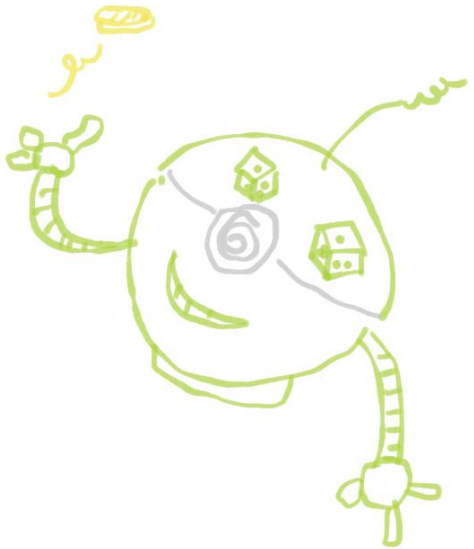
CS5491: Artificial Intelligence

Belief Networks



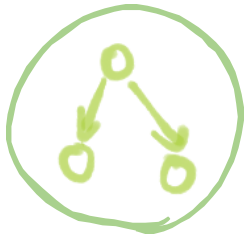
Instructor: Kai Wang

Recap: More Logics

Logic	Primitives	Available knowledge
propositional	facts	true/false/unknown
first-order	facts, objects, relations	true/false/unknown
temporal	facts, objects, relations, times	true/false/unknown
probabilistic theory	facts	degree of belief 0,…,1
fuzzy logic	facts + degree of truth	known internal value

Today

Probabilistic Reasoning

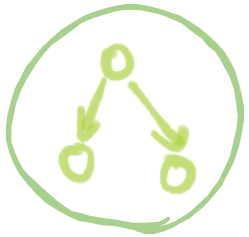


Bayesian Networks

Inference

Today

Probabilistic Reasoning



Bayesian Networks

Inference

Uncertainty

- Let action A_t = leave for airport t minutes before flight
  Will A_t get me there on time?

- Problems
  - partial observability (road state, other drivers' plans, etc.)
  - noisy sensors (WBAL traffic reports)
  - uncertainty in action outcomes (flat tire, etc.)
  - immense complexity of modelling and predicting traffic

- Hence a purely logical approach either
  - risks falsehood: "A_{25} will get me there on time"
  - leads to conclusions that are too weak for decision making:
    "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact etc etc."

Methods for Handling Uncertainty

- Default or nonmonotonic logic:
  - Assume my car does not have a flat tire
  - Assume A_{25} works unless contradicted by evidence

  Issues: What assumptions are reasonable? How to handle contradiction?

- Rules with fudge factors:

  $Sprinkler \mapsto_{0.99} WetGrass$

  $WetGrass \mapsto_{0.7} Rain$

  Issues: Problems with combination, e.g., *Sprinkler* causes *Rain*?

- Probability

  Given the available evidence,

  A_{25} will get me there on time with probability 0.04

  Mahaviracarya (9th C.), Cardamo (1565) theory of gambling

- Fuzzy logic handles **degree of truth** NOT uncertainty, e.g., *WetGrass* is true to degree 0.2)

Probability

- Probabilistic assertions **summarize** effects of
  - laziness: failure to enumerate exceptions, qualifications, etc.
  - ignorance: lack of relevant facts, initial conditions, etc.

- Subjective or Bayesian probability:
  Probabilities relate propositions to one's own state of knowledge
  - e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$

- Might be learned from past experience of similar situations
  - e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$

- Analogous to logical entailment status, not truth

Making Decisions under Uncertainty

- Suppose I believe the following:

$$P(A_{25} \text{ gets me there on time} \ldots) = 0.04$$
$$P(A_{90} \text{ gets me there on time} \ldots) = 0.70$$
$$P(A_{120} \text{ gets me there on time} \ldots) = 0.95$$
$$P(A_{1440} \text{ gets me there on time} \ldots) = 0.9999$$

- Which action to choose?
- Depends on my preferences for missing flight vs. airport cuisine, etc.
- Utility theory is used to represent and infer preferences
- Decision theory = utility theory + probability theory

Prior Probability

- Prior or unconditional probabilities of propositions
  e.g., *P*(*Cavity=true*) =0.2 and *P*(*Weather=sunny*) =0.72
  correspond to belief prior to arrival of any (new) evidence

- Probability distribution gives values for all possible assignments:
  $\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

- Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
  P(*Weather, Cavity*) = a 4×2 matrix of values:

Weather=	*sunny*	*rain*	*cloudy*	*snow*
Cavity=true	0.144	0.02	0.016	0.02
Cavity=false	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Conditional Probability

- Conditional or posterior probabilities
  e.g., $P(cavity|toothache) = 0.8$
  i.e., **given that** $toothache$ **is all I know**
  NOT "if $toothache$ then 80% chance of $cavity$"

- (Notation for conditional distributions:
  P$(Cavity|Toothache)$ = 2-element vector of 2-element vectors)

- If we know more, e.g., *cavity* is also given, then we have
  $P(cavity|toothache, cavity) = 1$
  Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

- New evidence may be irrelevant, allowing simplification, e.g.,
  $P(cavity|toothache, RavensWin) = P(cavity|toothache) = 0.8$
  This kind of inference, sanctioned by domain knowledge, is crucial

Conditional Probability

- Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

- Product rule gives an alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

- A general version holds for whole distributions, e.g.,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, **not** matrix multiplication)

- Chain rule is derived by successive application of product rule:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})\end{aligned}$$

Inference by Enumeration

- Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \sum_{\omega : \omega \models \varphi} P(\omega)$$

(catch = dentist's steel probe gets caught in cavity)

Inference by Enumeration

- Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \sum_{\omega : \omega \models \varphi} P(\omega)$$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by Enumeration

- Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- For any proposition φ, sum the atomic events where it is true:

$$P(\varphi) = \sum_{\omega : \omega \models \varphi} P(\omega)$$

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by Enumeration

- Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Can also compute conditional probabilities:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Denominator can be viewed as a normalization constant

$$\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache)$$
$$= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$
$$= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

- General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Inference by Enumeration

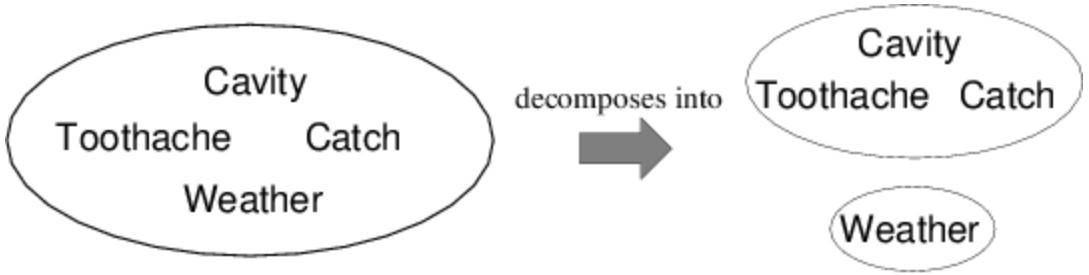
- Let **X** be all the variables.
  Typically, we want the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**
- Let the hidden variables be **H** = **X** – **Y** – **E**
- Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\sum_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

- The terms in the summation are joint entries because **Y**, **E**, and **H** together exhaust the set of random variables
- Obvious problems
  - Worst-case time complexity $O(d^n)$ where d is the largest arity
  - Space complexity $O(d^n)$ to store the joint distribution
  - How to find the numbers for $O(d^n)$ entries???

Independence

- A and B are independent iff
  $\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$



- $\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$
- 32 entries reduced to 12; for n independent biased coins, $2^n \rightarrow n$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Conditional Independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) $P(catch|toothache, cavity) = P(catch|cavity)$

- The same independence holds if I haven't got a cavity:
  (2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

- *Catch* is conditionally independent of *Toothache* given *Cavity*:

- Equivalent statements:
  $\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$
  $\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Conditional Independence

- Write out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(Toothache, Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity) \\
&= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity) \\
&= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)
\end{aligned}$$

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

- **Conditional independence is our most basic and robust form of knowledge about uncertain environments.**

Bayes Rule

- Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\Longrightarrow \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

- Or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Bayes Rule

- Useful for assessing diagnostic probability from causal probability

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

- E.g., let *M* be meningitis, *S* be stiff neck:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

- Note: posterior probability of meningitis still very small!

Bayes' Rule and Conditional Independence

- Example of a naive Bayes model

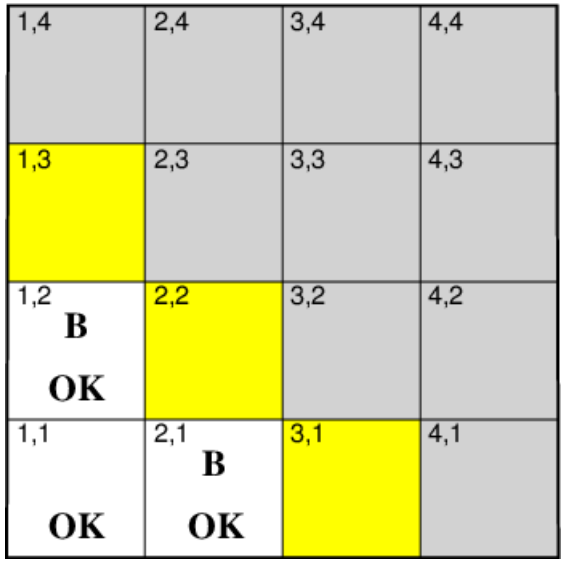
$$\mathbf{P}(Cavity|toothache \wedge catch)$$
$$= \alpha \mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)$$
$$= \alpha \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$$

- Generally:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$

- Total number of parameters is **linear** in n

Wumpus World



- P_{ij} = *true* iff $[i, j]$ contains a pit
- B_{ij} = *true* iff $[i, j]$ is breezy

  Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the Probability Model

- The full joint distribution is $\mathbf{P}(P_{1,1},\ldots,P_{4,4},B_{1,1},B_{1,2},B_{2,1})$
- Apply product rule: $\mathbf{P}(B_{1,1},B_{1,2},B_{2,1}\mid P_{1,1},\ldots,P_{4,4})\mathbf{P}(P_{1,1},\ldots,P_{4,4})$

  This gives us: $P(Effect|Cause)$

- First term: 1 if pits are adjacent to breezes, 0 otherwise
- Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1},\ldots,P_{4,4}) = \prod_{i,j=1,1}^{4,4}\mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and Query

- We know the following facts:

  $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$

  $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

- Query is $\mathbf{P}(P_{1,3}|known, b)$

- Define *Unknown* = P_{ij}s other than $P_{1,3}$ and *Known*

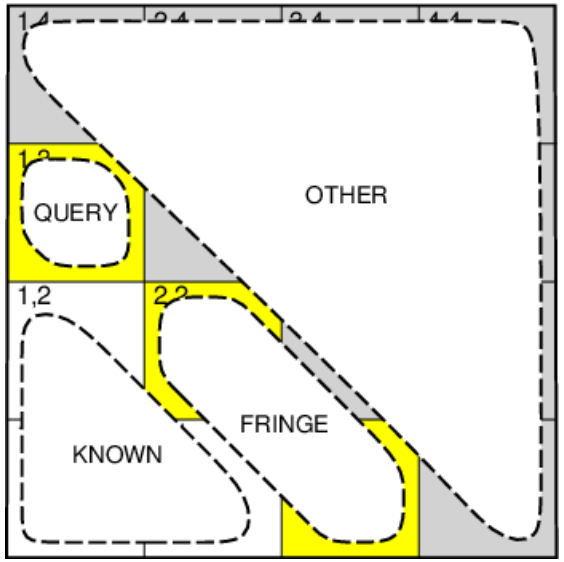
- For inference by enumeration, we have

$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

- Grows exponentially with number of squares!

Using Conditional Independence

- Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares



- Define $Unknown = Fringe \cup Other$
  $\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$
- Manipulate query into a form where we can use this!

