoGas|Gauge=empty, Lights=on, Starts=false)$
- Conjunctive queries
  - $P(X_i,X_j|E=e)$
- Optimal decisions
  - Need utility information, but also need $P(outcome \mid action, evidence)$
- Value of information — "What info do I need now?"
- Sensitivity analysis — "Which values matter most?"
- Explanation — "Why do I need a new starter?"

Example: Home security

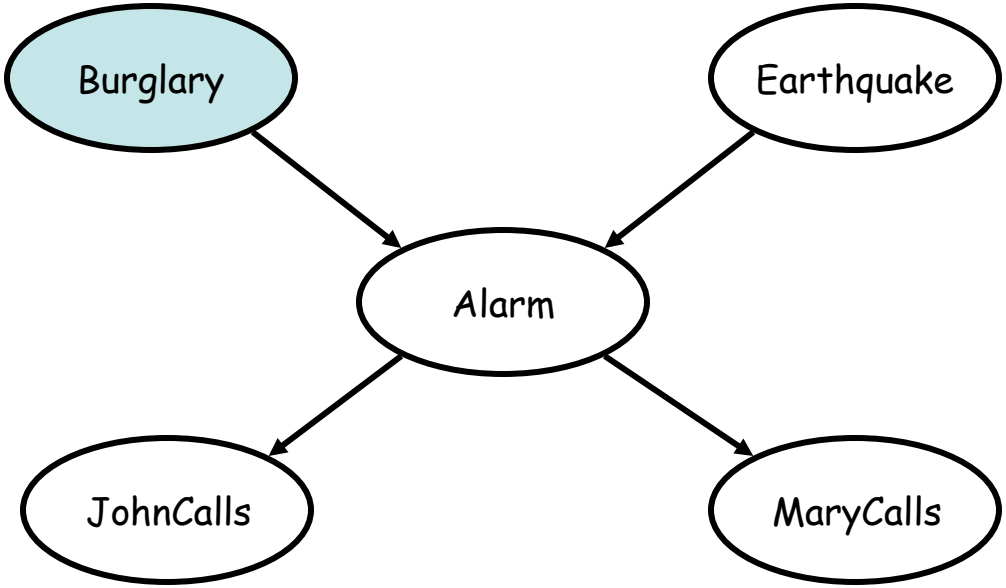


Types of nodes in inference



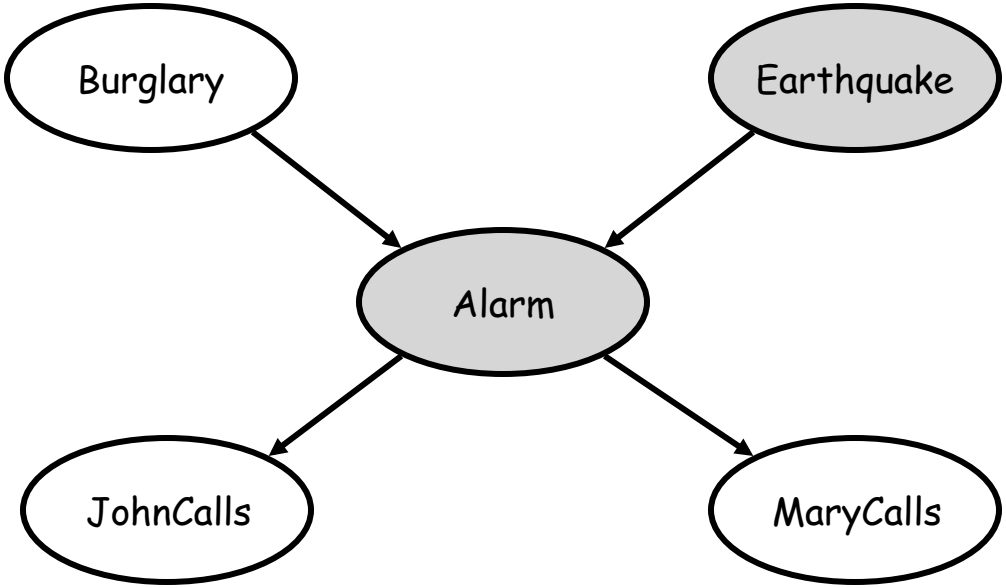
