CS 561 Artificial Intelligence
Lecture # 4 Inference in Bayesian Networks

Rashmi Dutta Baruah

Department of Computer Science & Engineering

IIT Guwahati

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी

Indian Institute of Technology Guwahati

Guwahati - 781039, INDIA

Outline

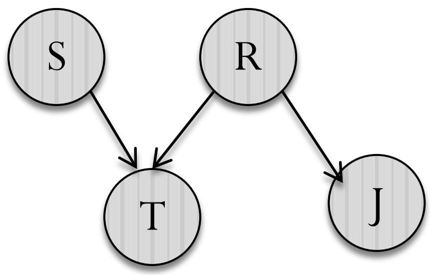
- Inference in Bayesian Network
  - Exact Inference
  - Approximate Inference
- Exact Inference
  - Inference by Enumeration
  - Variable Elimination
- Approximate inference: Sampling
- Sampling
  - Direct Sampling methods
    - Forward sampling
    - Rejection sampling
    - Likelihood sampling
  - Markov chain sampling

Conditional Independence

- **Example:** One morning Tracey leaves her house and realise that her grass is wet. Is it due to overnight rain or did she forget to turn off the sprinkler last night?
- Next she notices that the grass of her neighbour, Jack, is also wet.
- This **explains away** to some extent the possibility that her sprinkler was left on, and she concludes therefore that it is probably been raining (it decreases her belief that the sprinkler is on).
- Using the following four propositional random variables, construct the BN and determine if S is d-separated from J when T is known.
  - R: Rain $\in \{0,1\}$ (Rain = 1 means that it has been raining, and 0 otherwise)
  - S: Sprinkler $\in \{0,1\}$
  - J: Jack's grass wet $\in \{0,1\}$
  - T: Tracy's Grass wet

Conditional Independence

- Four propositional random variables are:
  - R: Rain $\in \{0,1\}$ (Rain = 1 means that it has been raining, and 0 otherwise)
  - S: Sprinkler $\in \{0,1\}$
  - J: Jack's grass wet $\in \{0,1\}$
  - T: Tracy's Grass wet