Using Conditional Independence

$$\mathbf{P}(P_{1,3}|known,b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

$$= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown)$$

$$= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other)$$

$$= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other)$$

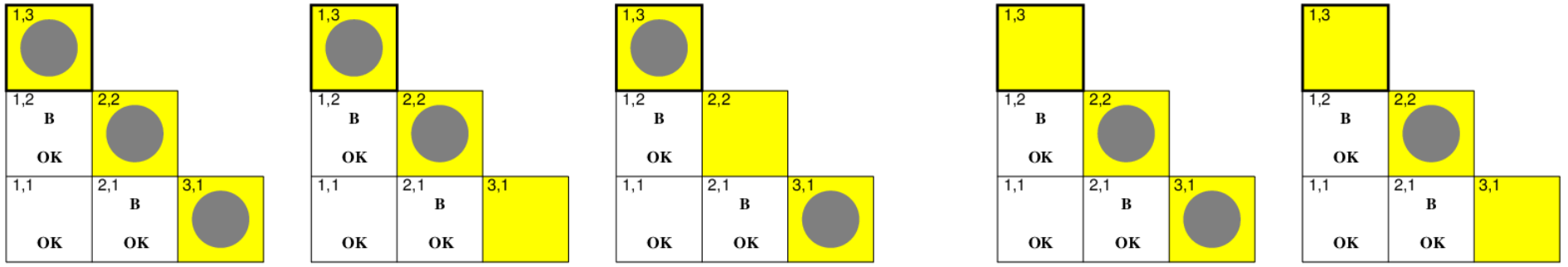
$$= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other)$$

$$= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other)$$

$$= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other)$$

$$= \alpha' \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)$$

Using Conditional Independence

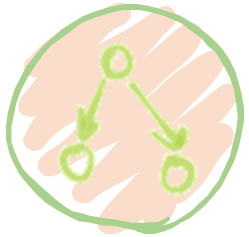


$$\mathbf{P}(P_{1,3}|known, b) = \alpha \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle$$
$$\approx \langle 0.31, 0.69 \rangle$$

$$\mathbf{P}(P_{2,2}|known, b) \approx \langle 0.86, 0.14 \rangle$$

Today

Probabilistic Reasoning



Bayesian Networks

Inference

Bayesian Networks

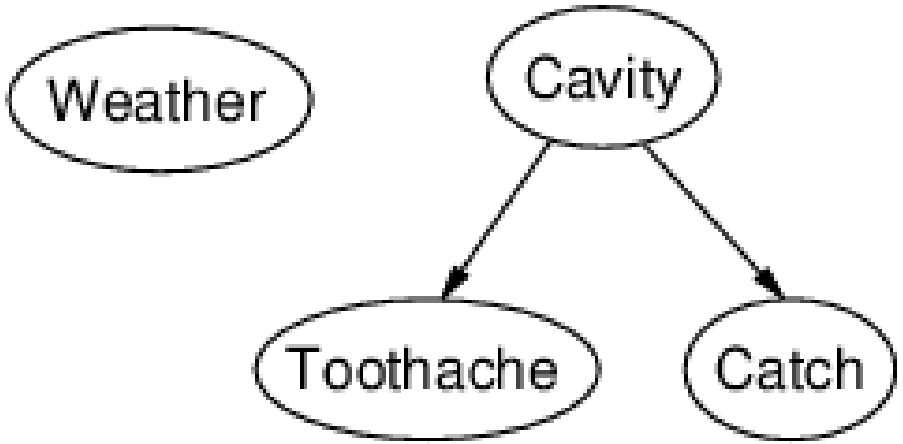
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈"directly influences")
  - a conditional distribution for each node given its parents:
    $\mathbf{P}(X_i|Parents(X_i))$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

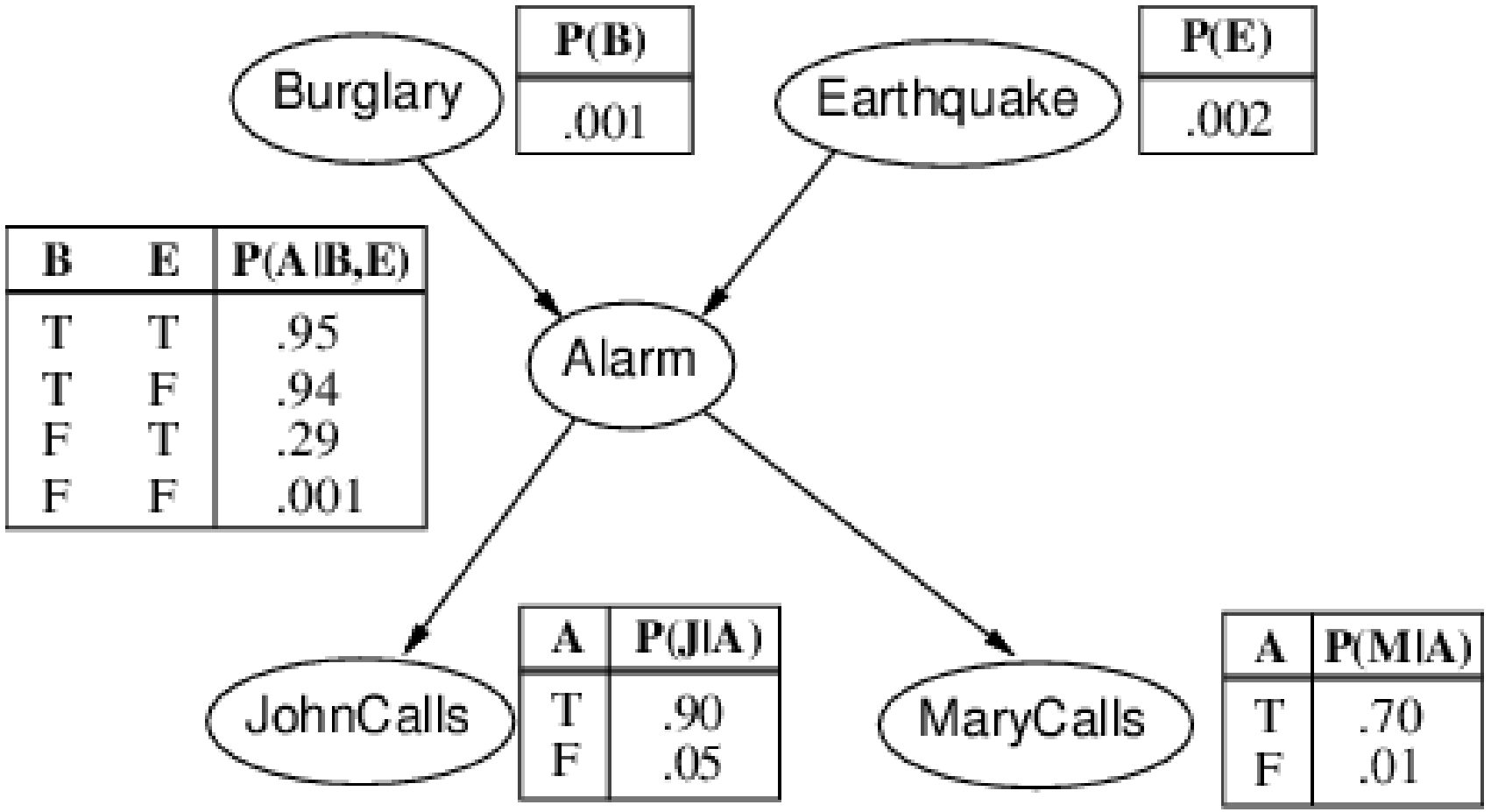
- *I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes.*

  Is there a burglar?

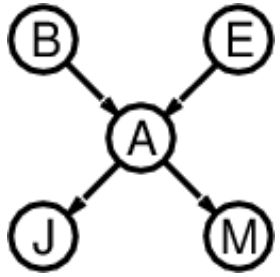
- Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- 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

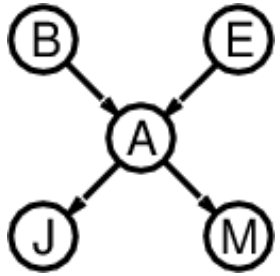

Burglary
P(B)
.001
Earthquake
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
Alarm
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

Compactness



- A conditional probability table for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values
- Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1 - p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 = 32$)

Global Semantics



- Global semantics defines the full joint distribution as the product of the local conditional distributions:

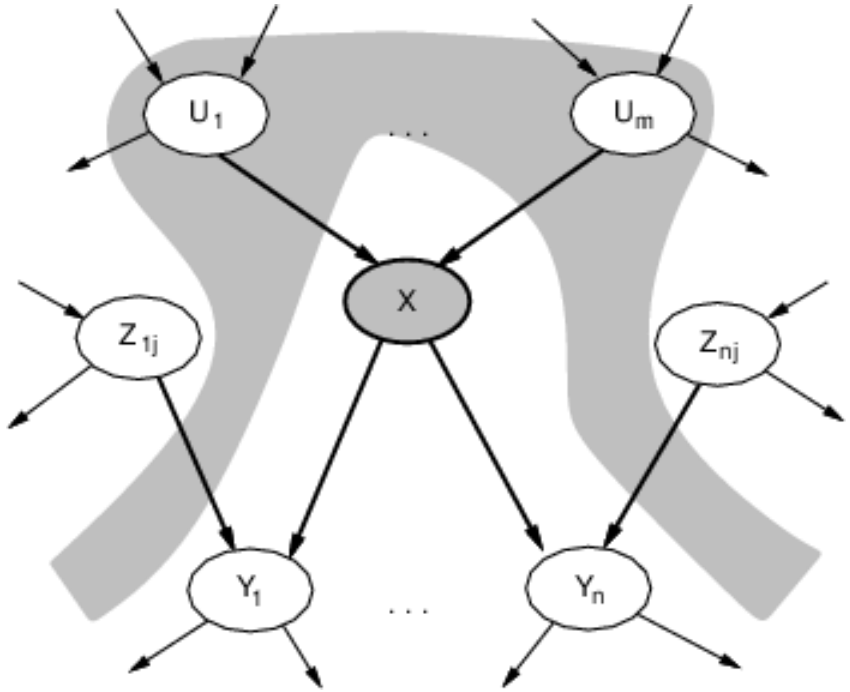
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- E.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

Local Semantics

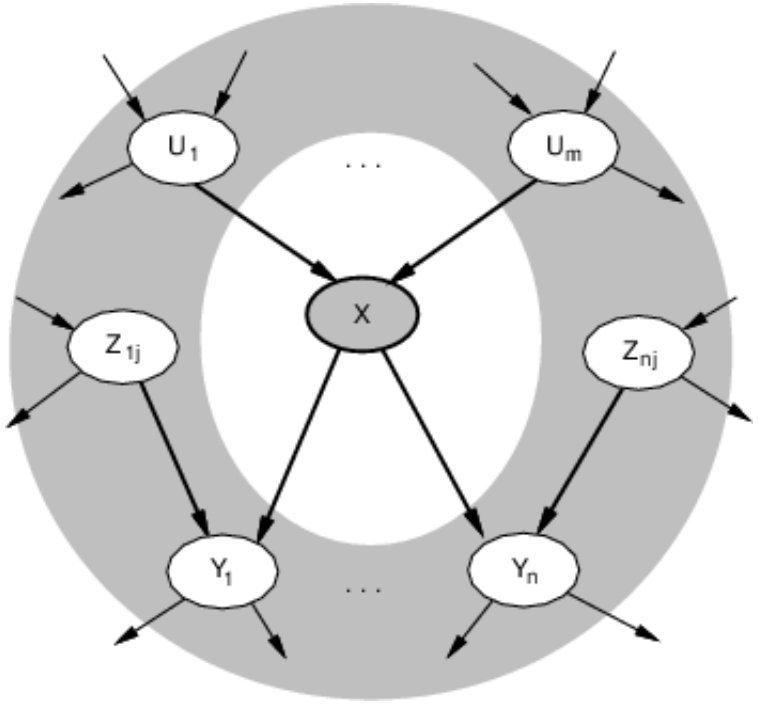
- Local semantics: each node is conditionally independent of its nondescendants given its parents



- Theorem: Local semantics ⇔ global semantics

Markov Blanket

- Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Constructing Bayesian Networks

- Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For i = 1 to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i|Parents(X_i)) = \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

- This choice of parents guarantees the global semantics:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}$$

$$= \prod_{i=1}^{n} \mathbf{P}(X_i|Parents(X_i)) \quad \text{(by construction)}$$

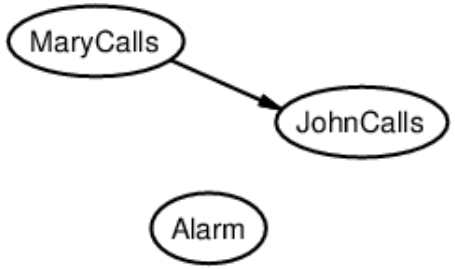
Example

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

- $P(J|M) = P(J)$?

Example

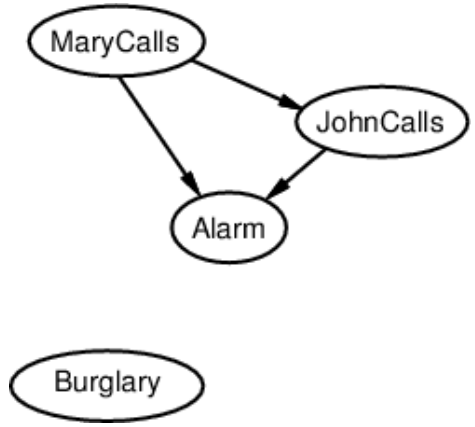
- Suppose we choose the ordering *M*, *J* , *A*, *B*, *E*



- $P(J|M) = P(J)$? No
- $P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$?

Example

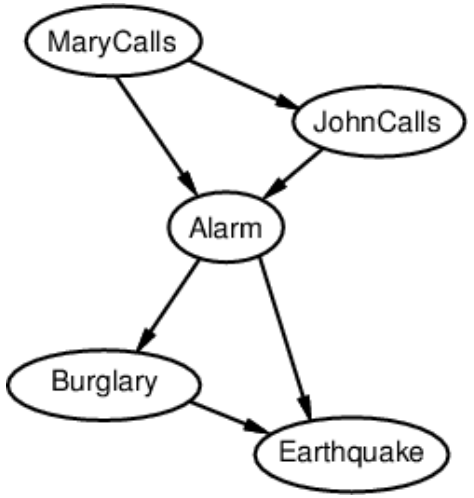
- Suppose we choose the ordering *M*, *J* , *A*, *B*, *E*



- $P(J|M) = P(J)$? 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

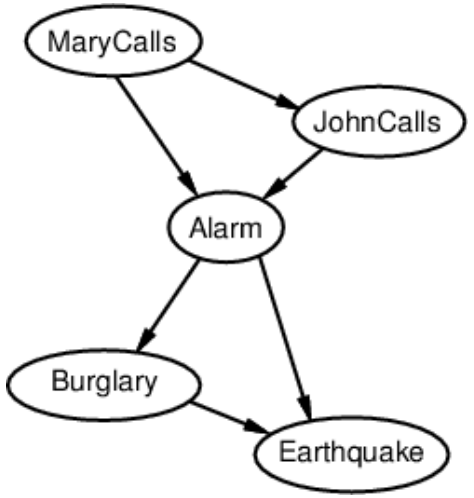
- Suppose we choose the ordering *M*, *J* , *A*, *B*, *E*



- $P(J|M) = P(J)$? 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
- $P(E|B,A,J,M) = P(E|A)$?
- $P(E|B,A,J,M) = P(E|A,B)$?