Evidence (or "observed") variables

Types of nodes in inference



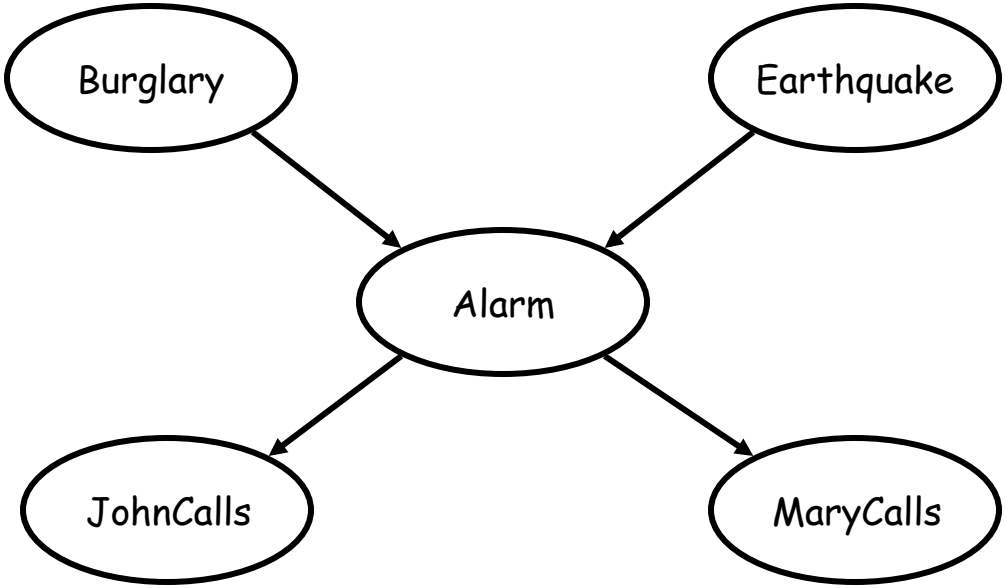
Query variables

Types of nodes in inference

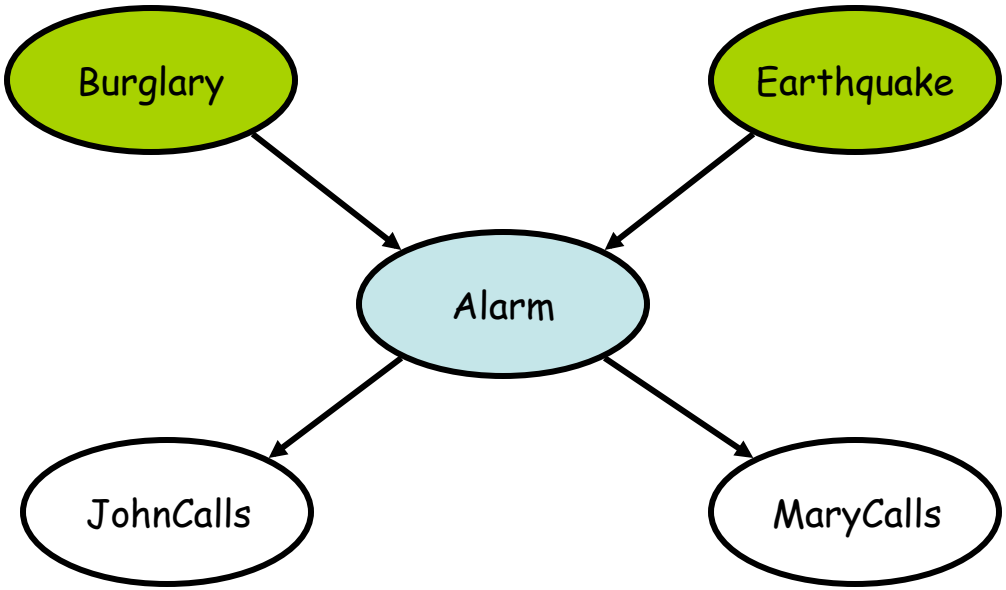


Hidden variables

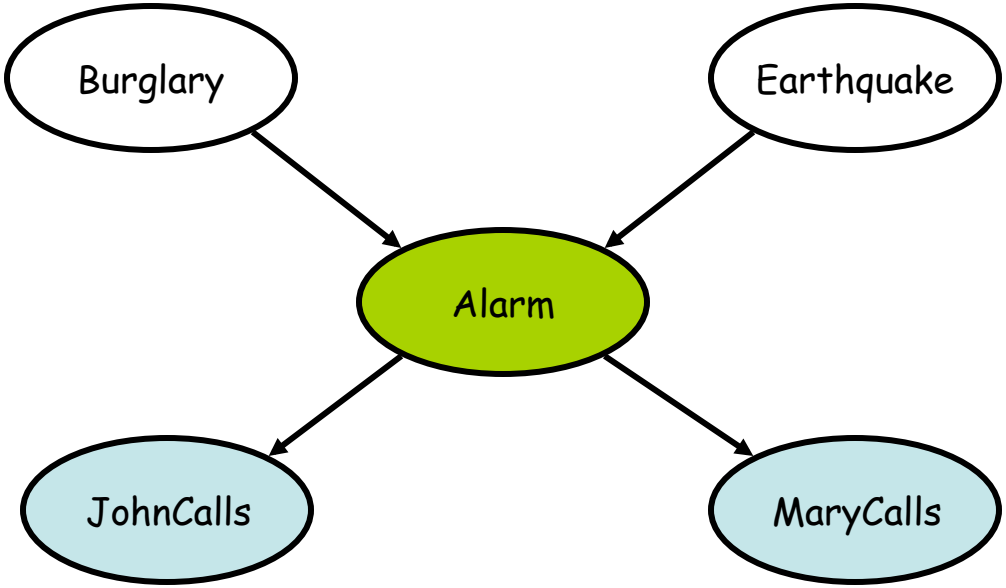
Simple inferences



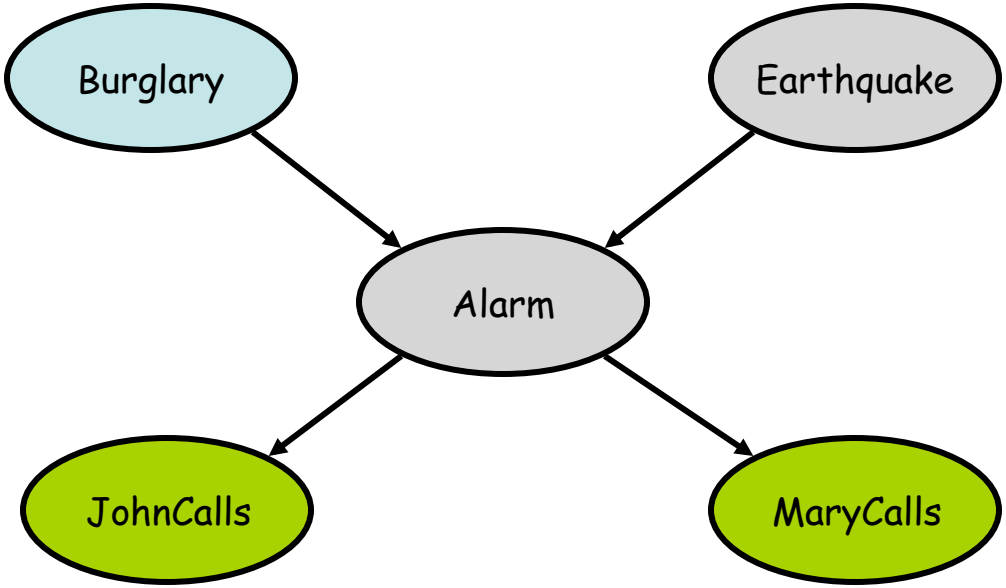
Simple inferences



Simple inferences



More difficult inferences

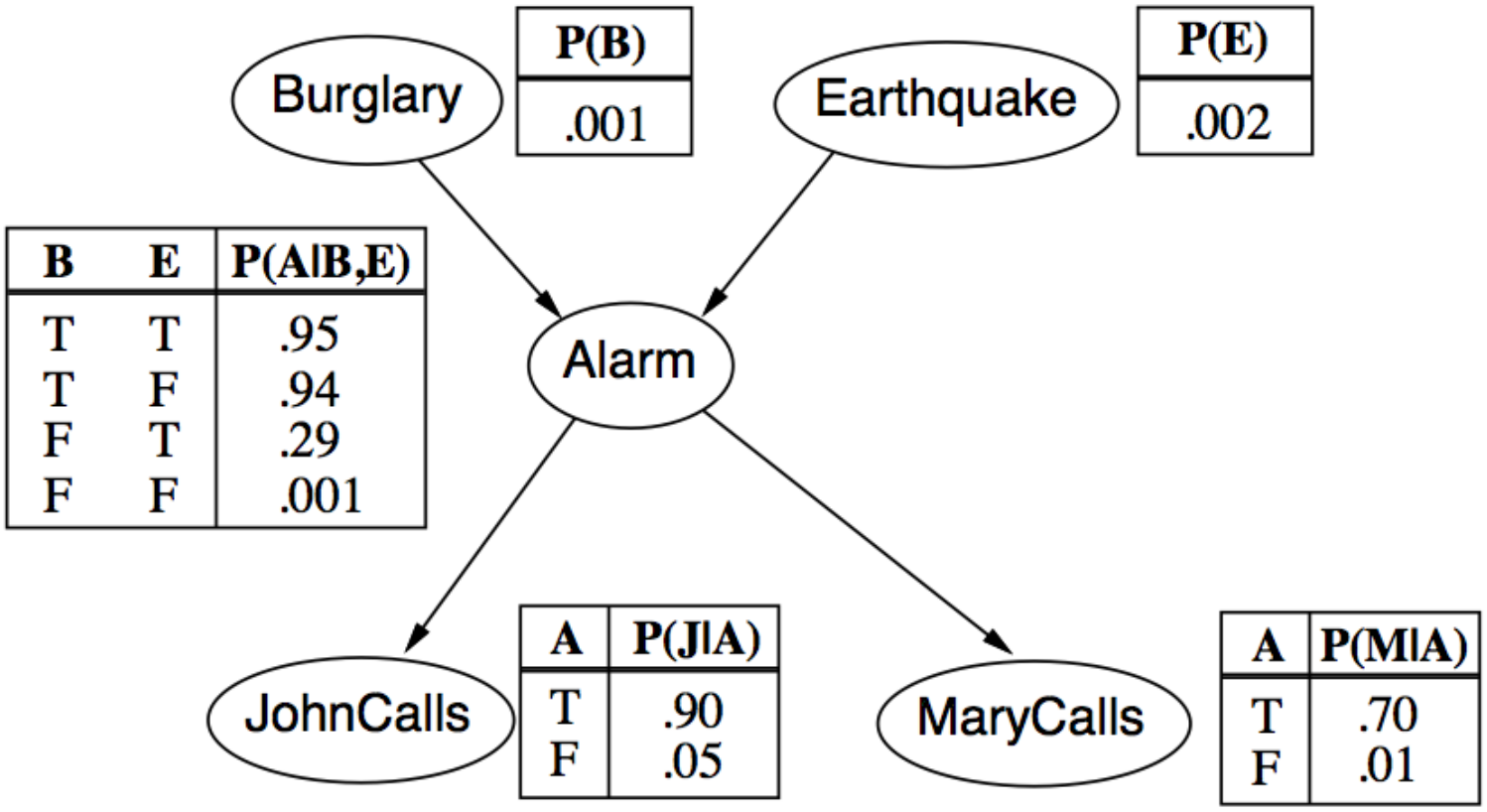


It's easy with the full joint distribution

