Intro to Artificial Intelligence

Lecture 4: Probabilistic inference

Ahmed Sallam {http://sallam.cf / http://sallamah.weebly.com}

Review

- Probability Theory
- Bayes Net
- Independence

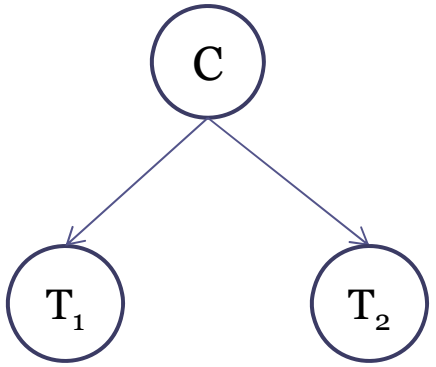
Today

- Probabilistic Inference
  - How to answer probability questions using Bayes nets?
- We will study two methods
  - Enumeration
  - Approximate inference

Adaptation

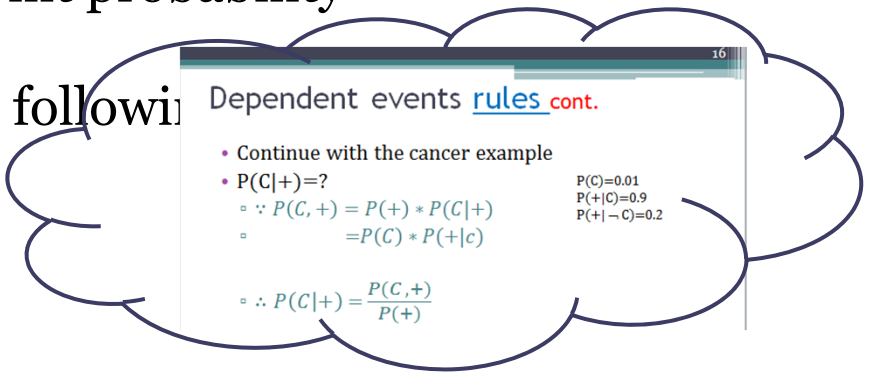
Conditional Probability

- Remember the Cancer example, where the following information are given:
  - P(C)=0.01
  - P(+|C)=0.9
  - P(-|¬C)=0.8



- Then we asked for P(C|+) ??
- And we used the Bayes rule to solve it:
  - $P(C|+) = \frac{P(C|+)*P(C)}{P(+)}$

- As you notice the nominator is the joint probability of the dependent events P(C,+)
- Thus we can rewrite the Bays rule as followi
  - $P(C|+) = \frac{P(C,+)}{P(+)}$



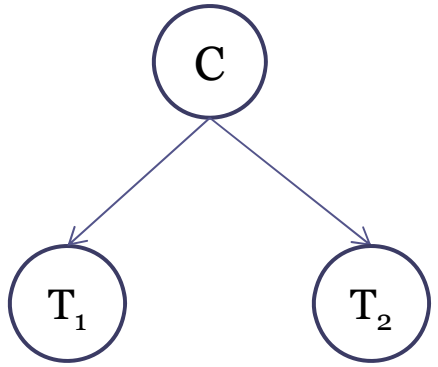
Conditional Probability cont.

- Conditional Probability general rule

- $P(A,B|C,D) = \frac{P(A,B,C,D)}{P(C,D)}$
  - Considering the dependency between events.

Boolean events writing notation

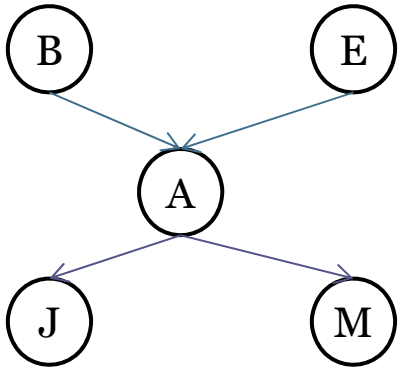
- Consider the Cancer example, suppose we have the following statements:
  - P(C)
  - $P(+_1 | \neg C)$
  - $P(-_2 | C)$



- We can write the probability statements in a more descriptive notation as following:
  - P(+c)
  - $P(+t_1|-c)$
  - $P(-t_2|+c)$