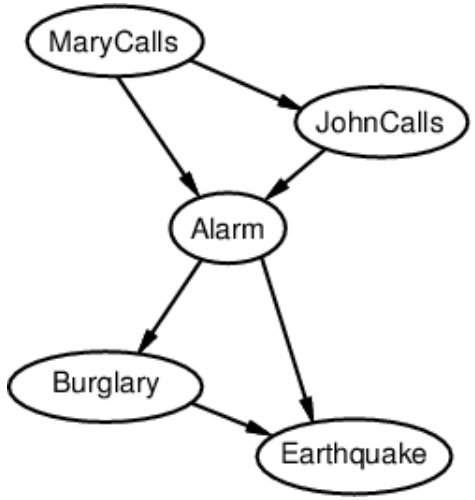
Example

- 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

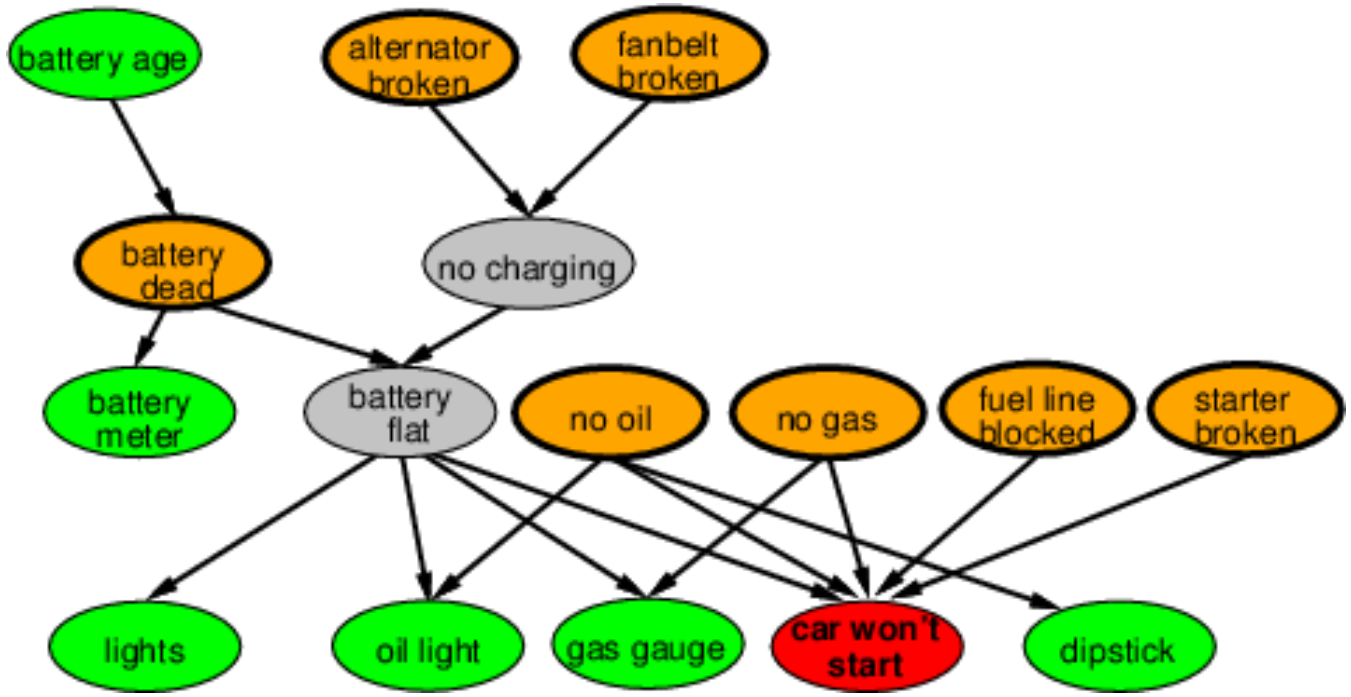
Example



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

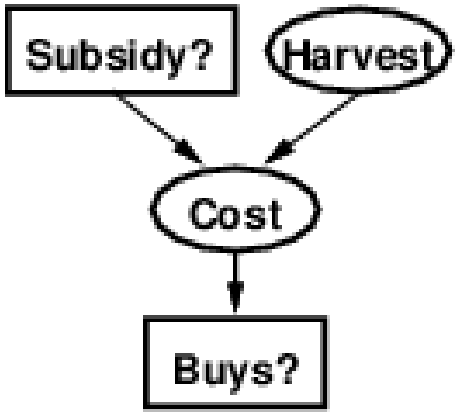
Example: Car Diagnosis

- Initial evidence: car won't start
- Testable variables (green), "broken, so fix it" variables (orange)
- Hidden variables (gray) ensure sparse structure, reduce parameters



Hybrid (Discrete+Continuous) 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?*)

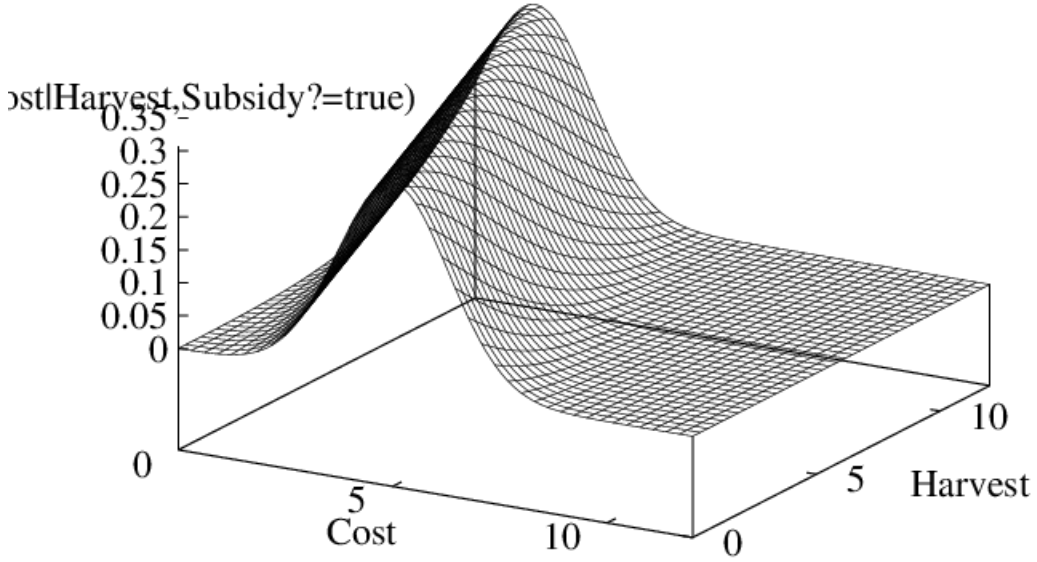
Continuous Child Variables

- 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.,:

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

Continuous Child Variables

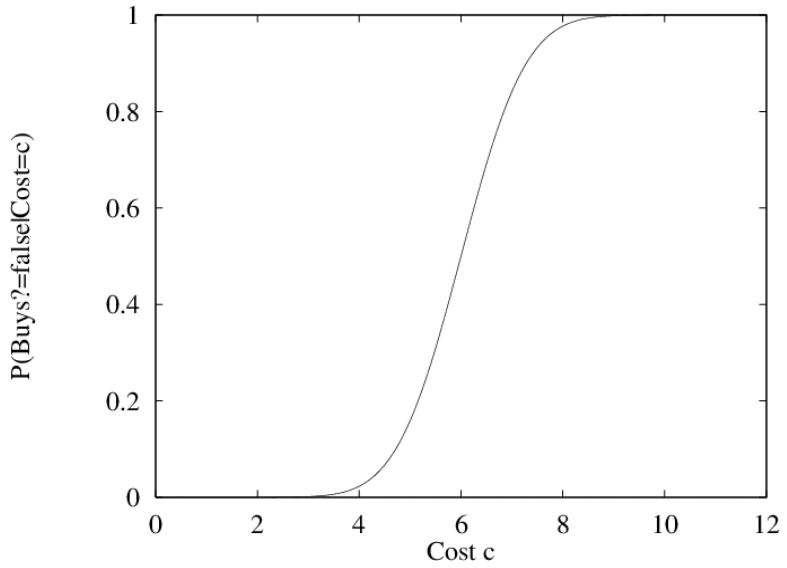


- All-continuous network with LG distributions
  ⟹ full joint distribution is a multivariate Gaussian

- Discrete+continuous LG network is a conditional Gaussian network i.e., a multivariate Gaussian over all continuous variables for each combination of discrete variable values

Discrete Variable w/ Continuous Parents

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



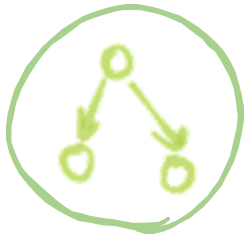
- Probit distribution uses integral of Gaussian:

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

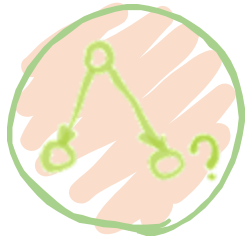
$$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$$

Today

Probabilistic Reasoning



Bayesian Networks



Inference

Inference Tasks

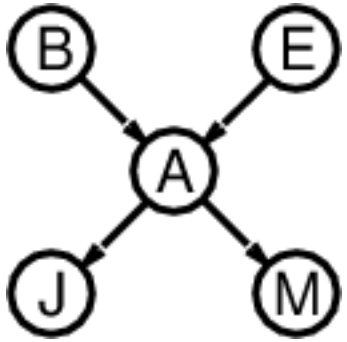
- Simple queries: compute posterior marginal $\mathbf{P}(X_i|\mathbf{E}=\mathbf{e})$
  e.g., $P(NoGas|Gauge=empty, Lights=on, Starts=false)$
- Conjunctive queries: $\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$
- Optimal decisions: decision networks include utility information;
  probabilistic inference required for $P(outcome|action, evidence)$
- Value of information: which evidence to seek next?
- Sensitivity analysis: which probability values are most critical?
- Explanation: why do I need a new starter motor?

Inference by Enumeration

- Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

- Simple query on the burglary network

$$\mathbf{P}(B|j,m)$$
$$= \mathbf{P}(B,j,m)/P(j,m)$$
$$= \alpha\mathbf{P}(B,j,m)$$
$$= \alpha \sum_e \sum_a \mathbf{P}(B,e,a,j,m)$$



- Rewrite full joint entries using product of CPT entries:

$$\mathbf{P}(B|j,m)$$
$$= \alpha \sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$$
$$= \alpha\mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)$$

Enumeration Algorithm

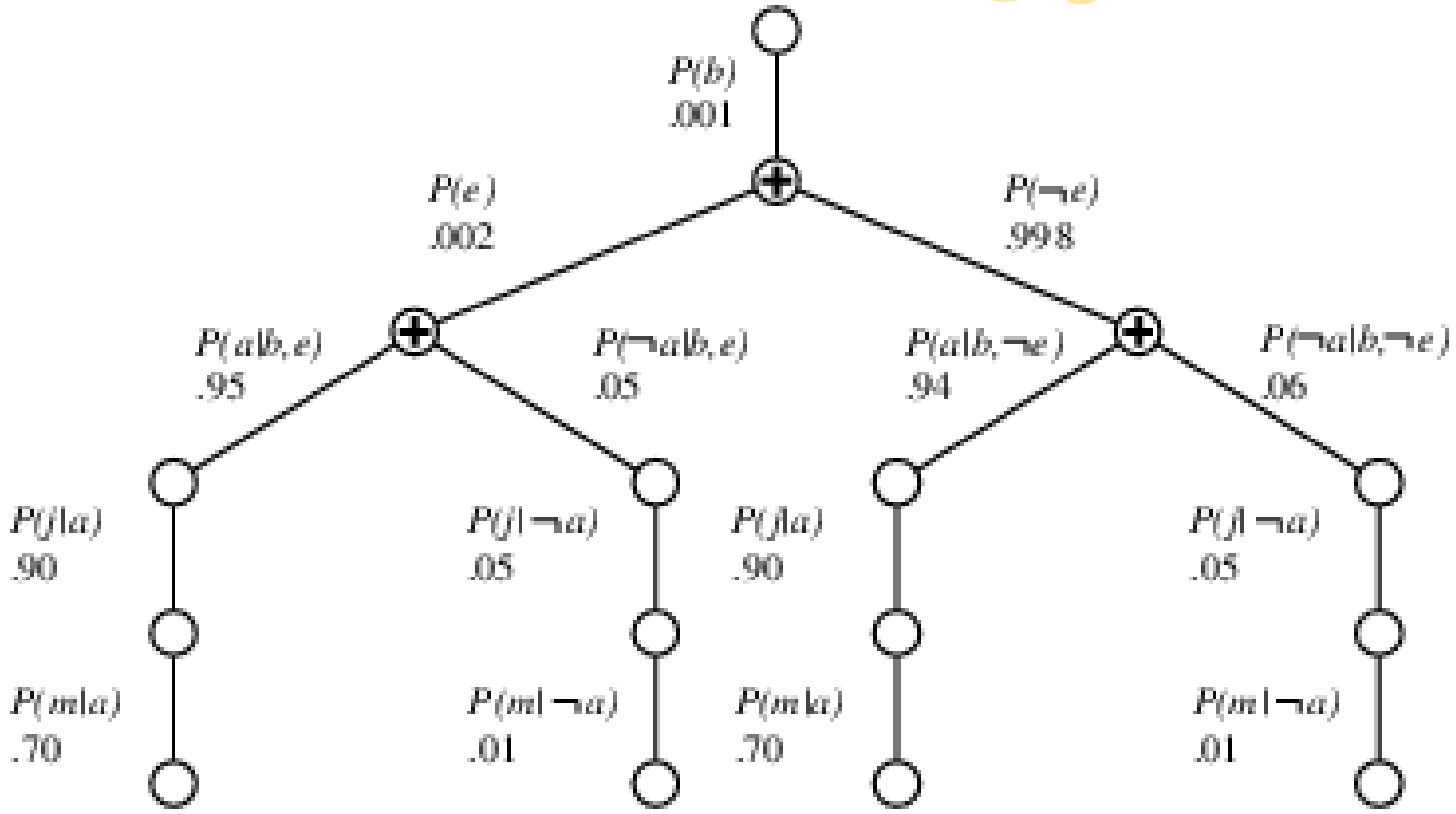
function ENUMERATION-ASK(*X*, **e**, *bn*) **returns** a distribution over *X*
  inputs: *X*, the query variable
    e, observed values for variables **E**
    bn, a Bayesian network with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$