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

V = Cavity; T = Toothache; C = Catch

...don't we have the full joint distribution?

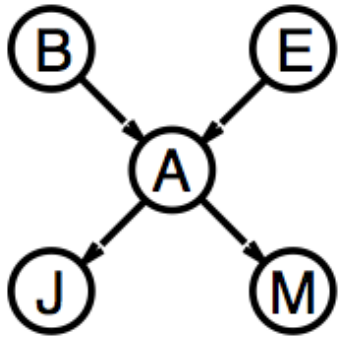


But how do we use it?

Inference by enumeration

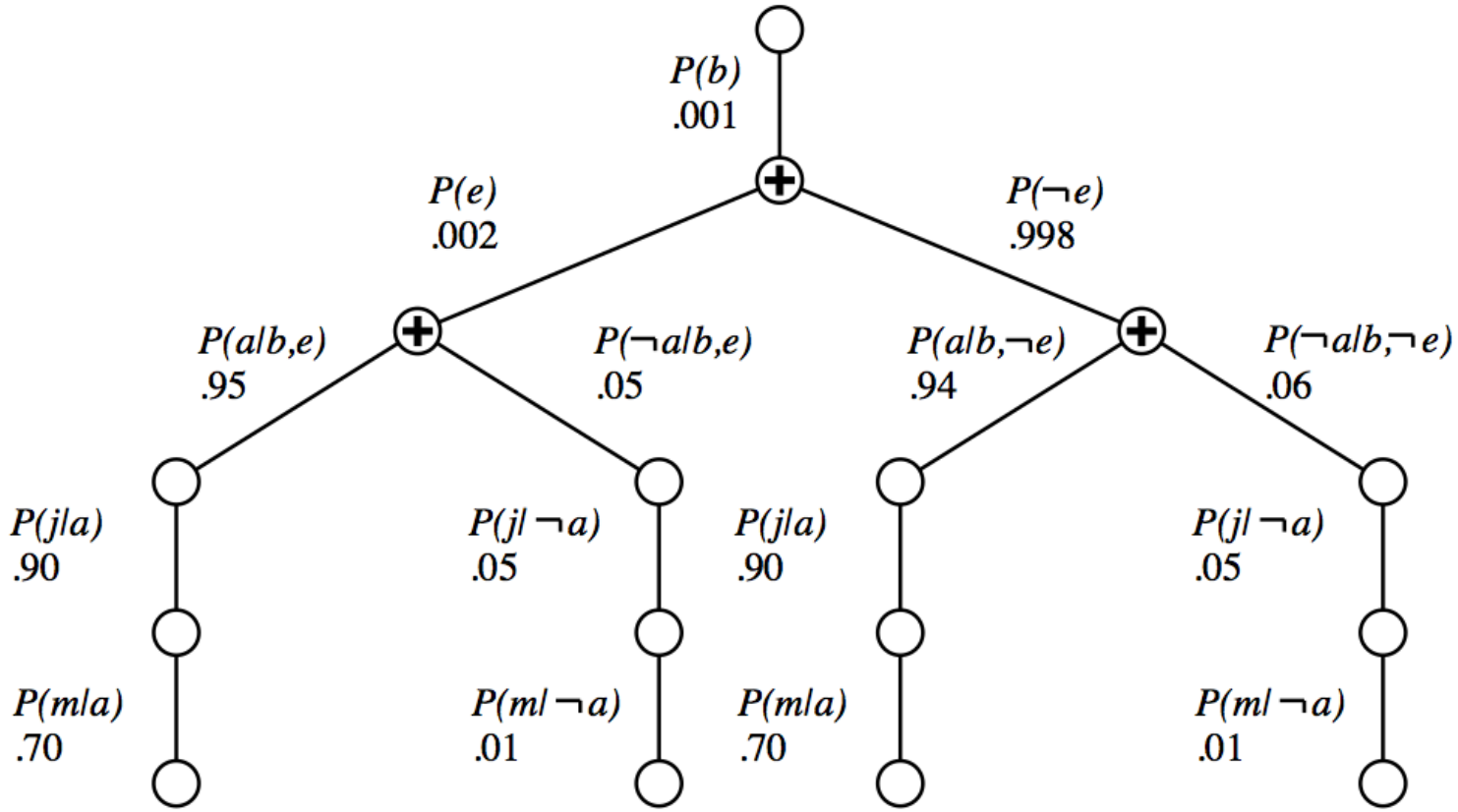
Recursive depth-first enumeration — $O(d^n)$

$$\mathbf{P}(B|j,m)$$



P(B|j,m) = <0.284, 0.716>

Enumeration tree

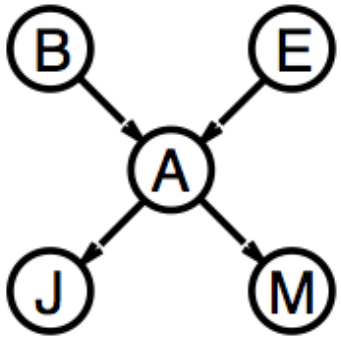


What's inefficient about this?