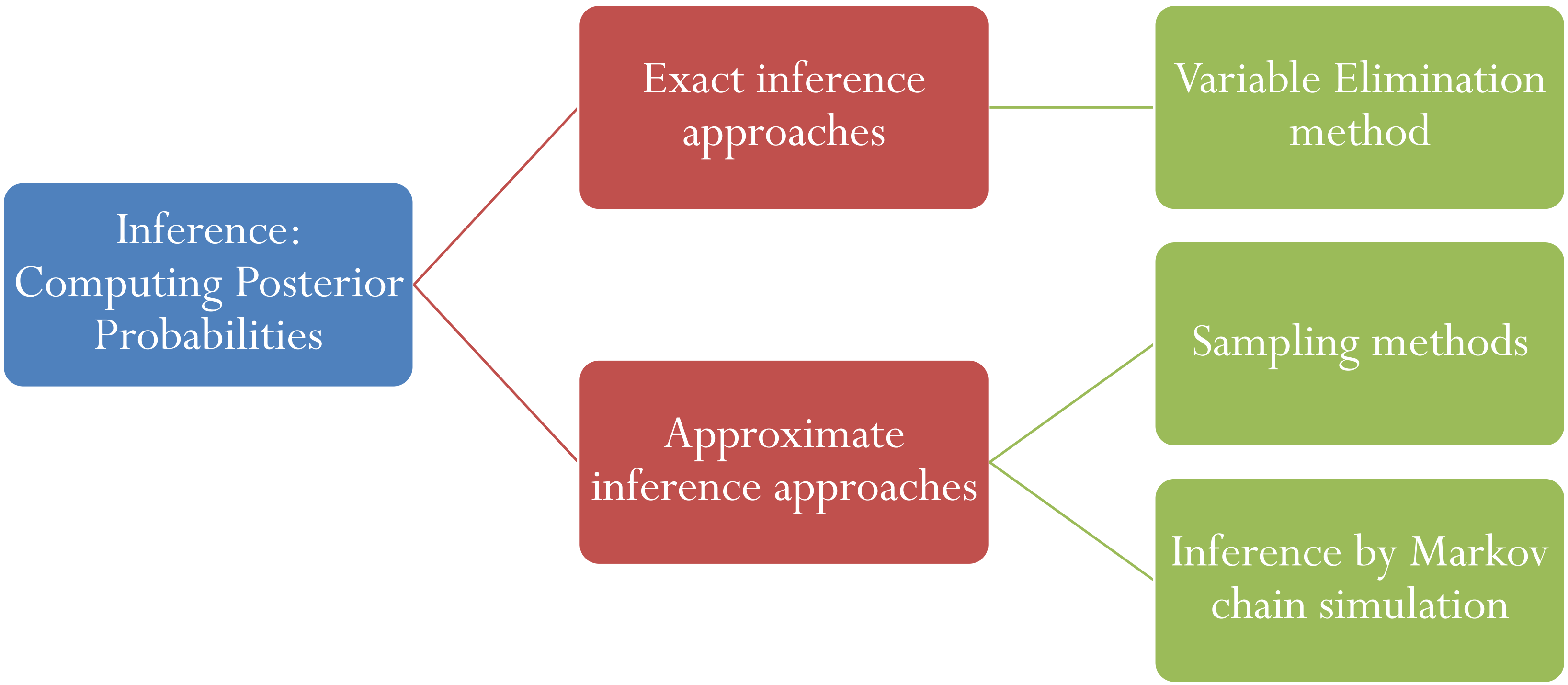


- The trail between S and J: S-T-R-J
- S-T-R converging connection and T is known so influence flows from S to R.
- T-R-J diverging connection and R is not known so influence flows from T to J.
- So, S and J are not d-separated given T

Inference in Bayesian Networks

- Given a Bayesian network, what queries one might ask?
  - Simple query: compute posterior probability i.e. $P(X_i \mid E=e)$
- X denotes query variable, E is set of evidence variables, E_1, …E_m, and e is particular observed event, Y denotes non-query, non-evidence variables (called hidden variables).
- Example: $P(Burglary \mid JohnCalls = true, MaryCalls = true)$
  - $X = Bruglary, E = \{JohnCalls, MaryCalls\}, Y = \{Alarm, Earthquake\}$

Inference in Bayesian Networks



Inference by Enumeration

- $\boldsymbol{P}(X \mid \boldsymbol{e}) = \alpha\, \boldsymbol{P}(X, \boldsymbol{e}) = \alpha \sum_{y} \boldsymbol{P}(X, \boldsymbol{e}, \boldsymbol{y})$: Sum over the variables not involved in the query.

$\boldsymbol{P}(\mathrm{B} \mid \mathrm{j,m}) = \alpha\, \boldsymbol{P}(B, j, m) = \alpha \sum \sum \boldsymbol{P}(B, j, m, e, a)$

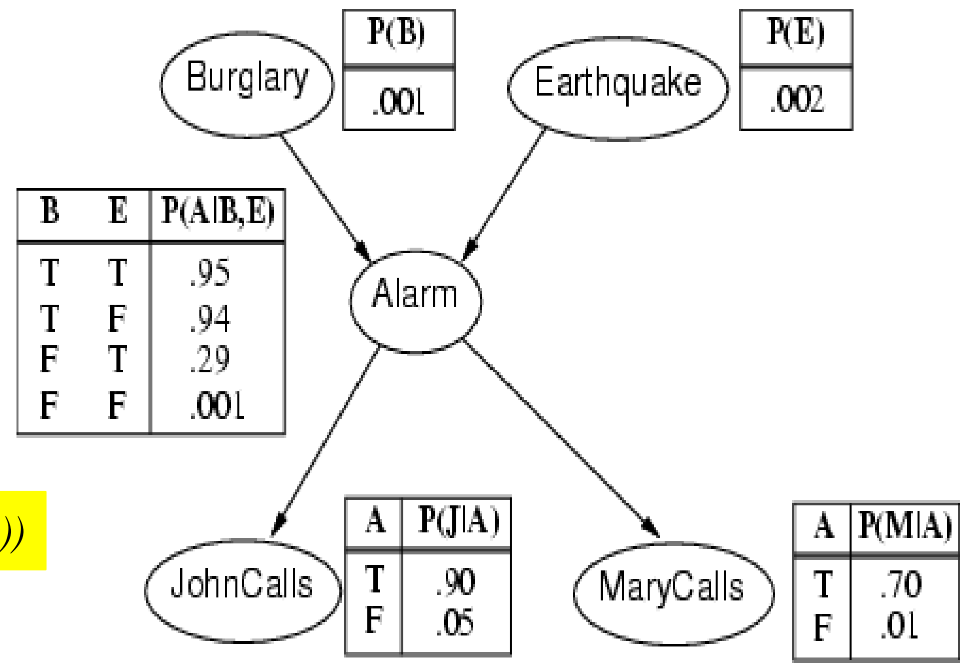
Using Bayesian network semantics we can get the expression in terms of CPTs.