  $\mathbf{Q}(X) \leftarrow$ a distribution over *X*, initially empty
  for each value x_i of *X* **do**
    extend **e** with value x_i for *X*
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(VARS[*bn*], **e**)
  return NORMALIZE($\mathbf{Q}(X)$)

---

function ENUMERATE-ALL(*vars*, **e**) **returns** a real number
  if EMPTY?(*vars*) **then return** 1.0
  $Y \leftarrow$ FIRST(*vars*)
  if *Y* has value *y* in **e**
    then return $P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST(*vars*), **e**)
    else return $\sum_y P(y \mid Pa(Y)) \times$ ENUMERATE-ALL(REST(*vars*), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is **e** extended with $Y = y$

Enumeration Tree



✦ Enumeration is inefficient: repeated computation

e.g., computes $P(j|a)P(m|a)$ for each value of e

Inference by Variable Elimination

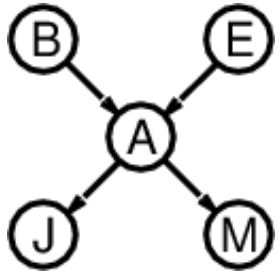
- Variable elimination: carry out summations right-to-left, storing intermediate results (factors) to avoid recomputation

$$\mathbf{P}(B|j,m)$$
$$= \alpha \underbrace{\mathbf{P}(B)}_{B} \sum_e \underbrace{P(e)}_{E} \sum_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M}$$
$$= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) P(j|a) f_M(a)$$
$$= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e) f_J(a) f_M(a)$$
$$= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a f_A(a,b,e) f_J(a) f_M(a)$$
$$= \alpha \mathbf{P}(B) \sum_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)}$$
$$= \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)}$$
$$= \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)$$

Variable Elimination Algorithm

function ELIMINATION-ASK(*X*, **e**, *bn*) **returns** a distribution over *X*
  inputs: *X*, the query variable
    e, evidence specified as an event
    bn, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$
  factors ← [ ]; *vars* ← REVERSE(VARS[*bn*])
  for each *var* **in** *vars* **do**
    factors ← [MAKE-FACTOR(var, **e**)|$factors$]
    if *var* is a hidden variable **then** *factors* ← SUM-OUT(*var*, *factors*)
  return NORMALIZE(POINTWISE-PRODUCT(*factors*))

Irrelevant Variables



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

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

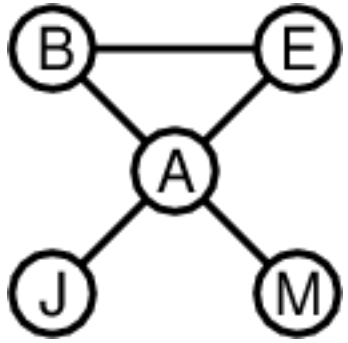
Sum over m is identically 1; M is **irrelevant** to the query

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

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

Irrelevant Variables

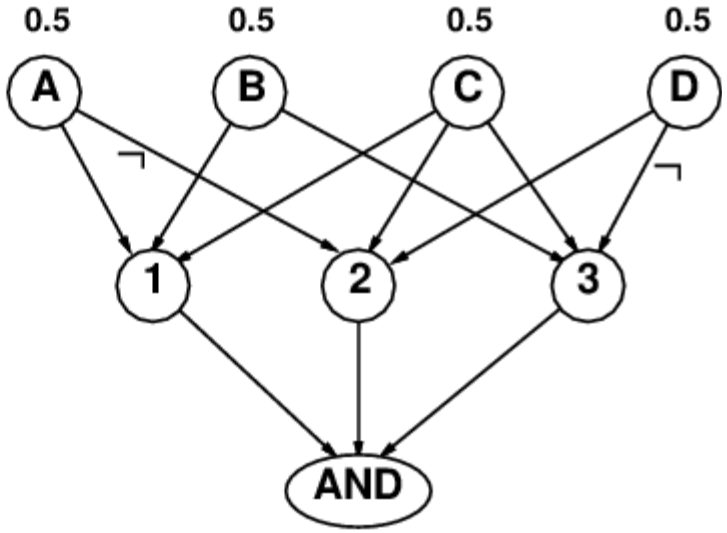
- Definition: moral graph of Bayes net: marry all parents and drop arrows
- Definition: **A** is m-separated from **B** by **C** iff separated by **C** in the moral graph
- Theorem 2: *Y* is irrelevant if m-separated from *X* by **E**



- For $P(JohnCalls|Alarm = true)$, both *Burglary* and *Earthquake* are irrelevant

Complexity of Exact Inference

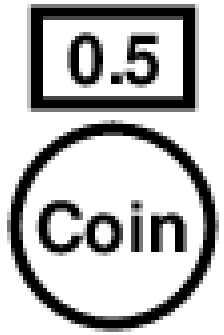
- Singly connected networks (or polytrees)
  - any two nodes are connected by at most one (undirected) path
  - time and space cost of variable elimination are $O(d^k n)$
- Multiply connected networks
  - can reduce 3SAT to exact inference ⇒ NP-hard
  - equivalent to **counting** 3SAT models ⇒ #P-complete



1. A v B v C
2. C v D v ¬A
3. B v C v ¬D

Inference by Stochastic Simulation

- Basic idea
  - Draw N samples from a sampling distribution S
  - Compute an approximate posterior probability $\hat{P}$
  - Show this converges to the true probability P



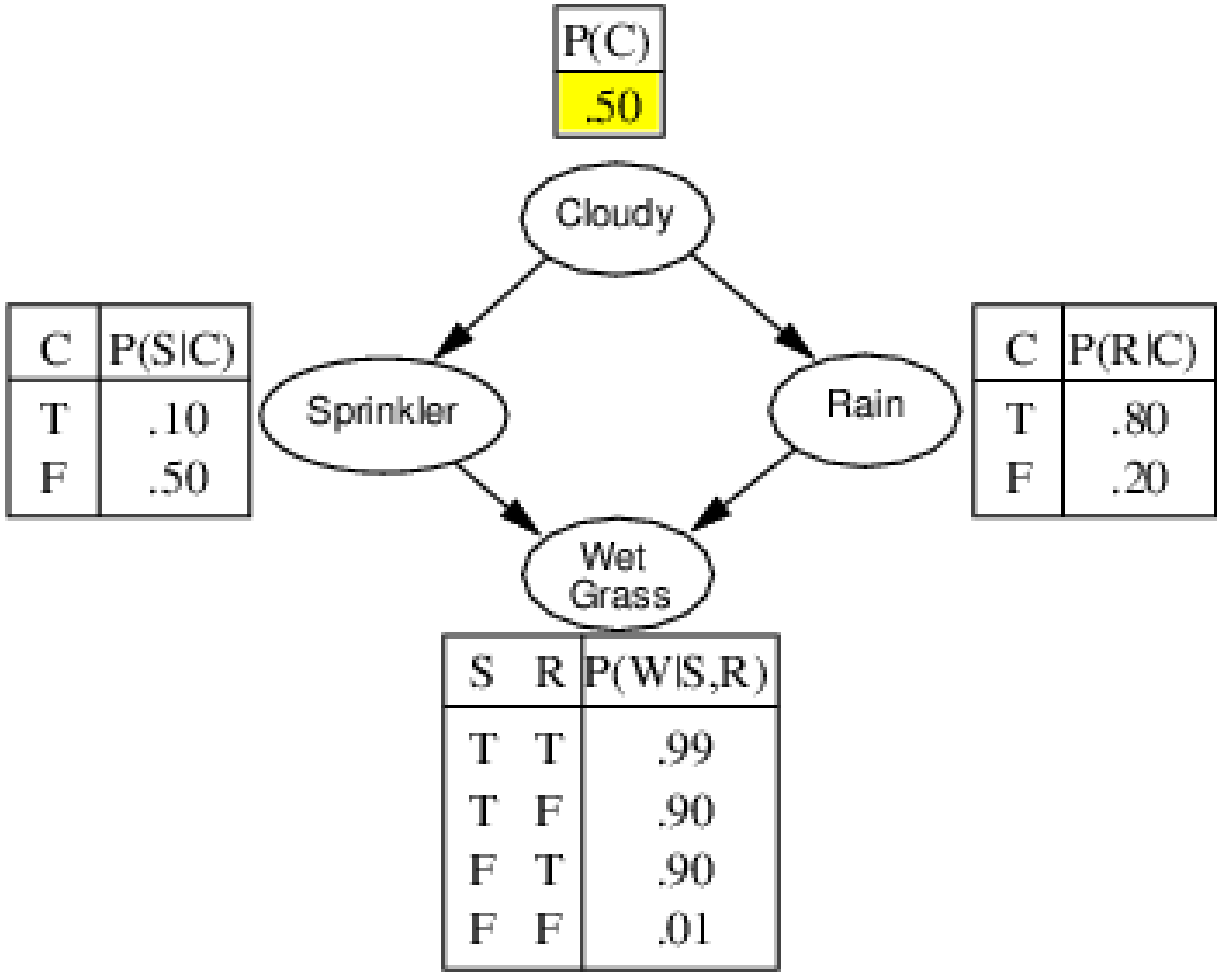
- Outline
  - Sampling from an empty network
  - Rejection sampling: reject samples disagreeing with evidence
  - Likelihood weighting: use evidence to weight samples
  - Markov chain Monte Carlo (MCMC): sample from a stochastic process whose stationary distribution is the true posterior

Sampling from an Empty Network

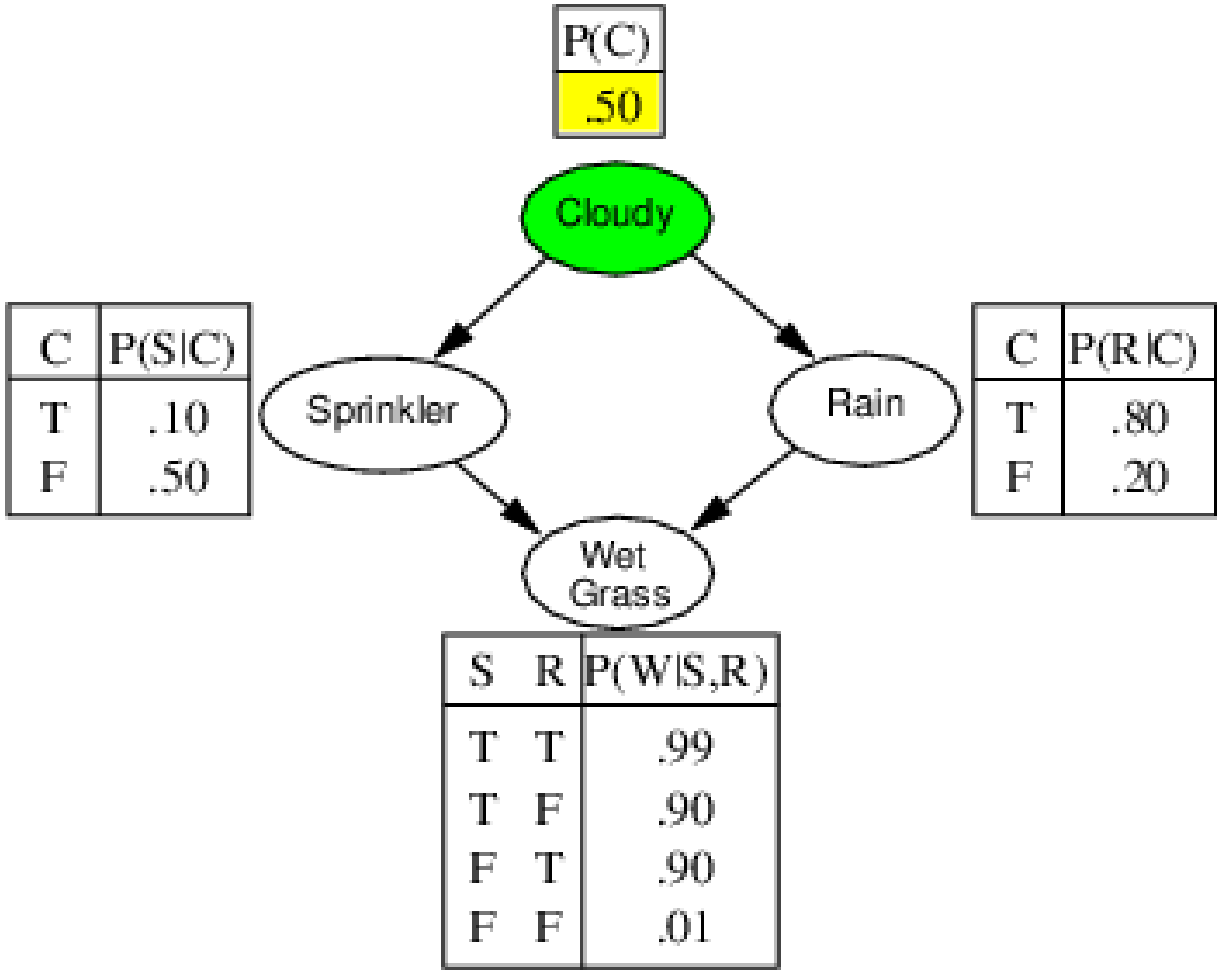
function PRIOR-SAMPLE(*bn*) **returns** an event sampled from *bn*
  inputs: *bn*, a belief network specifying joint distribution $\mathbf{P}(X_1, \ldots, X_n)$

  x ← an event with n elements
  for $i = 1$ **to** n **do**
    x_i ← a random sample from $\mathbf{P}(X_i \mid parents(X_i))$
      given the values of $Parents(X_i)$ in **x**
  return x