Inference terminologies

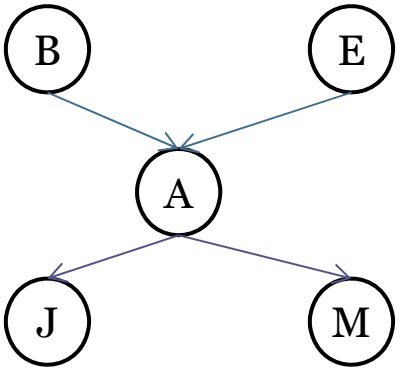
- Earthquake example
  - B stands for Burglary
  - E stands for Earthquake
  - A stands for Alarm setting off by B or E
  - J for Jone (has called to report that the alarm is setting off)
  - M for Mary (has called to report that the alarm is setting off)



- B, and E are usually given and we ask for J and M
  - B and E is called Evidence
  - J and M are the Query we usually asked bout
  - A is called hidden (Not given or Query)
- The Answer to find J and M is a joint probability distribution over the query variables
  - P(Q1,Q2, …..|E1, =e1, E2=e2)
  - Is called a posterior given the evidence.
- Another query which Q value are max given the evidence
  - max P(Q1=q1, Q2=q2, …….| E1=e1, ….)
- The direction is not restricted, J and M might be the evidence and B, E are the query.

Inference terminologies cont.

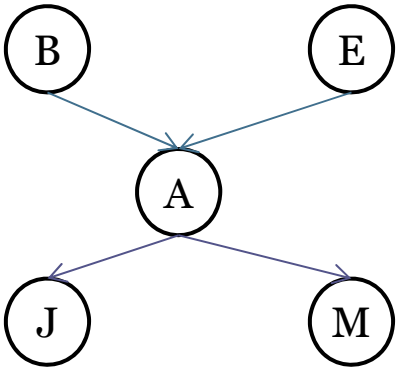
- Suppose Mary called to report that the alarm is off, and we want to know weather there is a burglary. What are the Evidence, Query, and Hidden nodes??
  - Evidence: M, A
  - Query: B
  - Hidden: J, E



Enumeration

First method for probabilistic inference

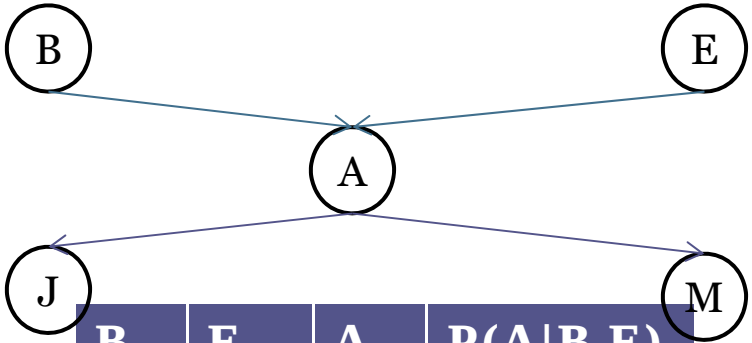
Enumeration



- Enumeration is to calculate all conditional probability.
- P(+b|+j,+m) = ??
  - $= \frac{P(+b,+j,+m)}{P(+j,+m)}$
  - P(+b,+j,+m) =P(+b).P(+j|A).P(+m|A) .P(A|+b,E) .P(E)
  - = P(E).P(+b).P(+a|+b,E).P(+j|+a).P(+m|+a) + P(E).P(+b).P(- a|+b,E).P(+j|- a).P(+m|- a)
  - = P(+e).P(+b) .P(+a|+b,+e).P(+j|+a).P(+m|+ a) + P(+e).P(+b) .P(- a|+b,+e).P(+j|- a).P(+m|- a) + P(- e).P(+b) .P(+a|+b,- e).P(+j|+a).P(+m|+a) + P(- e).P(+b) .P(- a|+b,- e).P(+j|- a).P(+m|- a)
- P(+b,+j,+m) = $\sum_e \sum_a P(+b, +j, +m, A, E)$
  - This is called enumeration of the hidden variables.

Enumeration cont.