Let us write this for *Burglary= true*

$P(\mathrm{b} \mid \mathrm{j,m}) =$

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

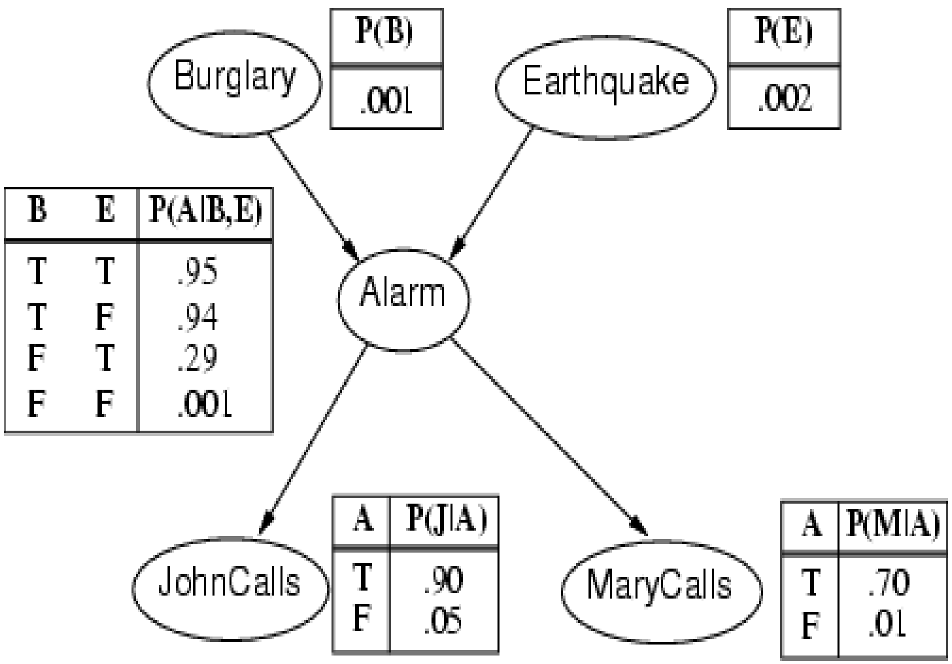
$= \alpha\, P(b) \sum_{e} P(e) \sum_{a} P(a \mid b,e)\, P(j \mid a)\, P(m \mid a)$



We know: $\boldsymbol{P}(X_1, \ldots, X_n) = \pi_{i=1} \boldsymbol{P}(X_i \mid Parents(X_i))$