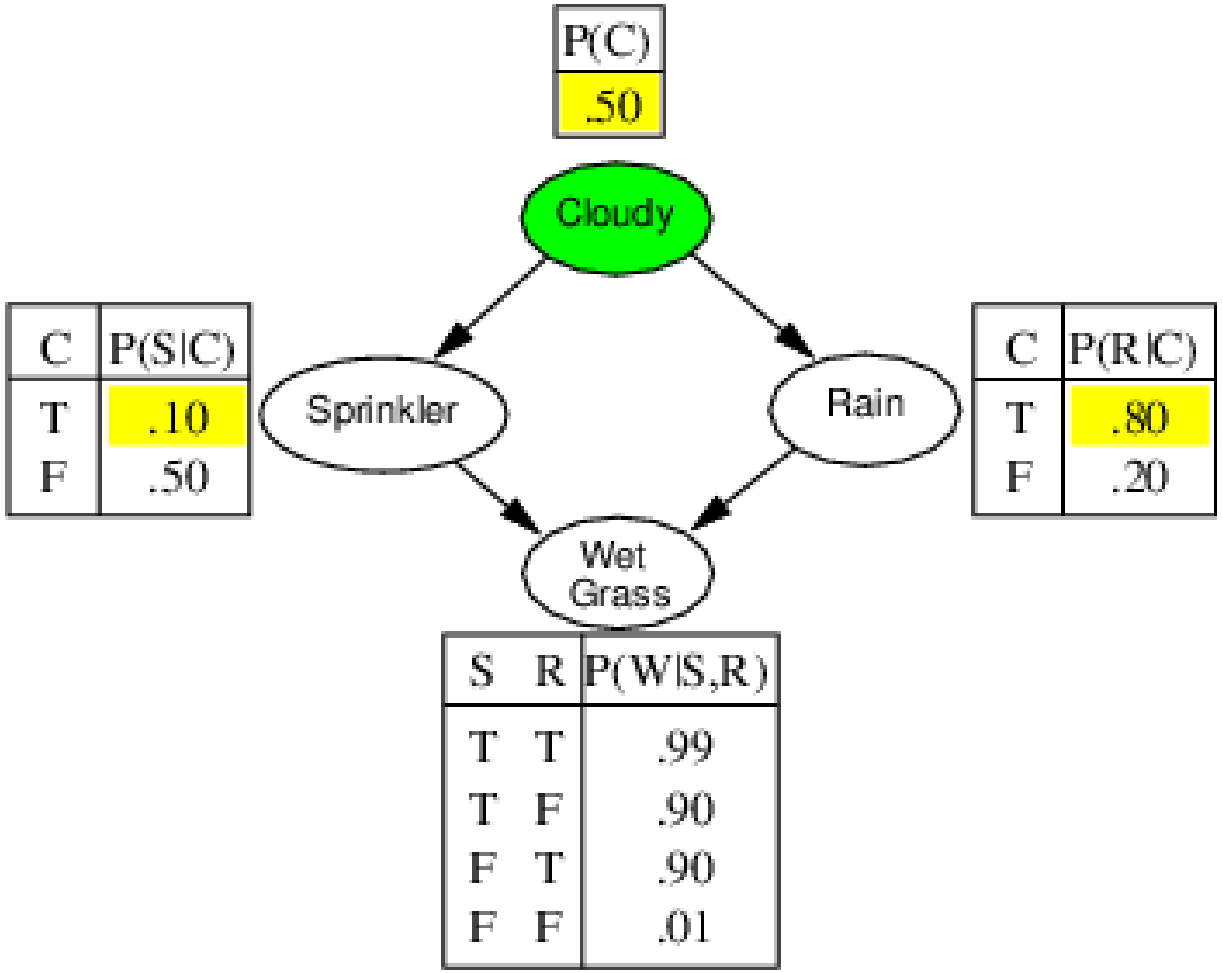
Example


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

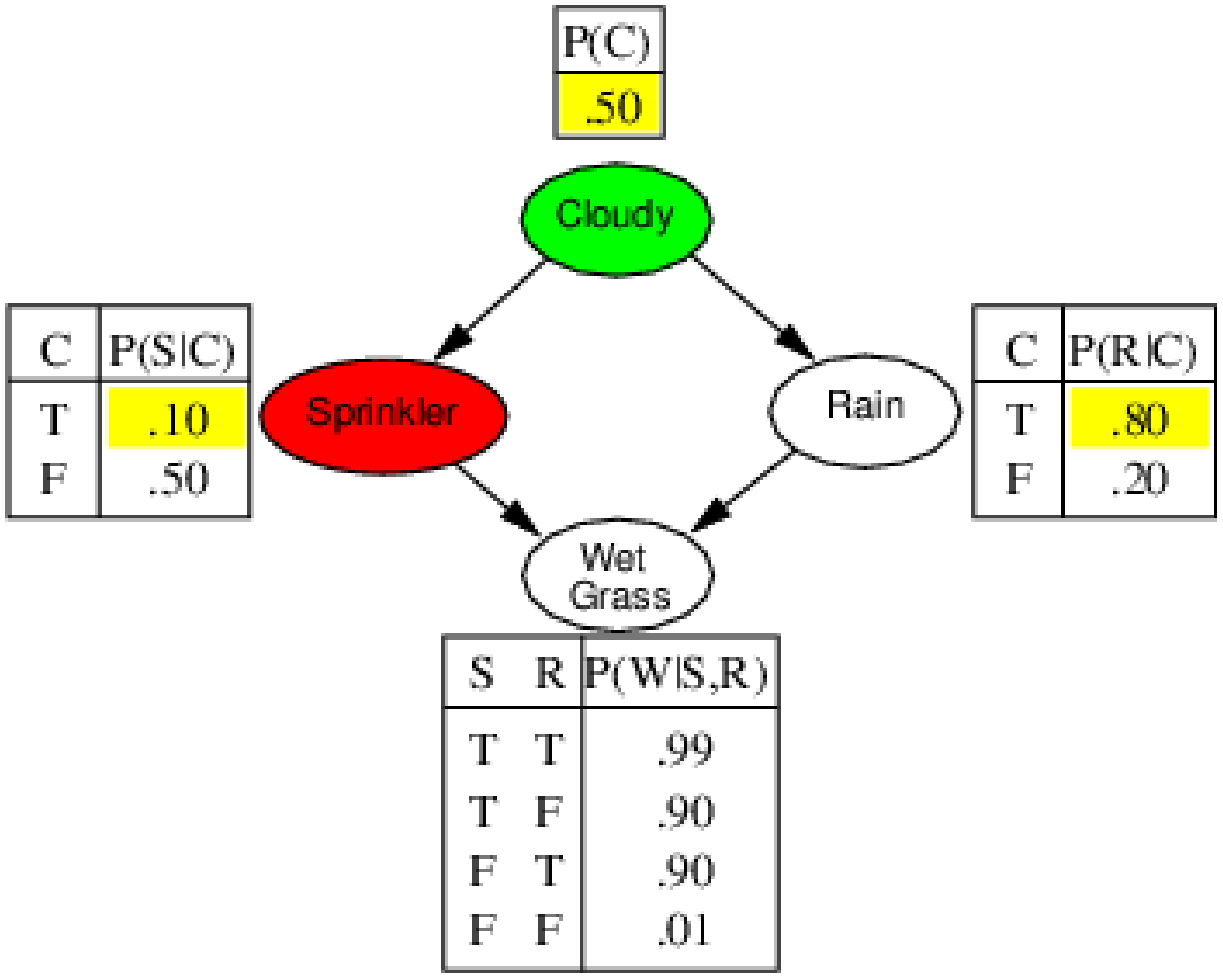
Example


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

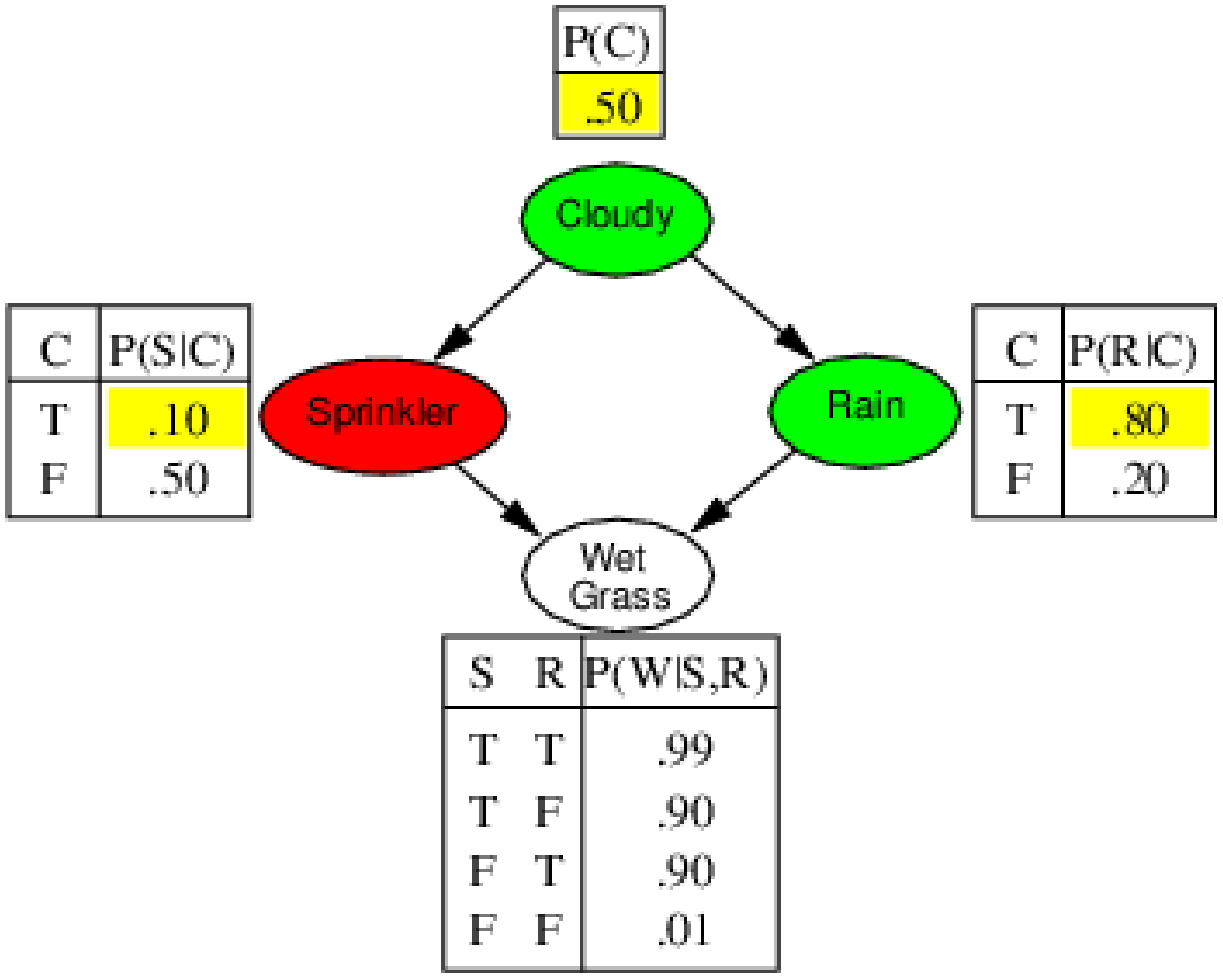
Example


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

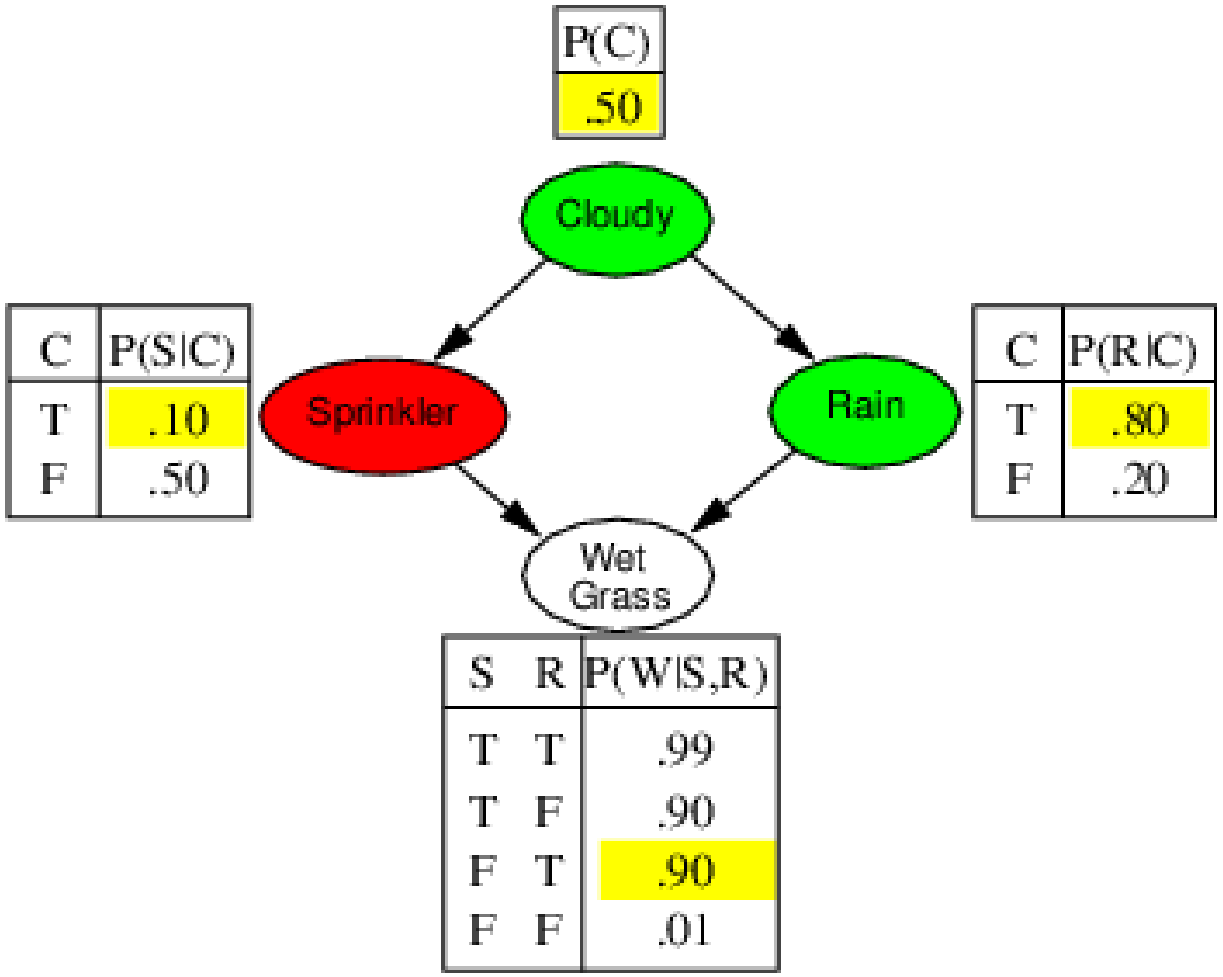
Example


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

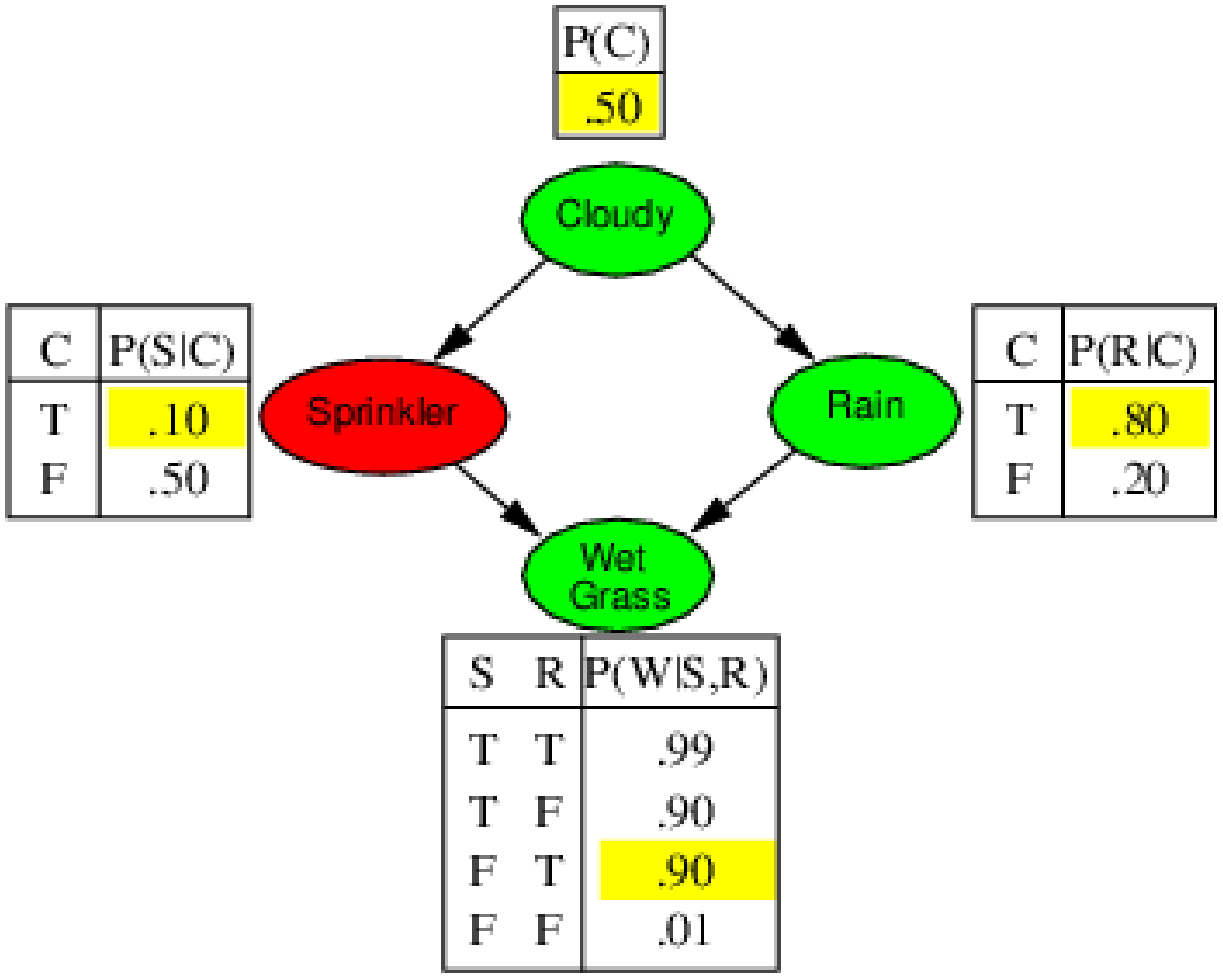
Example


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

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

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

Sampling from an Empty Network

- Probability that PRIORSAMPLE generates a particular event

$$S_{PS}(x_1 \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i)) = P(x_1 \ldots x_n)$$

  i.e., the true prior probability

- E.g., $S_{PS}(t,f,t,t) = 0.5 \times 0.9 \times 0.8 \times 0.9 = 0.324 = P(t,f,t,t)$

- Let $N_{PS}(x_1 \ldots x_n)$ be the number of samples generated for event $x_1, \ldots, x_n$

- Then we have

$$\begin{aligned} \lim_{N\to\infty} \hat{P}(x_1, \ldots, x_n) &= \lim_{N\to\infty} N_{PS}(x_1, \ldots, x_n)/N \\ &= S_{PS}(x_1, \ldots, x_n) \\ &= P(x_1 \ldots x_n) \end{aligned}$$

- That is, estimates derived from PRIORSAMPLE are consistent

- Shorthand: $\hat{P}(x_1, \ldots, x_n) \approx P(x_1 \ldots x_n)$

Rejection Sampling

- $\hat{\mathbf{P}}(X|\mathbf{e})$ estimated from samples agreeing with **e**

function REJECTION-SAMPLING(X, **e**, bn, N) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: **N**, a vector of counts over X, initially zero