Removing inefficiencies

- Two problems with enumeration
  - Repeated computation
  - Irrelevant variables (any node that is not an ancestor of a query or evidence node)
- However, repeated computations can be eliminated by variable elimination

Removing irrelevant variables



- Consider the query
  P(JohnCalls|Burglary=true)

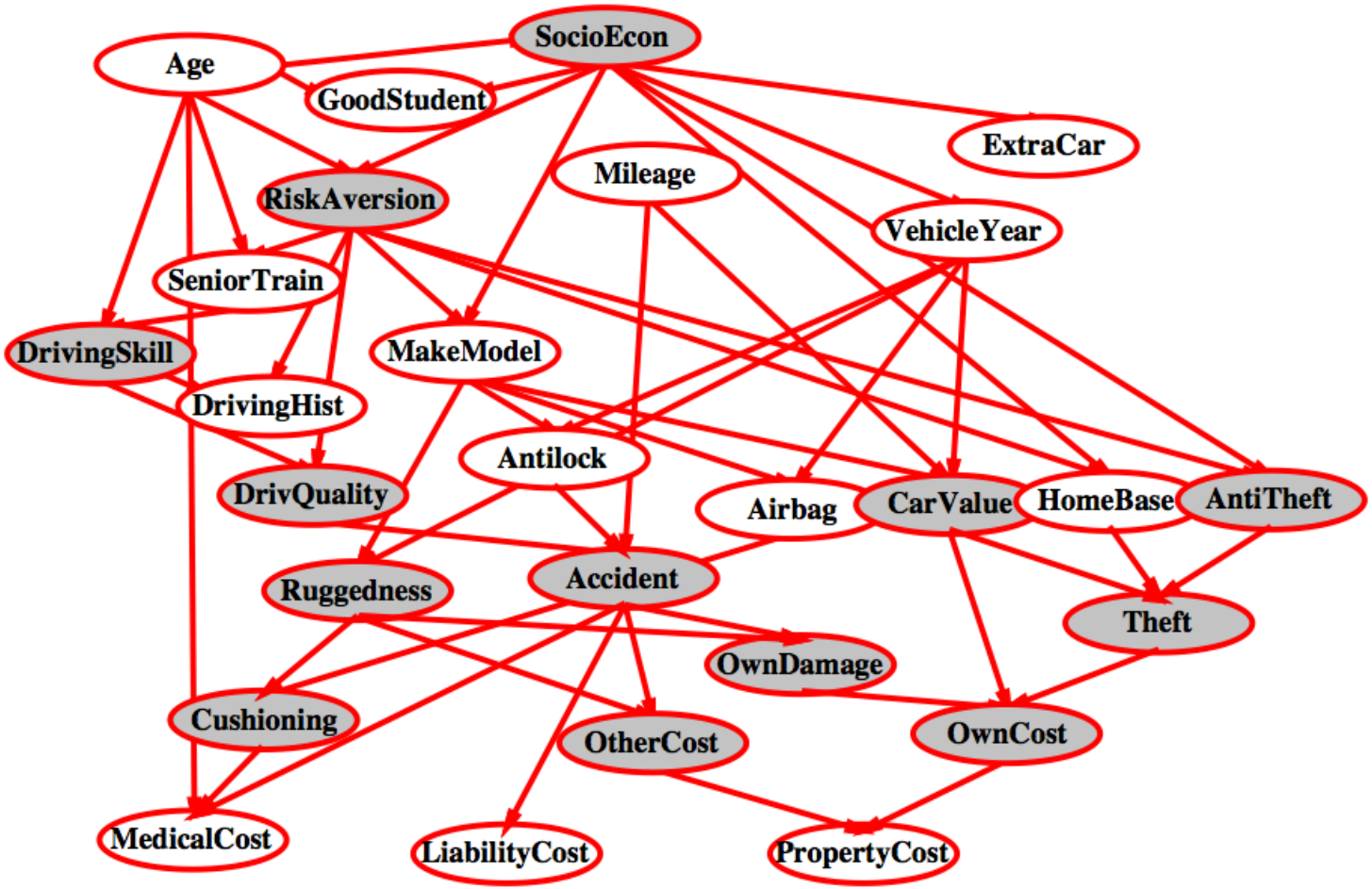
$$P(J|b) = \alpha P(b) \sum_e P(e) \sum_a P(a|b,e) P(J|a) \sum_m P(m|a)$$

- Summing over *m* is identically 1
  M is irrelevant to this specific query

Thm 1: Y is irrelevant unless $Y \in Ancestors(\{X\} \cup \mathbf{E})$

Here, $X = JohnCalls$, $\mathbf{E} = \{Burglary\}$, and
$Ancestors(\{X\} \cup \mathbf{E}) = \{Alarm, Earthquake\}$
so $MaryCalls$ is irrelevant

Not all networks are so simple



Complexity of exact inference

- Singly connected networks or polytrees
  - At most one undirected path between any two nodes in the network
  - Time and space complexity in linear in *n*
- Multiply connected networks
  - Time and space complexity is exponential even when the number of parents per nodes is bounded
  - Consider — Special case of Bayesian network inference is inference in propositional logic. Inference is as hard as finding the number of satisfying assignments (#P-hard)
- Thus, may want to consider lower-complexity methods for inference that are approximate

Approximate Inference

- Inference by stochastic simulation
- Simple sampling
- Rejection sampling
- Likelihood weighting
- Markov chain Monte Carlo (MCMC)

Inference terminology

- **Stochastic Process:** a type of non-deterministic process that is the core of approximate inference techniques.

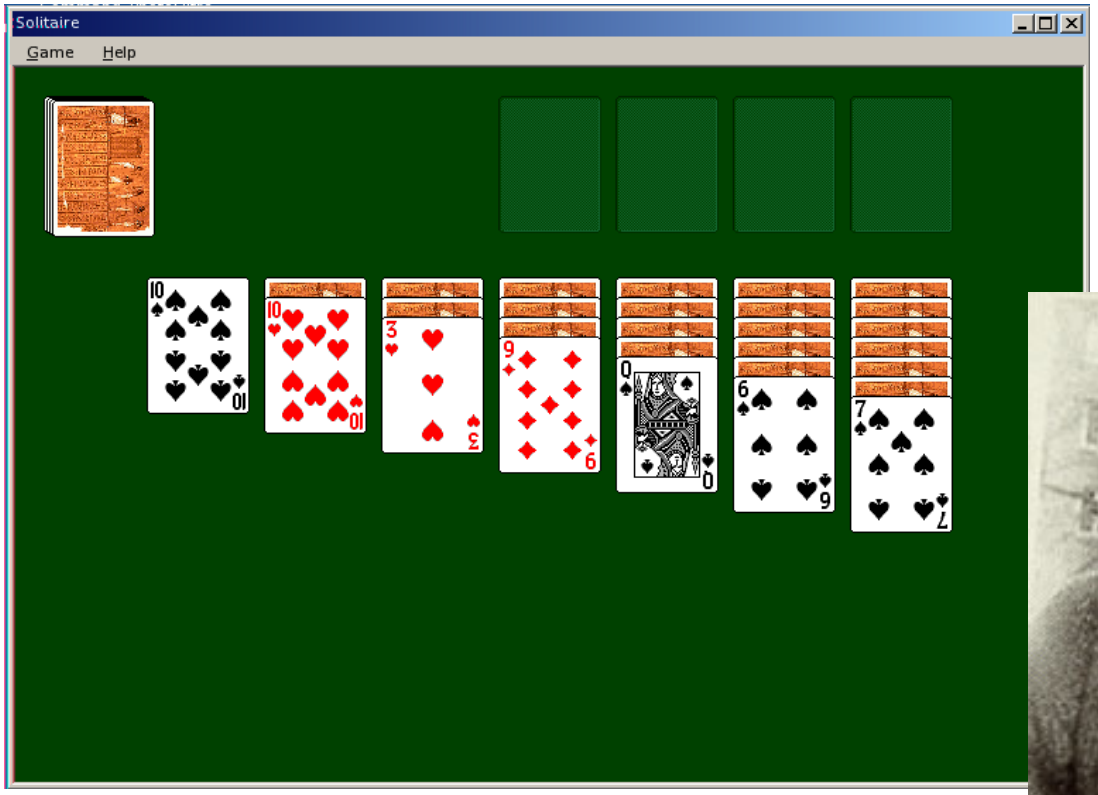
- **Markov Process:** a type of sequential process in which the next state depends only on the prior state