Inference by Enumeration

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



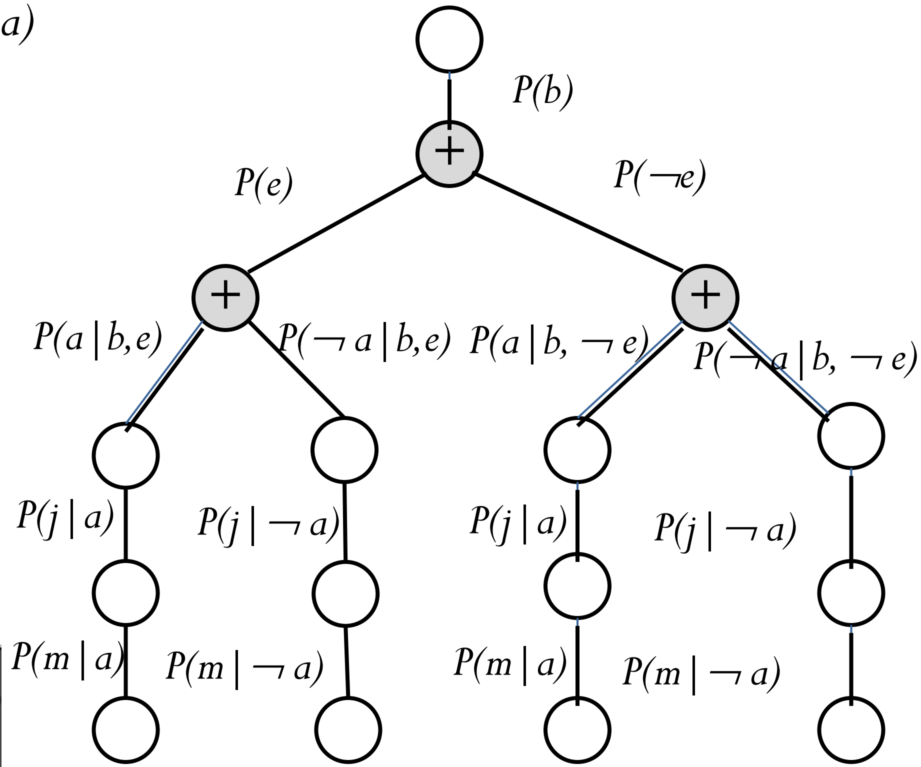
P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01



Proceed top down, multiplying values along each path and summing at the "+" nodes. Note repetition of the paths for *j* and *m* (repeated computation). For large networks takes long time.

Inference by Variable elimination

- Eliminates repeated calculations
- Variable elimination:
  - expression is evaluated from right-to-left order
  - intermediate results are stored
  - summations over each variable are done only for those portions of the expression that depend on the variable.

$$\mathbf{P}(B \mid j,m) = \alpha\, \mathbf{P}(B) \sum_{e} P(e) \sum_{a} P(a \mid B,e)\, P(j \mid a)\, P(m \mid a)$$

$$f_1(B) \qquad f_2(E) \quad f_3(A,B,E) \quad f_4(A) \quad f_5(A) \longleftarrow \quad \textit{factor}$$

$$= \alpha\, f_1(B) \times \sum_{e} f_2(E) \times \sum_{a} f_3(A,B,E) \times f_4(A) \times f_5(A)$$

Matrix

Point-wise product operation

$$Eg.\ f_4(A) = \begin{pmatrix} P(j|a) \\ P(j|\sim a) \end{pmatrix} = \begin{pmatrix} 0.90 \\ 0.05 \end{pmatrix}$$

Inference by variable elimination

- Sum out variables (right-to-left) from point-wise products of factors to produce new factors, eventually yielding a factor that is the solution.
  - sum out A from the product f_3, f_4, and f_5 giving f_6

    $$f_6(B,E) = \sum_a f_3(A,B,E) \times f_4(A) \times f_5(A)$$
    $$= (f_3(a,B,E) \times f_4(a) \times f_5(a)) + (f_3(\sim a,B,E) \times f_4(\sim a) \times f_5(\sim a))$$

    Now we are left with expression

    $$P(B \mid j,m) = \alpha f_1(B) \times \sum_e f_2(E) \times f_6(B,E)$$
  - sum out E from the product of f_2 and f_6

    $$f_7(B) = \sum_e f_2(E) \times f_6(B,E) = f_2(e) \times f_6(B,e) + f_2(\sim e) \times f_6(B,\sim e)$$