  for $j = 1$ to N **do**
    x ← PRIOR-SAMPLE(bn)
    if x is consistent with **e then**
      N[x] ← **N**[x]+1 where x is the value of X in **x**
  return NORMALIZE(**N**[X])

- E.g., estimate $\mathbf{P}(Rain|Sprinkler = true)$ using 100 samples
  - 27 samples have $Sprinkler = true$
    - of these, 8 have $Rain = true$ and 19 have $Rain = false$
- $\hat{\mathbf{P}}(Rain|Sprinkler = true) = \text{NORMALIZE}(\langle 8, 19\rangle) = \langle 0.296, 0.704\rangle$
- Similar to a basic real-world empirical estimation procedure

Analysis of Rejection Sampling

$\hat{\mathbf{P}}(X|\mathbf{e}) = \alpha \mathbf{N}_{PS}(X,\mathbf{e})$ (algorithm defn.)
$= \mathbf{N}_{PS}(X,\mathbf{e})/N_{PS}(\mathbf{e})$ (normalized by $N_{PS}(\mathbf{e})$)
$\approx \mathbf{P}(X,\mathbf{e})/P(\mathbf{e})$ (property of PRIORSAMPLE)
$= \mathbf{P}(X|\mathbf{e})$ (defn. of conditional probability)

- Hence rejection sampling returns consistent posterior estimates
- Problem: hopelessly expensive if $P(\mathbf{e})$ is small
- $P(\mathbf{e})$ drops off exponentially with number of evidence variables!

Likelihood Weighting

- Idea: fix evidence variables, sample only nonevidence variables, and weight each sample by the likelihood it accords the evidence

function LIKELIHOOD-WEIGHTING(*X*, **e**, *bn*, *N*) **returns** an estimate of $P(X|\mathbf{e})$
  local variables: **W**, a vector of weighted counts over *X*, initially zero

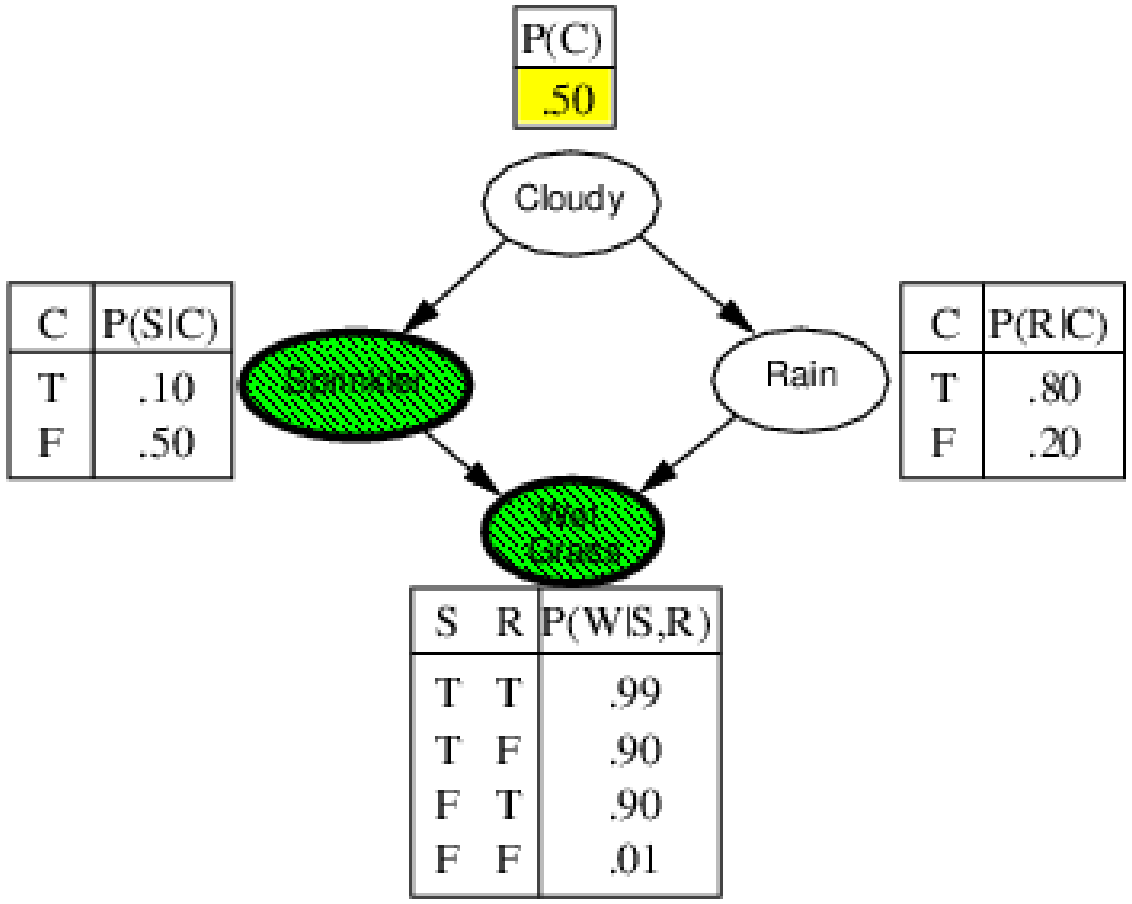
  for *j* = 1 to *N* **do**
    x, *w* ← WEIGHTED-SAMPLE(*bn*)
    W$[x]$ ← **W**$[x]$ + w where *x* is the value of *X* in **x**
  return NORMALIZE(**W**$[X]$)