B	P(B)
+b	0.001
- b	0.999



E	P(E)
+e	0.002
- e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	- j	0.1
- a	+j	0.05
- a	- j	0.95

B	E	A	P(A|B,E)
+b	+m	+a	0.95
+b	+m	- a	0.05
+b	- m	+a	0.94
+b	- m	- a	0.06
- b	+m	+a	0.29
- b	+m	- a	0.71
- b	- m	+a	0.001
- b	- m	- a	0.999

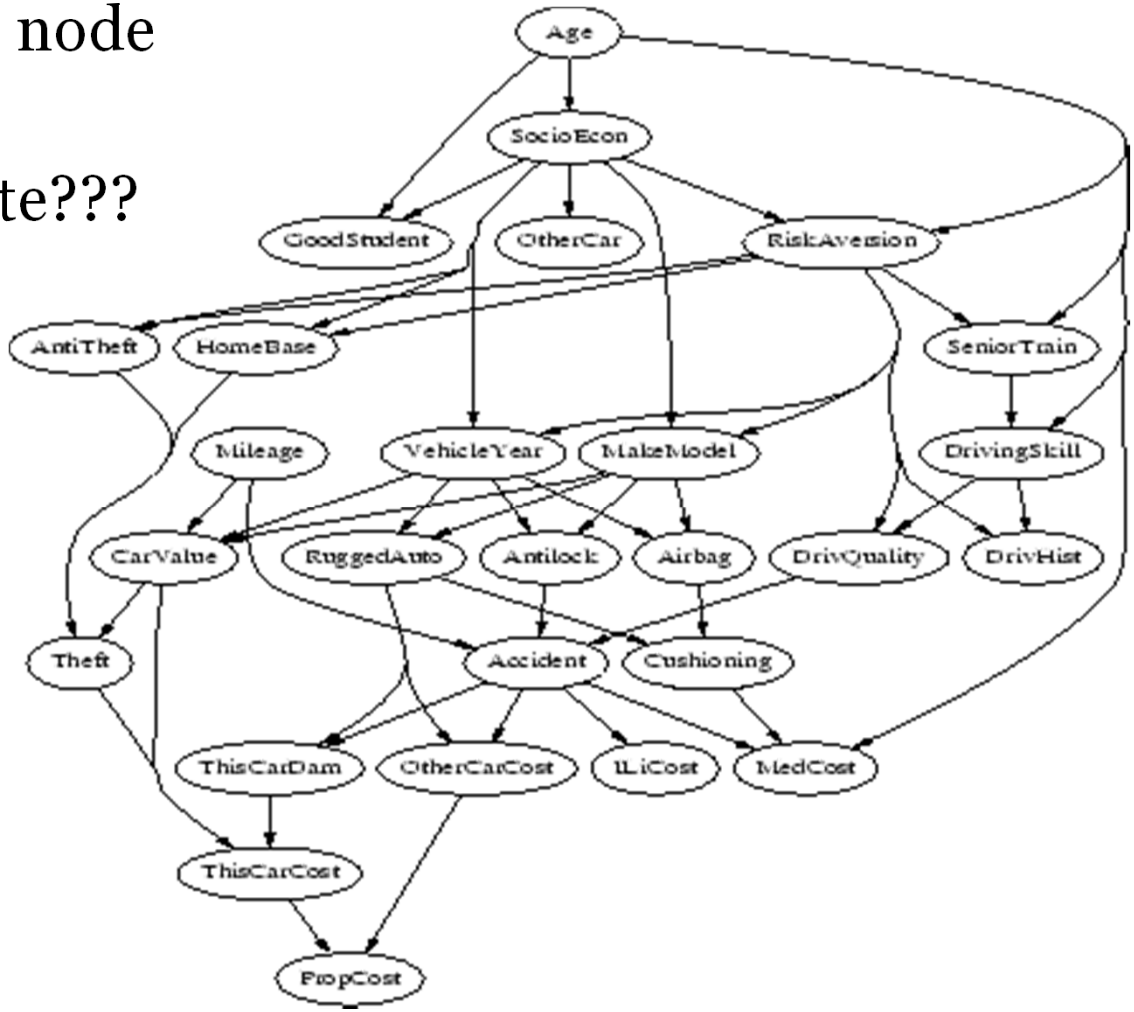
A	M	P(M|A)
+a	+m	0.7
+a	- m	0.3
- a	+m	0.01
- a	- m	0.99

- P(+e).P(+b).P(+a|+b,+e).P(+j|+a).P(+m|+a) = 0.002 . 0.001 . 0.95 . 0.9 . 0.7
- = 0.000001197 remember this only one possibility where (+a,+e)

Enumeration difficulty

- Insurance for car owners problem
- It has 27 Boolean node

- How to Enumerate???