    Now the expression becomes

    $$P(B \mid j,m) = \alpha f_1(B) \times f_7(B)$$

Inference by variable elimination

- Basic operations
  - Point-wise product
    - two factors f_1 and f_2 yields a new factor f whose variables are the union of the variables in f_1 and f_2
    - f's elements are given by product of the corresponding elements in the two factors
    - Example:
      - given two factors $f_1(A,B)$ and $f_2(B,C)$, the pointwise product $f_1 \times f_2 = f_3$ (A,B,C) has 2^{1+1+1} entries (table in next slide)
  - Summing out variable
    - It is done by adding up the submatrices formed by fixing the variable to each of its value in turn.

Inference by variable elimination

A	B	$f_1(A,B)$	B	C	$f_2(B,C)$	A	B	C	$f_3(A,B,C)$
T	T	0.3	T	T	0.2	T	T	T	$0.3 \times 0.2 = 0.06$
T	F	0.7	T	F	0.8	T	T	F	0.3 x 0.8 = 0.24
F	T	0.9	F	T	0.6	T	F	T	0.7 x 0.6 = 0.42
F	F	0.1	F	F	0.4	T	F	F	0.7 x 0.4 = 0.28
						F	T	T	0.9 x 0.2 = 0.18
						F	T	F	0.9 x 0.8 = 0.72
						F	F	T	0.1 x 0.6 = 0.06
						F	F	F	0.1 x 0.4 = 0.04

- To sum out A from f_3(A,B,C), we write
- $f(B,C) = \sum_{a} f_3(A,B,C) = f_3(a,B,C) + f_3(\sim a,B,C)$

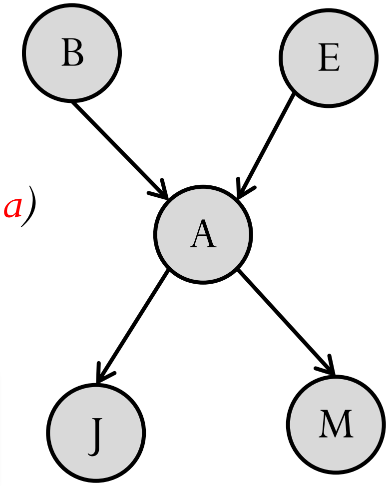
$$= \begin{pmatrix} .06 & .24 \\ .42 & .28 \end{pmatrix} + \begin{pmatrix} .18 & .72 \\ .06 & .04 \end{pmatrix} = \begin{pmatrix} .24 & .96 \\ .48 & .32 \end{pmatrix}$$

Summary Exact Inference

- Inference – compute posterior probability distribution for a set of query variables given a set of evidence variables that are observed.
- Exact methods :
  - **Inference by enumeration**
  - Simple query: $\mathbf{P}(B \mid j,m) = \alpha\, \mathbf{P}(B,j,m) = \alpha \sum \sum \mathbf{P}(B,j,m,e,a)$

  $P(B \mid j,m) = \alpha \sum_{e} \sum_{a} P(B)\, P(E)\, P(a \mid B,e)\, P(j \mid a)\, P(m \mid a)$

  Worst case time complexity: for n Boolean variable $O(n2^n)$, can be improved by moving the summation outside but still will be $O(2^n)$



  - **Variable elimination**
    - d^k entries computed for a factor over k variables with domain sizes d
    - ordering of elimination of hidden variables does matter- bad elimination order can generate large factors
    - Worst case running time exponential in the size of the Bayes' net (large multiply connected networks)

Approximate Inference

- Sampling based inference

- Basic idea: Generate random samples and compute required probabilities from samples.
  - Draw N samples from distribution
  - Estimate $P(X \mid \boldsymbol{E})$ from samples

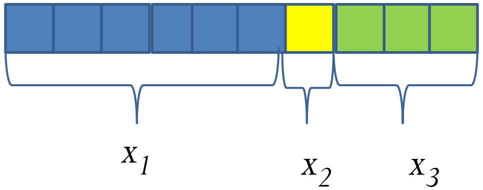
- What do we need to know?
  - How to generate a new sample ?
  - How many samples do we need ?
  - How to estimate P(X|E) ?