---

function WEIGHTED-SAMPLE(*bn*, **e**) **returns** an event and a weight

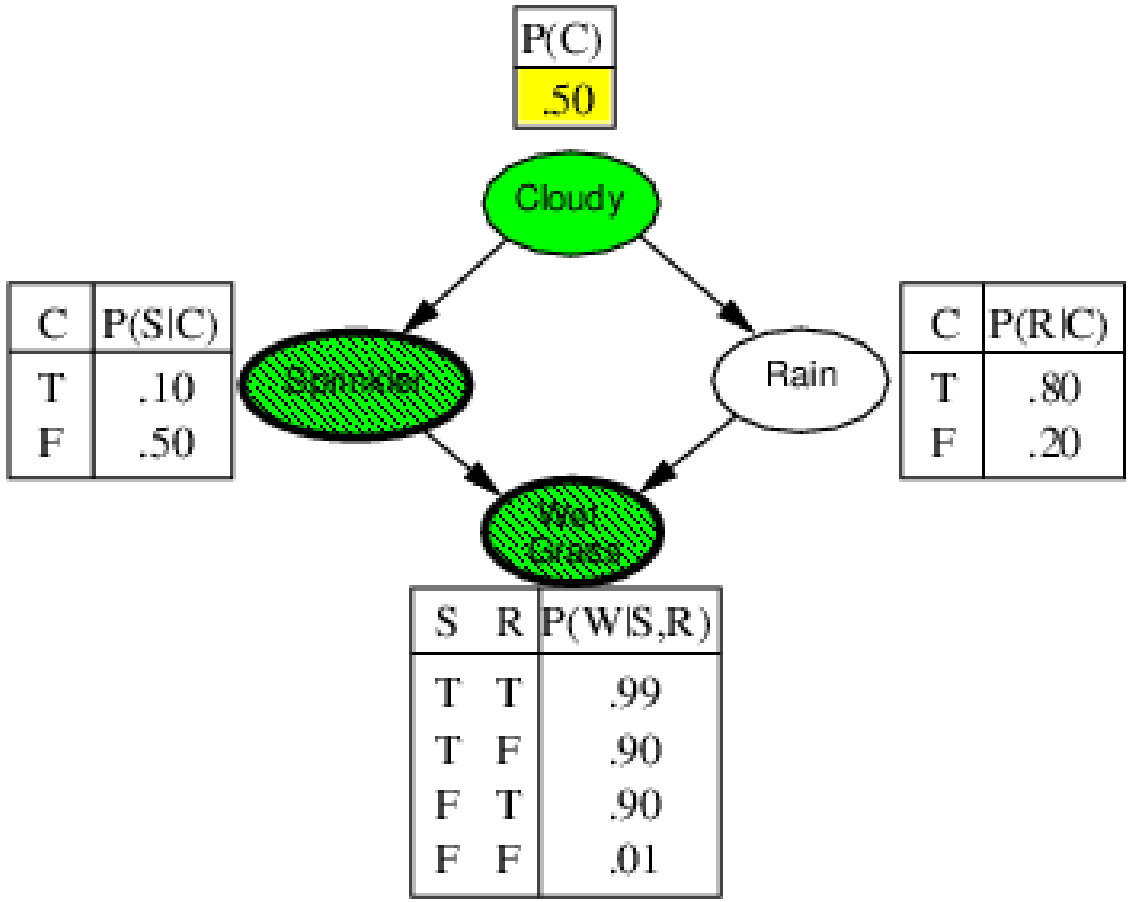
  x ← an event with n elements; *w* ← 1
  for *i* = 1 **to** n **do**
    if X_i has a value x_i in **e**
      then *w* ← $w \times P(X_i = x_i \mid parents(X_i))$
      else x_i ← a random sample from **P**$(X_i \mid parents(X_i))$
  return **x**, *w*