- **Monte Carlo Algorithm:** an algorithm that relies on non-determinism to simulate a system

Why approximate inference?

- Inference in singly connected networks is linear!
- ...but many networks are not singly connected
- Inference in multiply connected networks is exponential, even when the number of parents/node is bounded
- May be willing to trade some small error for more tractable inference

Solitaire and Stanislaw Ulam



Stochastic simulation

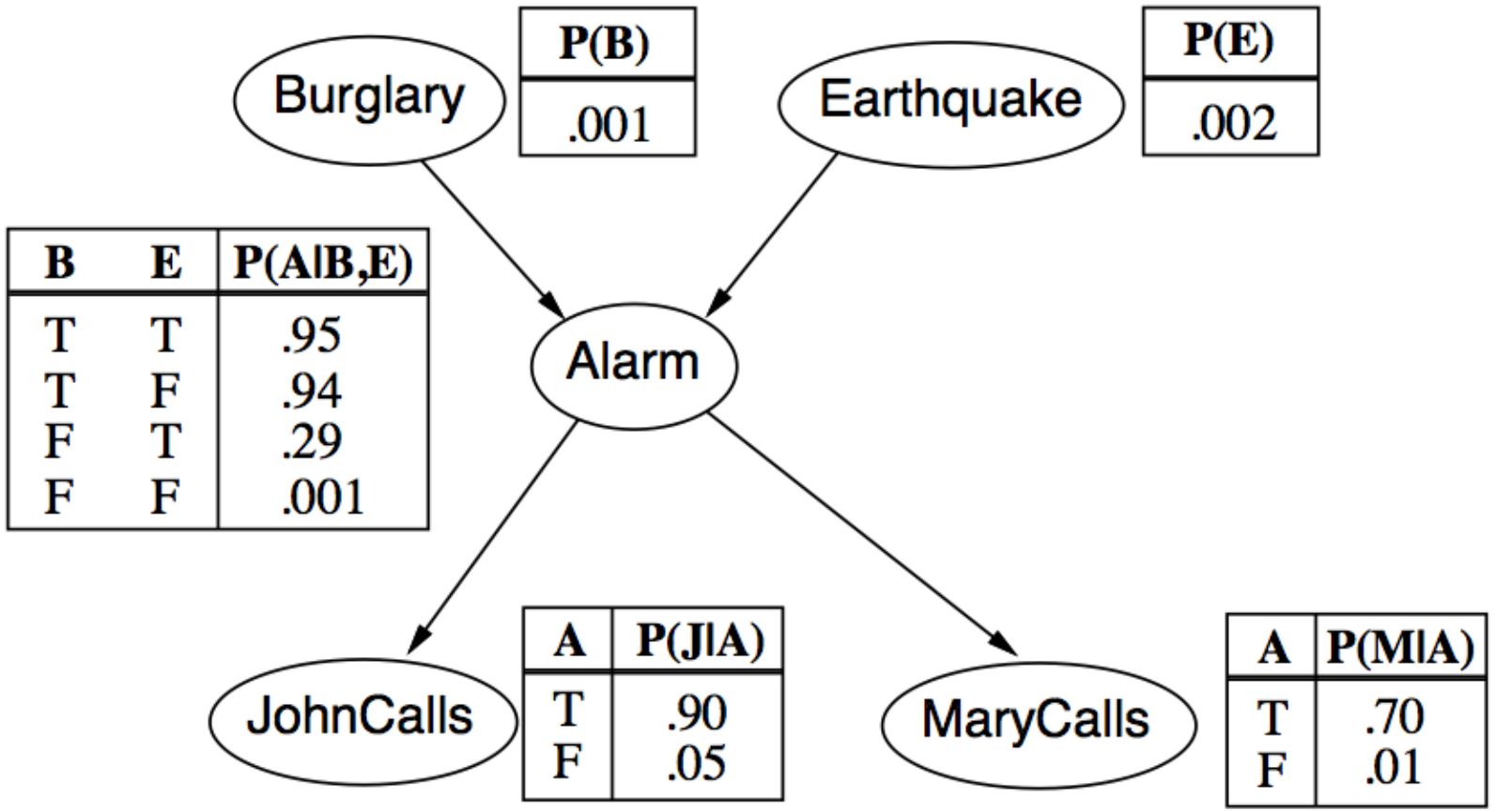
- Core idea
  - Draw samples from a sampling distribution defined by the network
  - Compute an approximate posterior probability in a way that converges to the true probability
- Methods
  - Simple sampling from an empty network
  - Rejection sampling — reject samples that don't agree with the evidence
  - Likelihood weighting — weight samples based on evidence
  - Markov chain Monte Carlo — sample from a stochastic process whose stationary distribution is the true posterior

What are samples?

	T		$\neg T$	
	C	$\neg C$	C	$\neg C$
V	0.108	0.012	0.072	0.008
$\neg V$	0.016	0.064	0.144	0.576