Sampling

- Known distribution, single variable
- Consider a random variable X with $dom(X) = \{x_1, x_2, x_3\}$
- To generate a random sample for X
  - Select a random number y in the range [0,1) (we select y from a uniform distribution to ensure that each number between 0 and 1 has same chance of being chosen)
  - Convert this value of y to a sample from given distribution by associating each outcome $\{x_1, x_2, x_3\}$ to a given sub-interval with size proportional to P(X).
  - Example: suppose `random()` returns y = 0.3, 0.25, 0.45, 0.65 …corresponding samples will be $x_1, x_1, x_1, x_2, \ldots$

X	P(X)
x_1	0.6
x_2	0.1
x_3	0.3

$$0 \leq y < 0.6 \rightarrow X = x_1$$
$$0.6 \leq y < 0.7 \rightarrow X = x_2$$
$$0.7 \leq y < 1 \rightarrow X = x_3$$



Sampling Methods

- Forward sampling
- Rejection sampling
- Likelihood weighting
- Gibbs Sampling (MCMC)

Sampling Methods

- Forward sampling / Prior sampling (without evidence)
  - Sample each variable in topological order
  - probability distribution from which the value is sampled is conditioned on the values already assigned to the variable's parents

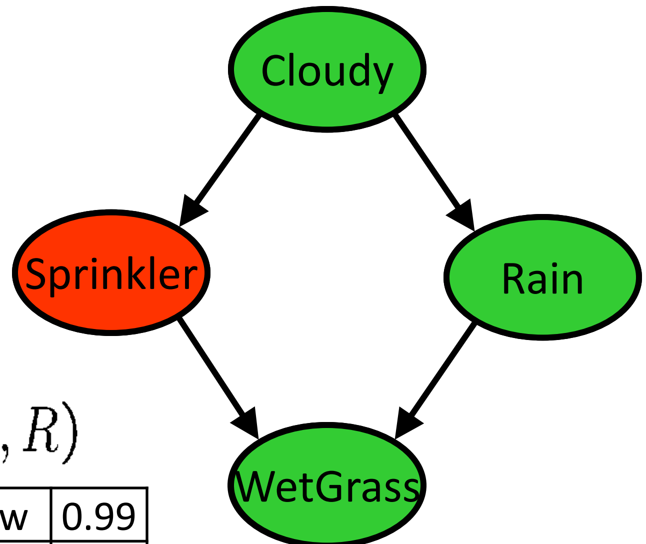
```
Input: Bayesian network
     X= {X_1,…,X_n}, n- #nodes, N - # samples
Output: N samples
Process nodes in topological order – first process
     the ancestors of a node, then the node itself
1.   For t = 1 to N
2.       For i = 1 to n
3.            X_i ← sample x_i^t from P(x_i|parents (X_i))
```

Forward Sampling

Assume ordering: Cloudy, Sprinkler, Rain, WetGrass

$P(C)$

+c	0.5
-c	0.5



$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5

$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, +w

-c, +s, -r, +w

...

Queries:

$P(w) \approx \#w/\#samples$

$P(w \mid c, s) \approx \frac{P(w,c,s)}{P(c,s)}$

Sampling methods

- Rejection Sampling (evidence available)
    - generate samples from the prior distribution specified by the network, reject all those samples that do not match the evidence.

```
Input: Bayesian network
    X= {X1,…,Xn}, n- #nodes
    E – evidence, N – # samples
Output: N samples consistent with E
1.  For t=1 to N
2.      For i=1 to n
3.          Xi ← sample xi^t from P(xi | parent(Xi))
4.          If Xi in E and Xi ≠ xi, reject sample:
5.                set i = 1 and go to step 2
```