Likelihood Weighting Example



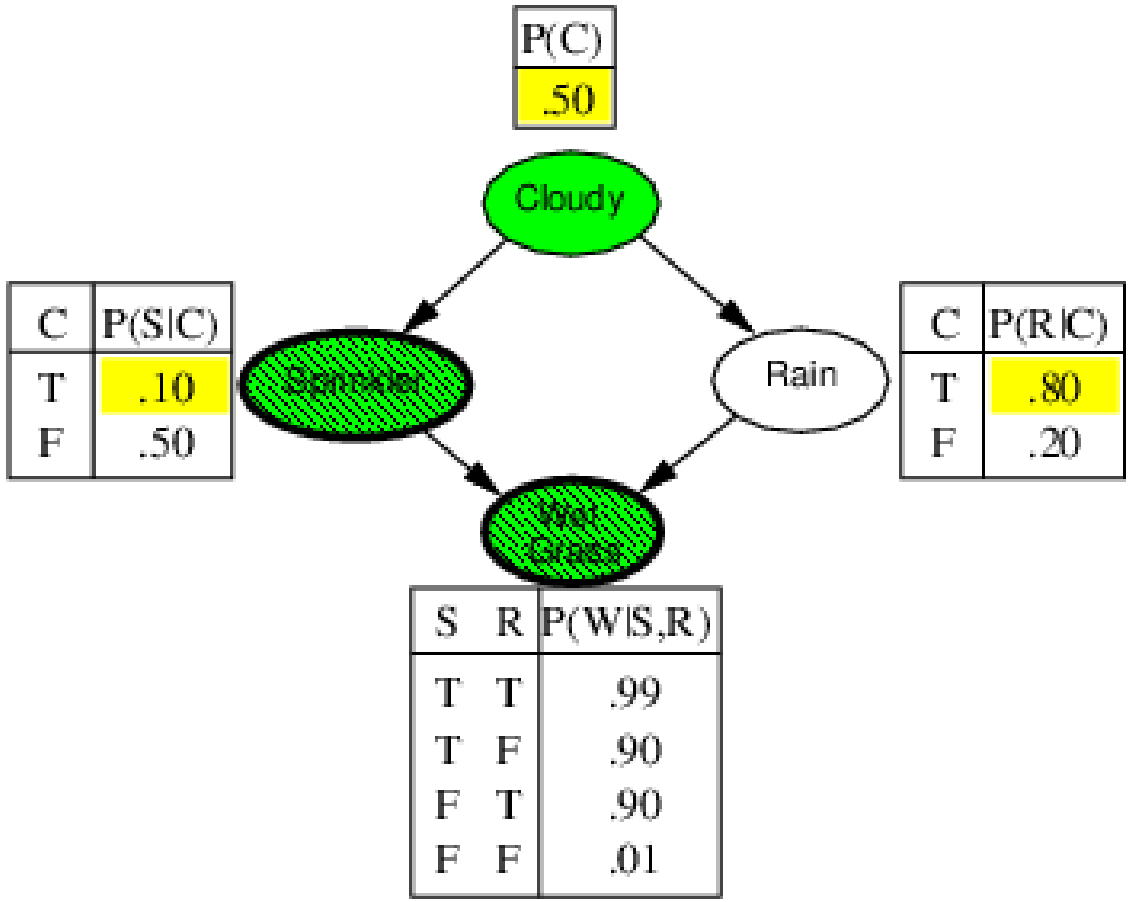
w=1.0

Likelihood Weighting Example



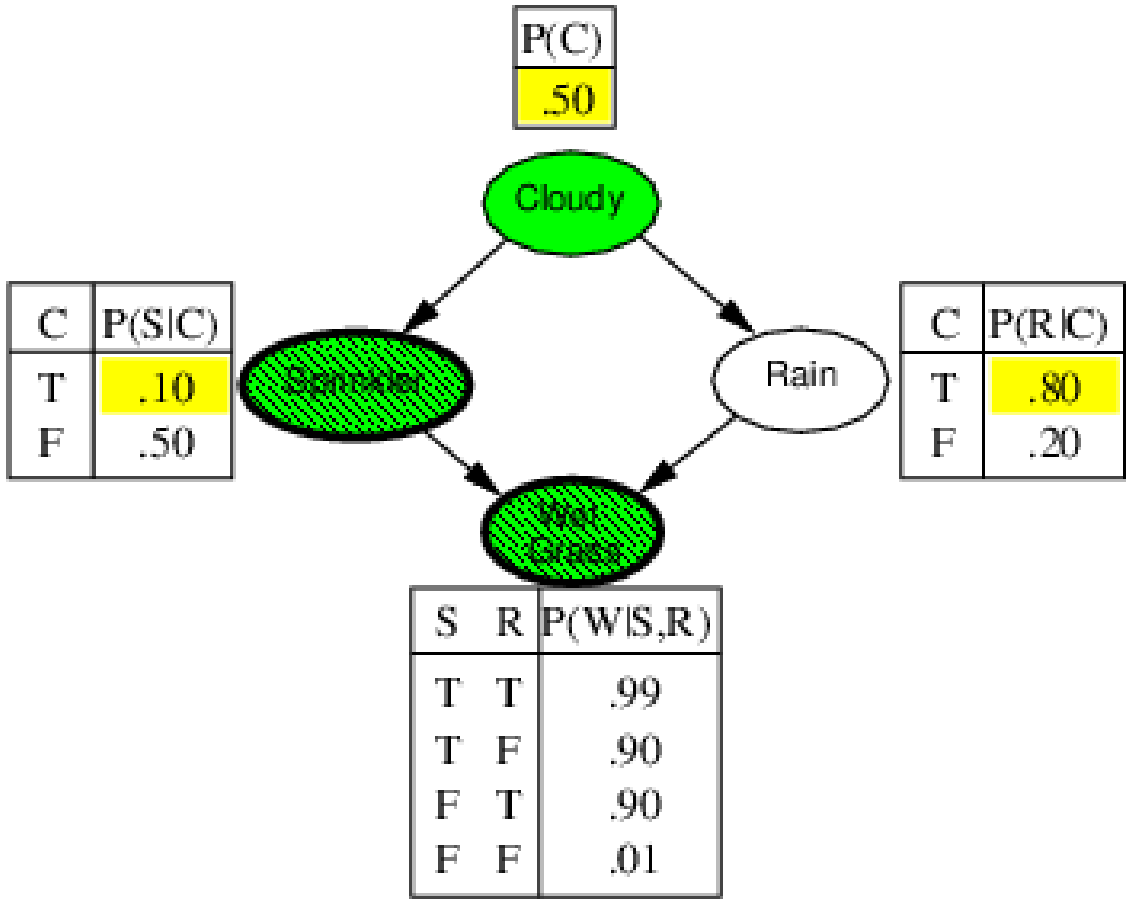
w=1.0

Likelihood Weighting Example



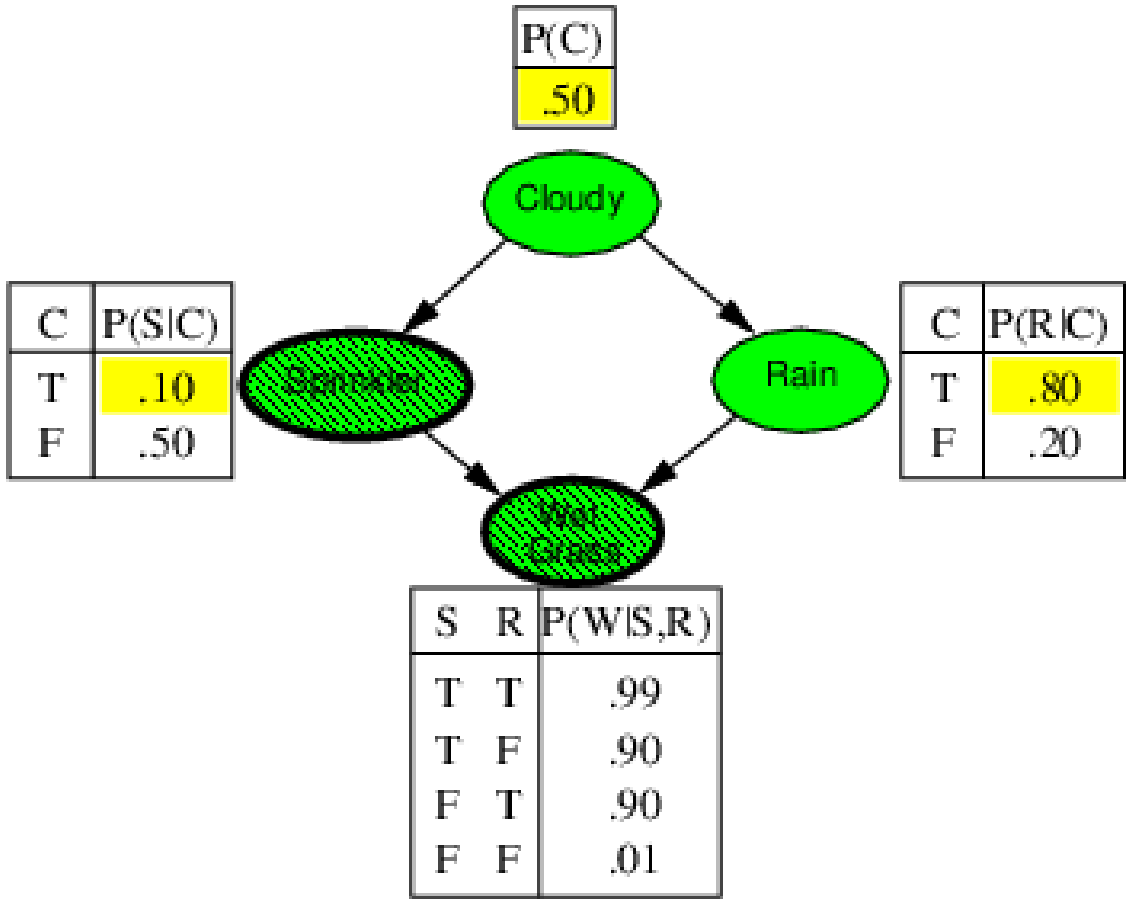
w=1.0

Likelihood Weighting Example



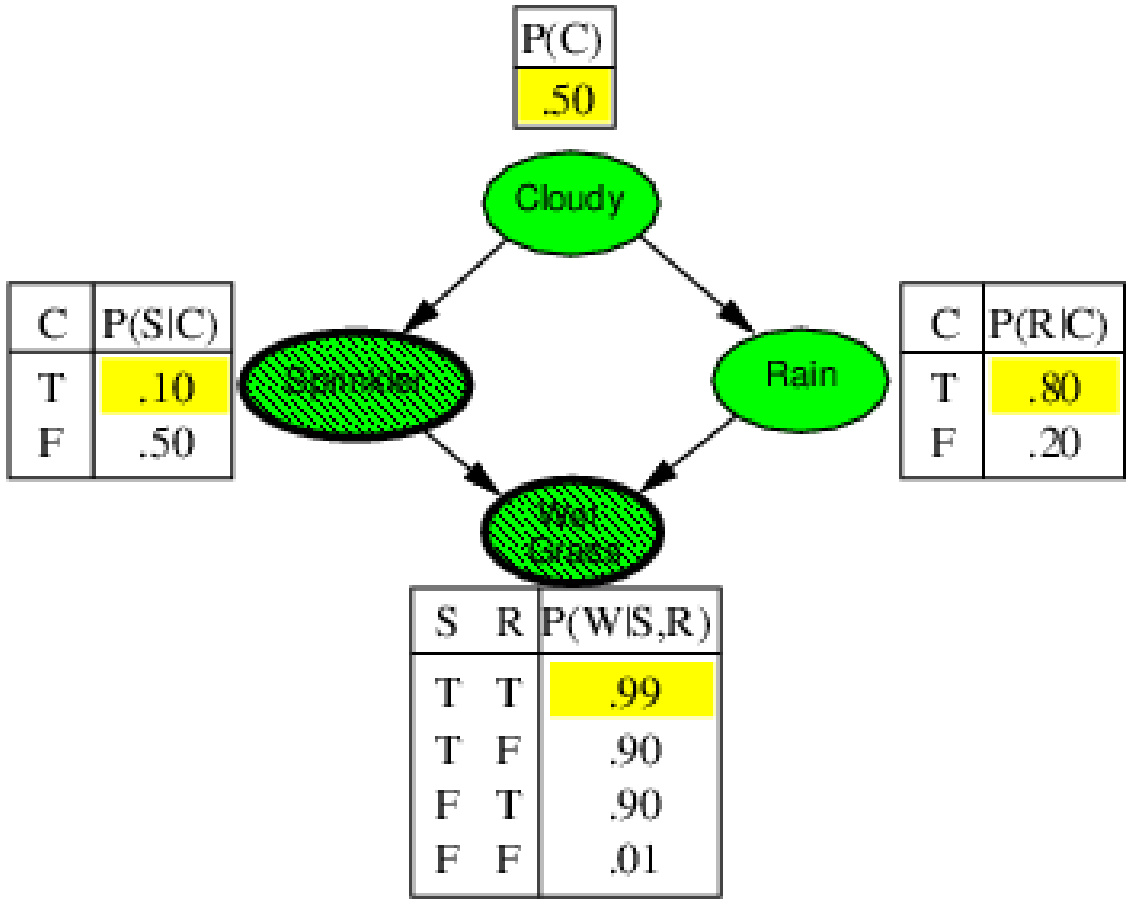
$w = 1.0 \times 0.1$

Likelihood Weighting Example



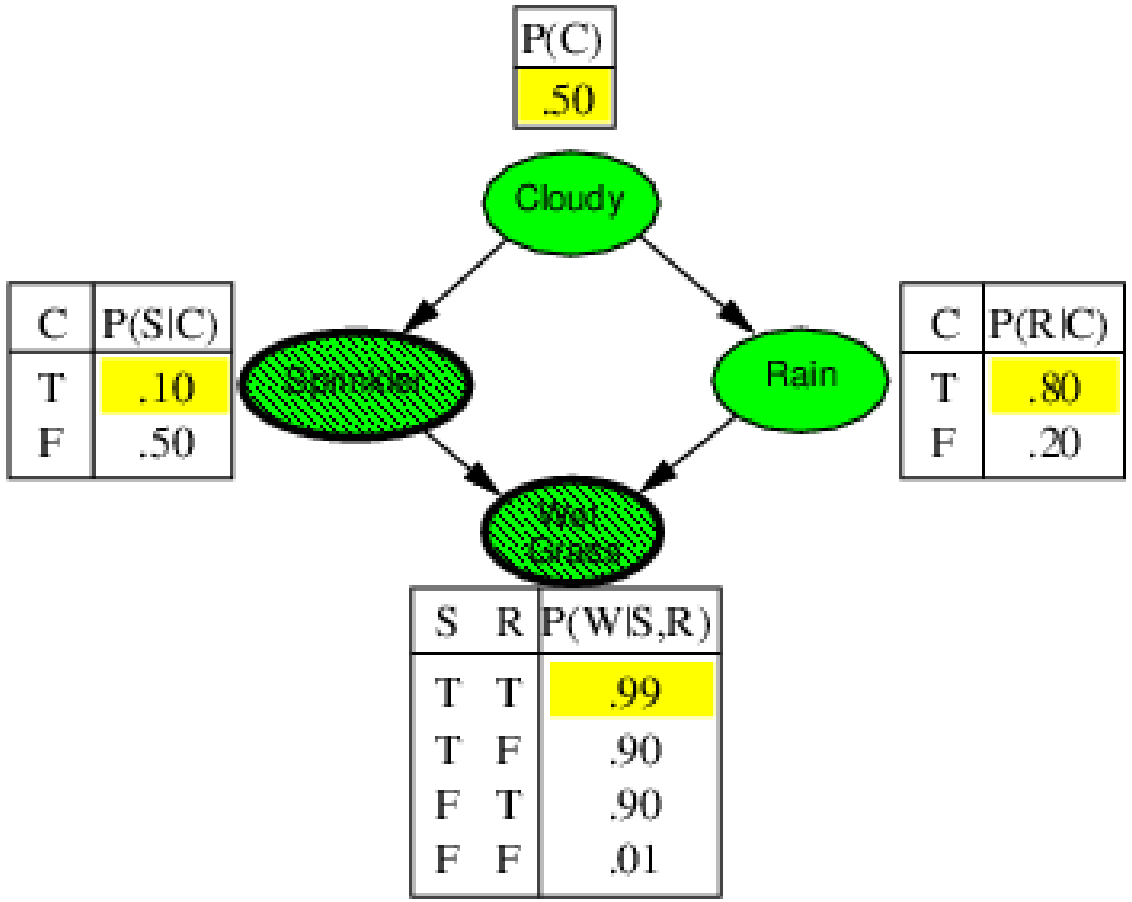
$w = 1.0 \times 0.1$

Likelihood Weighting Example



$w = 1.0 \times 0.1$

Likelihood Weighting Example



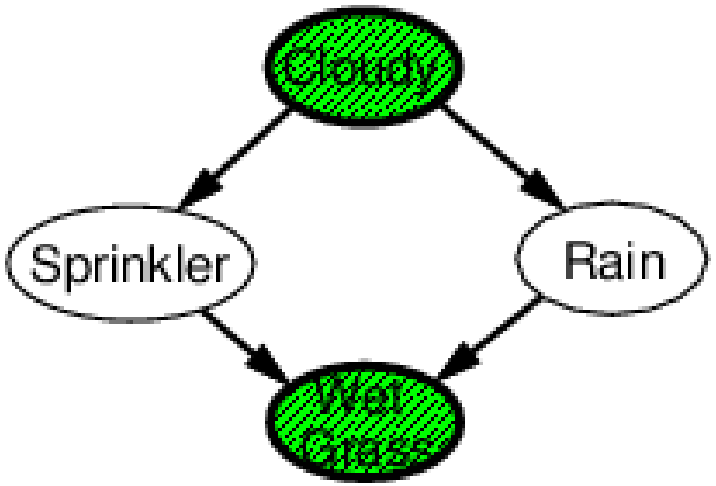
$w = 1.0 \times 0.1 \times 0.99 = 0.099$

Likelihood Weighting Analysis

- Sampling probability for WEIGHTEDSAMPLE is

$$S_{WS}(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{l} P(z_i|parents(Z_i))$$

- Note: pays attention to evidence in **ancestors** only
  ⇒ somewhere "in between" prior and posterior distribution



- Weight for a given sample **z, e** is

$$w(\mathbf{z},\mathbf{e}) = \prod_{i=1}^{m} P(e_i|parents(E_i))$$

- Weighted sampling probability is

$$S_{WS}(\mathbf{z},\mathbf{e})w(\mathbf{z},\mathbf{e})$$
$$= \prod_{i=1}^{l} P(z_i|parents(Z_i)) \prod_{i=1}^{m} P(e_i|parents(E_i))$$
$$= P(\mathbf{z},\mathbf{e})$$ (by standard global semantics of network)

- Hence likelihood weighting returns consistent estimates but performance still degrades with many evidence variables because a few samples have nearly all the total weight

Approximate Inference using MCMC

"State" of network = current assignment to all variables

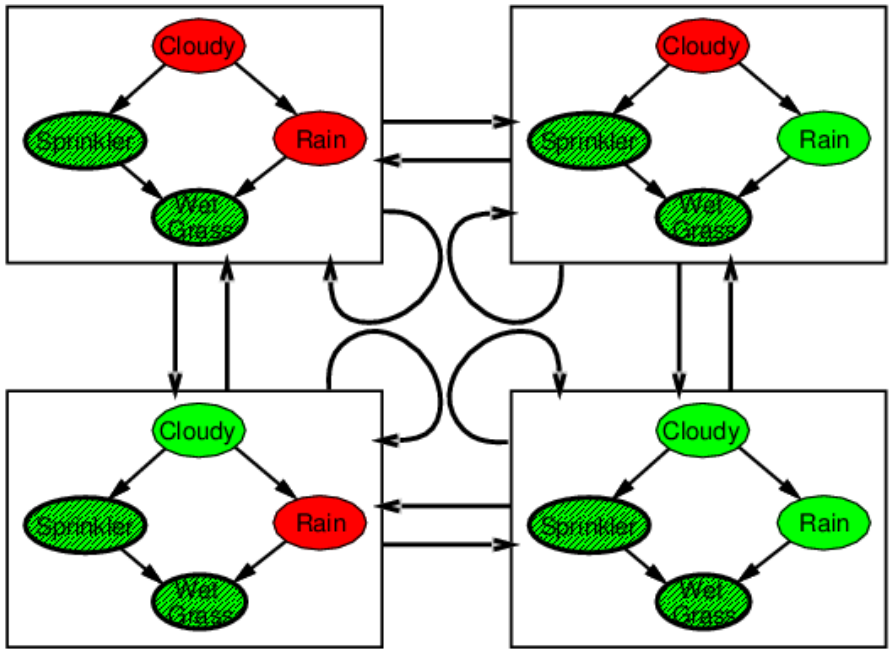
- Generate next state by sampling one variable given Markov blanket
  Sample each variable in turn, keeping evidence fixed

```
function MCMC-ASK(X, e, bn, N) returns an estimate of P(X|e)
  local variables: N[X], a vector of counts over X, initially zero
                   Z, the nonevidence variables in bn
                   x, the current state of the network, initially copied from e
  initialize x with random values for the variables in Y
  for j = 1 to N do
    for each Z_i in Z do
      sample the value of Z_i in x from P(Z_i|mb(Z_i))
        given the values of MB(Z_i) in x
      N[x] ← N[x] + 1 where x is the value of X in x
  return NORMALIZE(N[X])
```

- Can also choose a variable to sample at random each time

The Markov Chain

- With *Sprinkler=true,WetGrass=true*, there are four states:



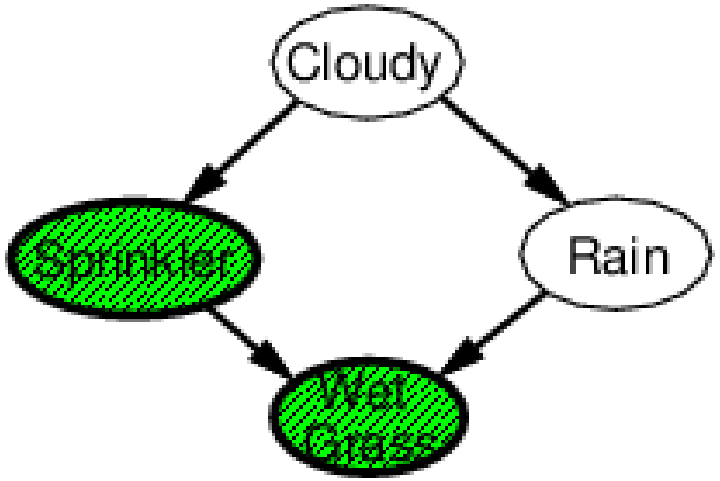
- Wander about for a while, average what you see

MCMC Example

- Estimate $\mathbf{P}(Rain|Sprinkler = true, WetGrass = true)$
- Sample *Cloudy* or *Rain* given its Markov blanket, repeat. Count number of times *Rain* is true and false in the samples.
- E.g., visit 100 states
  31 have *Rain =true*, 69 have *Rain=false*
- $\hat{\mathbf{P}}(Rain|Sprinkler = true, WetGrass = true)$
  $= \text{NORMALIZE}(\langle 31, 69 \rangle) = \langle 0.31, 0.69 \rangle$
- Theorem: chain approaches stationary distribution:
  long-run fraction of time spent in each state is exactly proportional to its posterior probability

Markov Blanket Sampling

- Markov blanket of *Cloudy* is *Sprinkler* and *Rain*
- Markov blanket of *Rain* is
  Cloudy, *Sprinkler*, and *WetGrass*
- Probability given the Markov blanket is calculated as follows:
  $P(x_i'|mb(X_i)) = P(x_i'|parents(X_i)) \prod_{Z_j \in Children(X_i)} P(z_j|parents(Z_j))$
- Easily implemented in message-passing parallel systems, brains
- Main computational problems
  - difficult to tell if convergence has been achieved
  - can be wasteful if Markov blanket is large:



Goals

- Understand how to conduct probabilistic reasoning.
- Understand how to construct Bayesian networks.
- Understand how to do exact inference by variable elimination.
- Understand how to approximately infer by LW and MCMC.
- Know how to implement the inference algorithms.