V = Cavity; T = Toothache; C = Catch

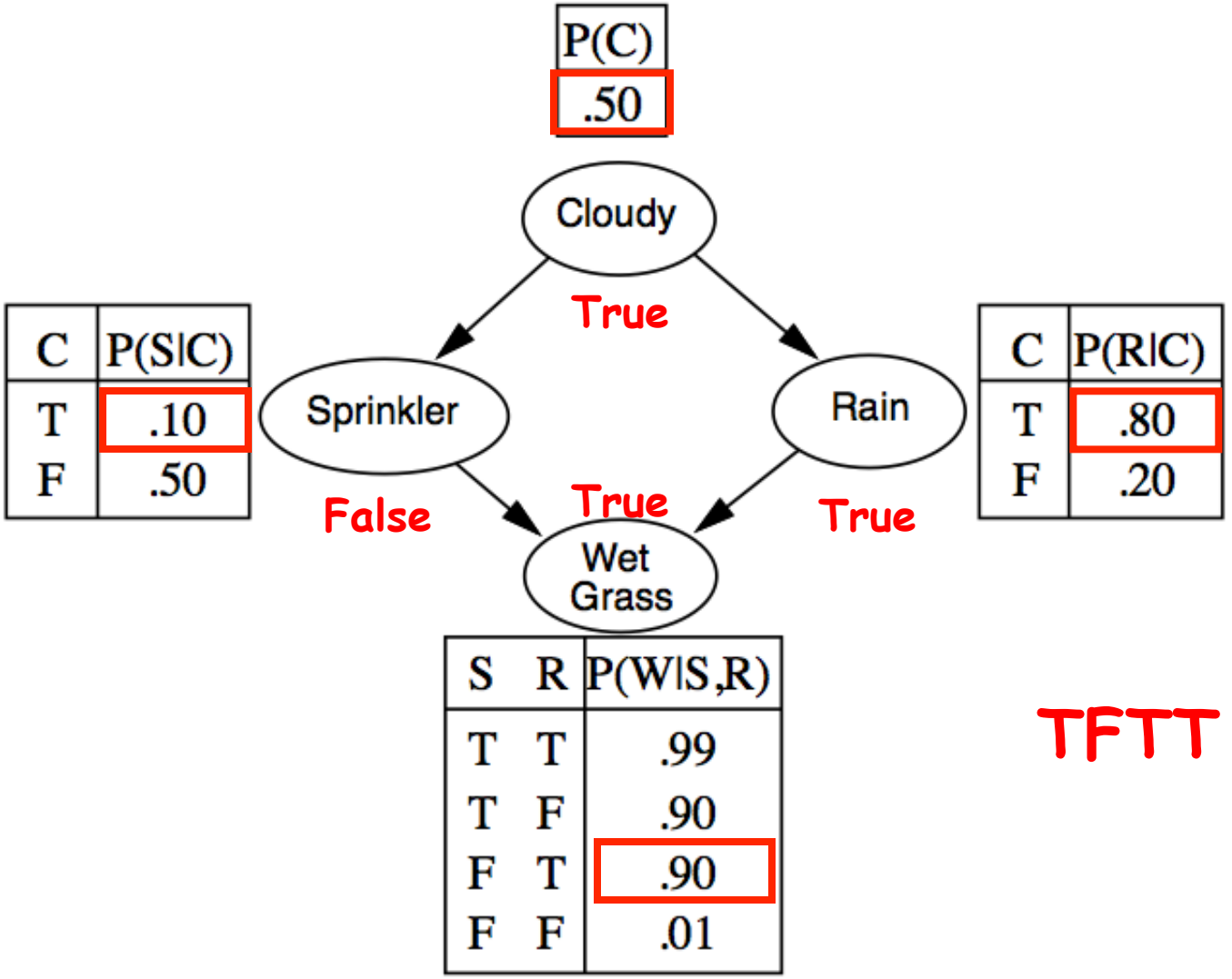
What are samples?



Simple sampling

- Given an empty network...
- And beginning with nodes without parents...
- We can sample from conditional distributions and instantiate all nodes.
- This will produce one element of the joint distribution.
- Doing this many times will produce an empirical distribution that approximates the full joint distribution.

Example


P(C)
.50
Cloudy
True
C P(S|C)
T .10
F .50
Sprinkler
False
Rain
True
C P(R|C)
T .80
F .20
True
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01
TFTT

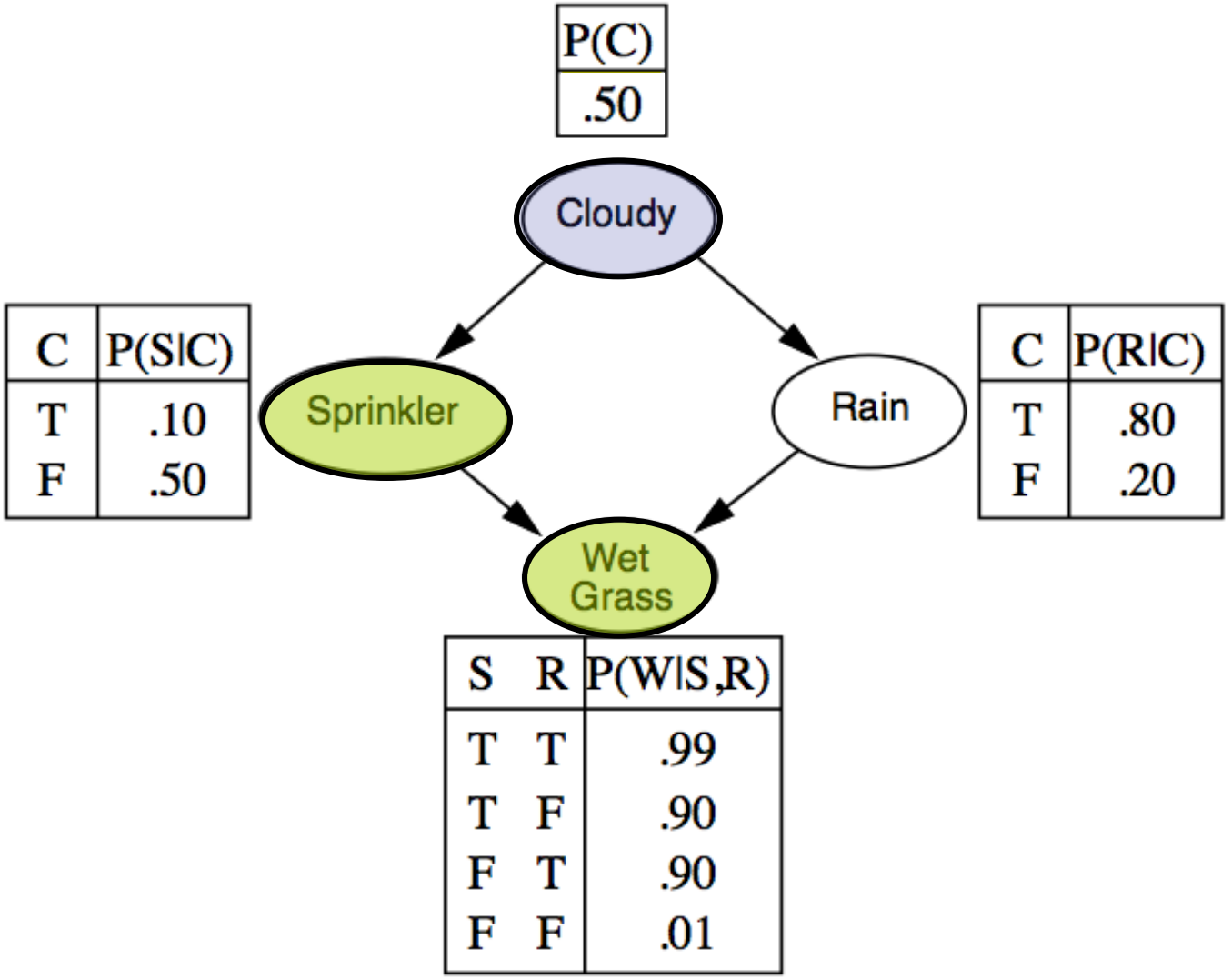
Bayesian networks are *generative*

- BNs can generate samples from the world they represent
- Generating samples is efficient (linear) even though general probabilistic inference is not
- Thus, we will attempt to use the efficient procedure to approximate the inefficient one