Rejection Sampling

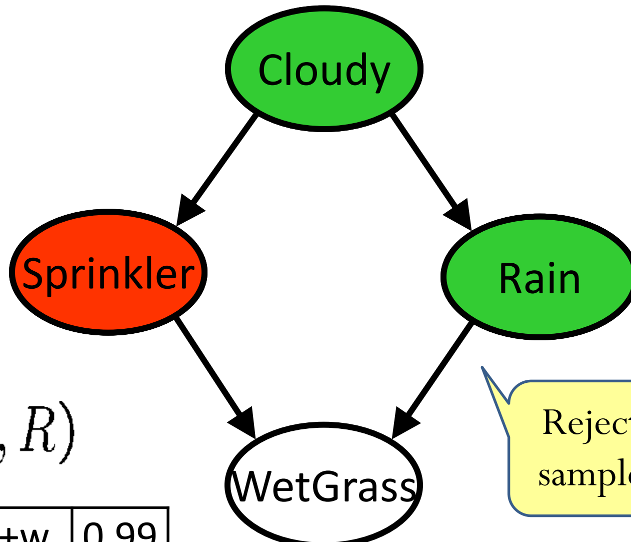
Assume ordering: Cloudy, Sprinkler, Rain, WetGrass

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

Reject sample

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, ---Reject

-c, +s, -r, +w

...

Evidence : $R = -r$

// generate sample k

1. Sample C from $P(C)$
2. Sample S from $P(S \mid C)$
3. Sample R from $P(R \mid C)$
4. If $R \neq$ -r, reject sample and start from 1, otherwise
5. Sample W from $P(W \mid S, R)$

Sampling Method

Problem of Rejection Sampling: for unlikely evidence, lots of samples rejected

- Likelihood Weighting
  - fix the values for evidence variables, and sample only non-evidence variable (not sampling from right distribution anymore)
  - Now weight the samples by evidence likelihood (probability of evidence given parents).

For $k = 1$ to N

For each each X_i in topological order $o = (X_1, \ldots, X_n)$:

$w_k = 1$

if $X_i \notin E$

$X_i \leftarrow$ sample x_i from $P(x_i \mid parents(X_i))$

else

assign $X_i = e_i$

$w_k = w_k \bullet P(e_i \mid parents(Xi))$

Assume ordering: Cloudy, Sprinkler, Rain, WetGrass

Likelihood Weighting

Query: P(_| W)

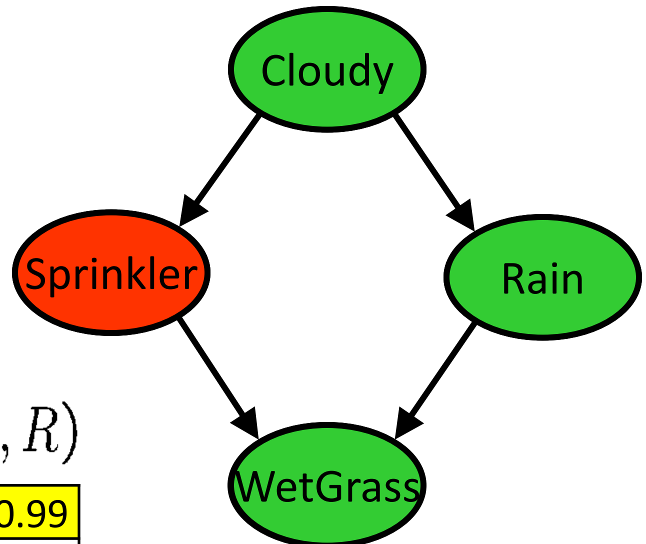
Evidence: W = +w

$P(C)$

+c	0.5
-c	0.5

$P(S|C)$

+c	+s	0.1
	-s	0.9
-c	+s	0.5
	-s	0.5



$P(R|C)$

+c	+r	0.8
	-r	0.2
-c	+r	0.2
	-r	0.8

$P(W|S,R)$

+s	+r	+w	0.99
		-w	0.01
	-r	+w	0.90
		-w	0.10
-s	+r	+w	0.90
		-w	0.10
	-r	+w	0.01
		-w	0.99

Samples:

+c, -s, +r, +w 0.90 (weight)

-c, +s, -r, +w 0.90

…