Benefits and problems of simple sampling

- Works well for an empty network
  - Simple
  - In the limit (many samples), the estimated distribution approaches the true posterior
- But in nearly all cases, we have evidence, rather than an empty network
- What can we do?
- Throw out cases that don't match the evidence

Example


P(C)
.50
Cloudy
C P(S|C)
T .10
F .50
Sprinkler
Rain
C P(R|C)
T .80
F .20
Wet Grass
S R P(W|S,R)
T T .99
T F .90
F T .90
F F .01

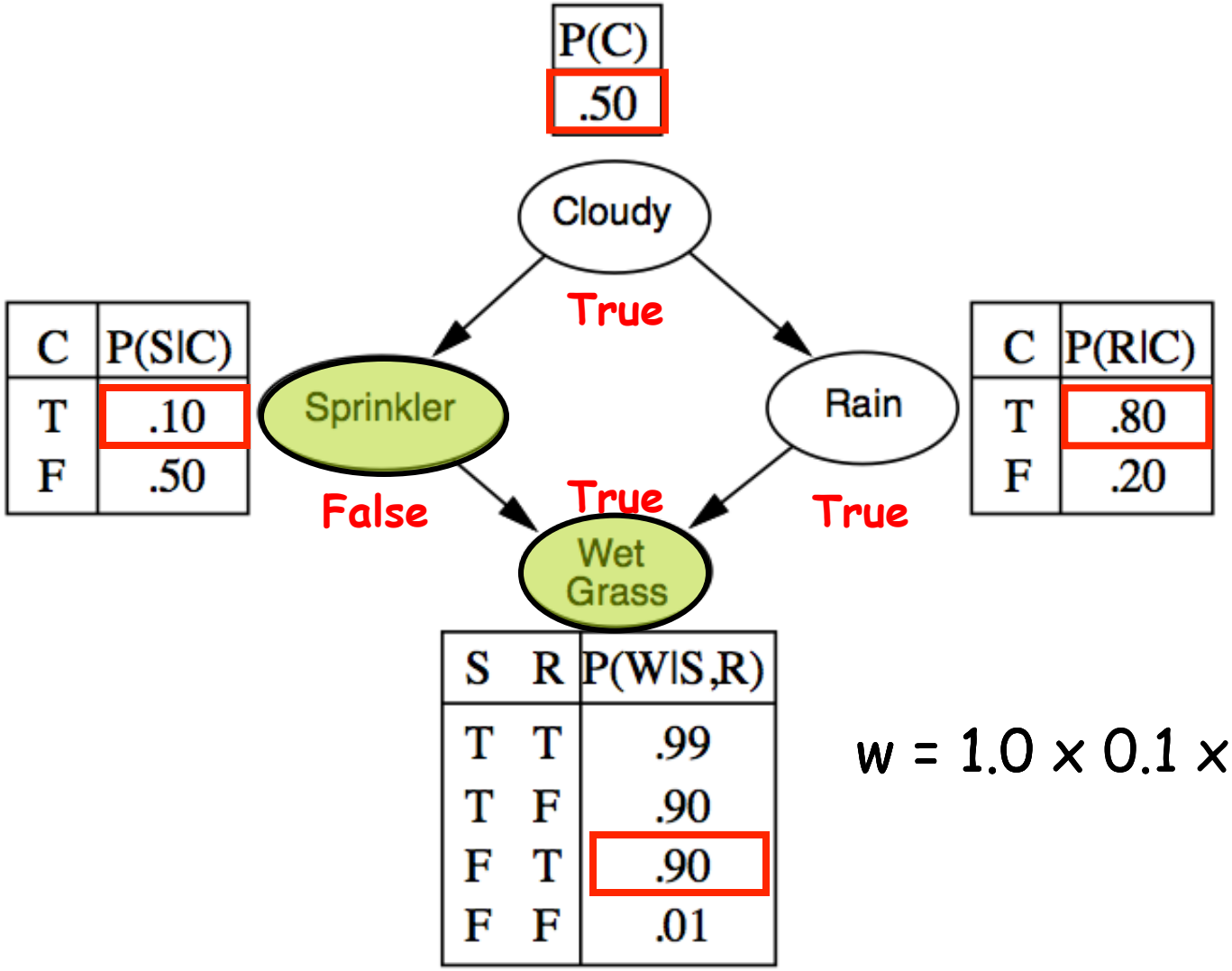
Rejection sampling

- Sample the network as before...
- But discard samples that don't correspond with the evidence.
- Similar to real-world estimation procedures, but the network is the stand-in for the world (much cheaper and easier).
- However, hopelessly expensive for large networks where P(e) is small.

Likelihood weighting

- Do simple sampling as before...
- But weight the likelihood of each sample based on the evidence

Evidence



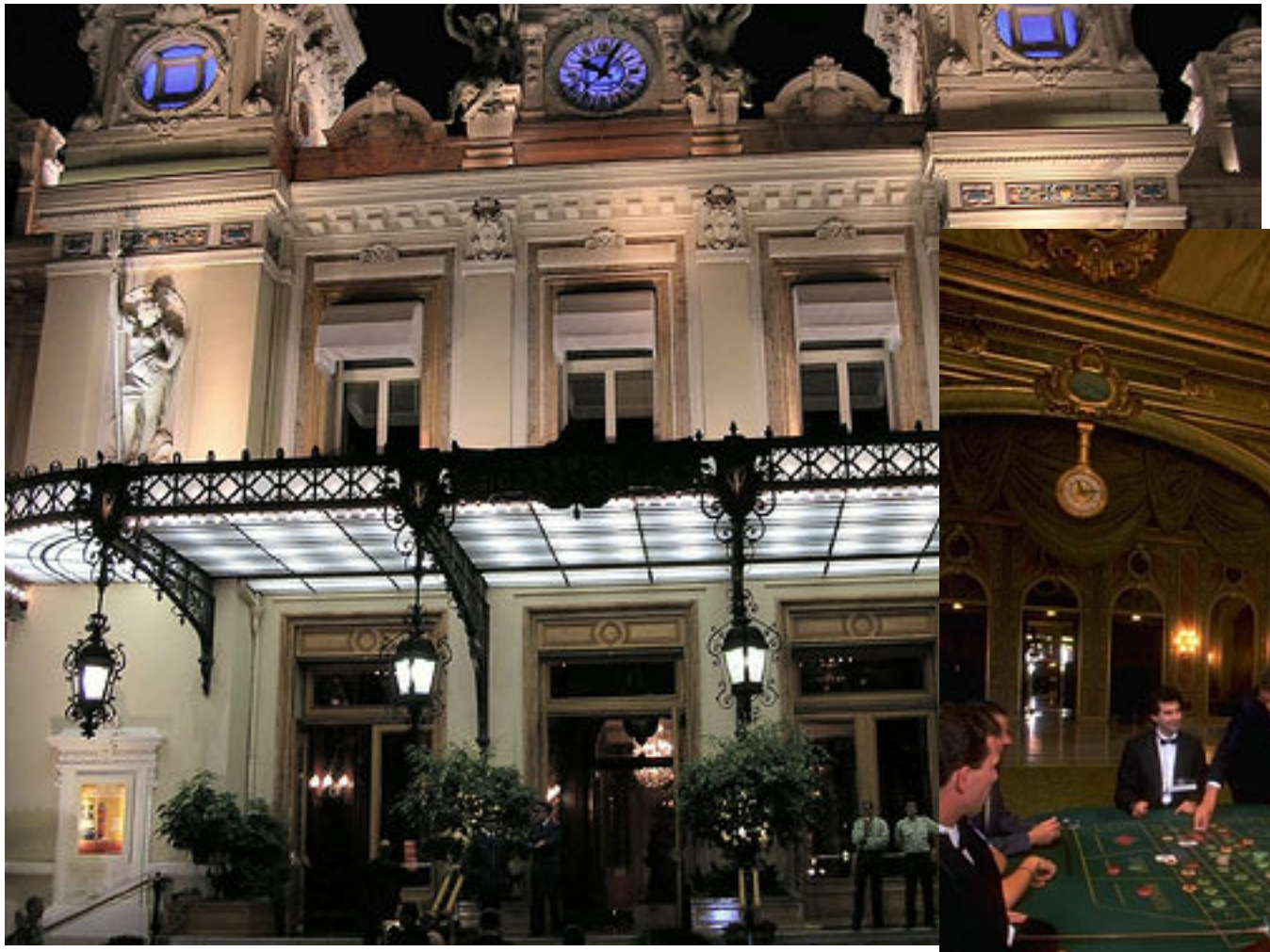
$w = 1.0 \times 0.1 \times 0.9 = 0.09$

Likelihood weighting

- Do simple sampling as before...
- But weight the likelihood of each sample based on the evidence
  - Inferred values only pay attention to the evidence in *ancestors*, not children, thus producing estimates somewhere in between the prior and the posterior
  - The weighting makes up the difference
- Problems
  - Performance degrades with many evidence variables

Markov chain Monte Carlo

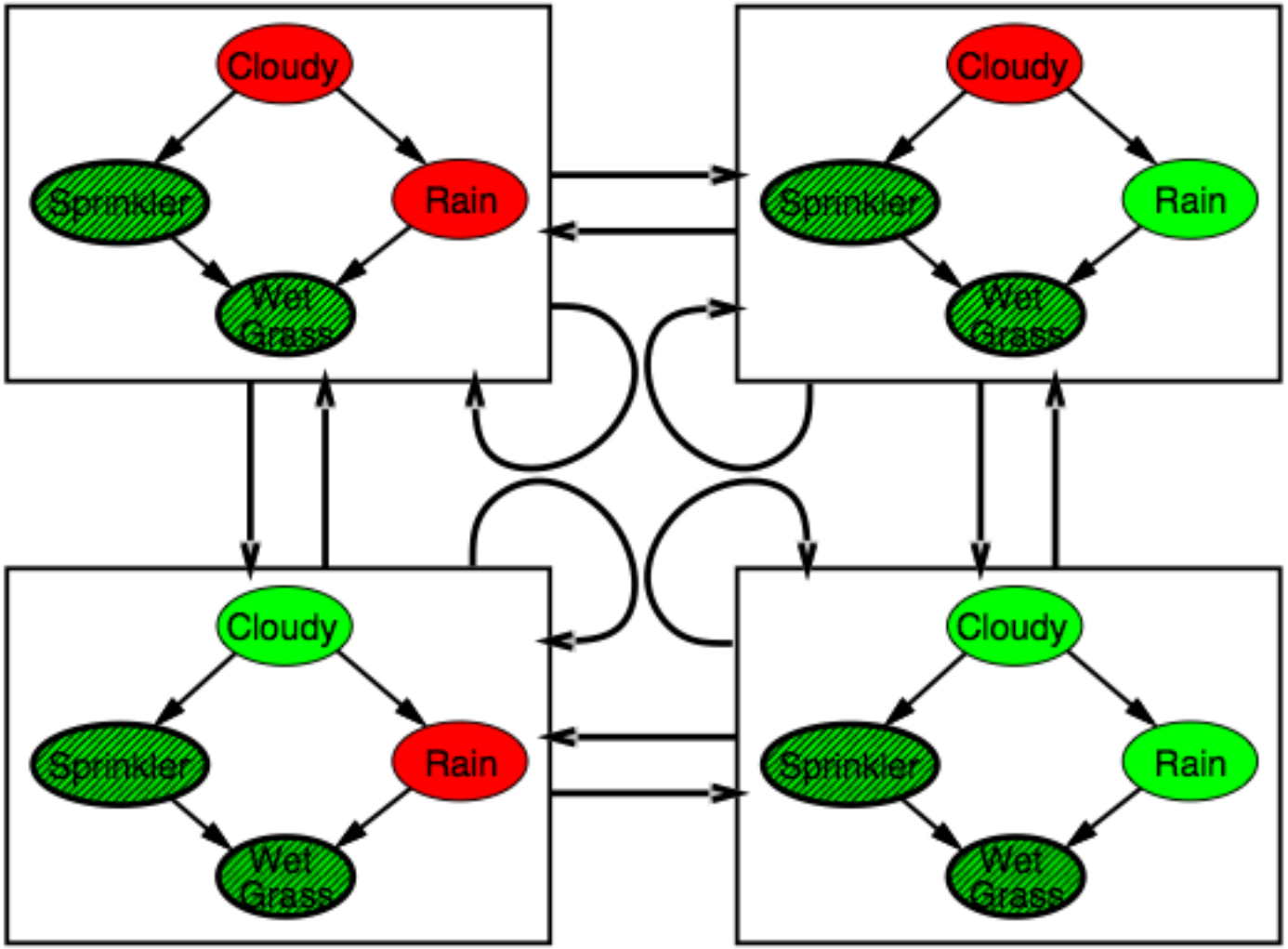
- Markov chain
  - Description of the state of a system at successive times
  - Markov property — State at time *t+1* depends only on the state at time *t*, not time *t-i* for *i>0*
- Monte Carlo–
  - A class of non-deterministic algorithms used to simulate the behavior of a system



MCMC

- The “state” of the system is the current assignment of all variables
- Algorithm
  - Initialize all variables randomly
  - Generate next state by sampling one variable given its Markov blanket
  - Sample each variable in turn, keeping other evidence fixed.
- Variable selection can be sequential or random

Markov chain



Probability given Markov blanket

$$P(x_i'|mb(X_i)) = P(x_i'|parents(X_i))\prod_{Z_j \in Children(X_i)} P(z_j|parents(Z_j))$$

MCMC Problems

- Difficult to tell if it has converged
- Multiple parameters (e.g., burn-in period)
- Can be wasteful if the Markov blanket is large because probabilities don’t change much

Today's topics: exact and approximate inference

- Exact
  - Inference with joint probability distributions
  - Exact inference in Bayesian networks
  - Inference by enumeration
  - Complexity of exact inference
- Approximate
  - Inference by stochastic simulation
  - Simple sampling
  - Rejection sampling
  - Markov chain Monte Carlo (MCMC)

Next Class

- Making simple decisions: Utility theory
- Secs. 16.1 – 16.4