Speeding up Enumeration

First method for probabilistic inference

Bulling out terms

- $\sum_e \sum_a P(E) \,.\, P(+b) \,.\, P(A|+b,E) \,.\, P(+j|A) \,.\, P(+m|A)$

- $= \sum_e P(E) \sum_a P(+b) \,.\, P(A|+b,E) \,.\, P(+j|A) \,.\, P(+m|A)$

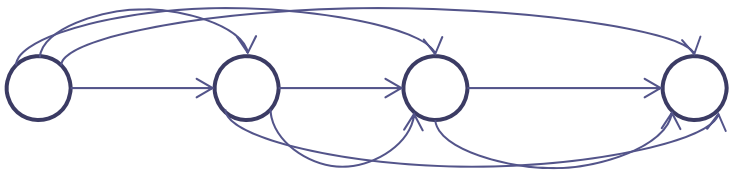
- $= P(+b) \sum_e P(E) \sum_a P(A|+b,E) \,.\, P(+j|A) \,.\, P(+m|A)$

- We succeed to reduce the calculations but we still have the number of rows in the table !!!

Maximize Independence

Enumerated terms

O(n)

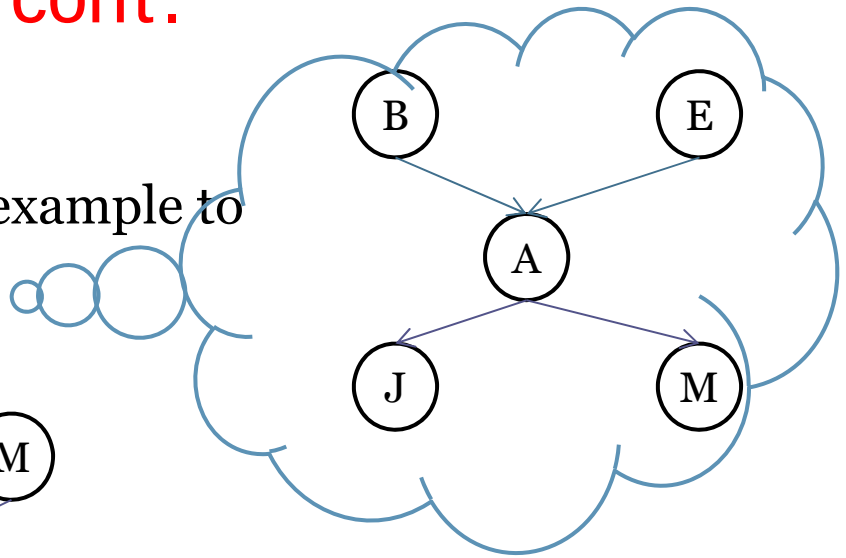


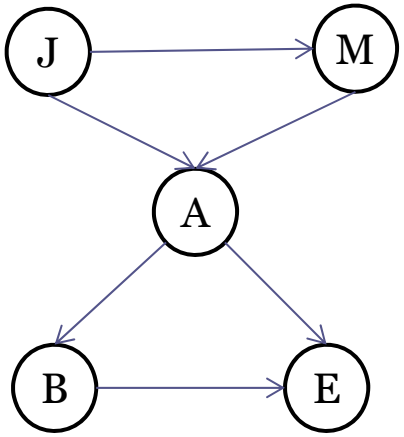
Enumerated terms

$O(2^n)$

Maximize independence cont.

- So we can reordering the earthquake example to maximize independence





- Thus, it's easier to do inference when they're written in the causal direction (network flow from causes to effects).

Variable elimination

R	P(R)
+r	0.1
- r	0.9

R	T	P(T|R)
+r	+t	0.8
+r	- t	0.2
- r	+t	0.1
- r	- t	0.9

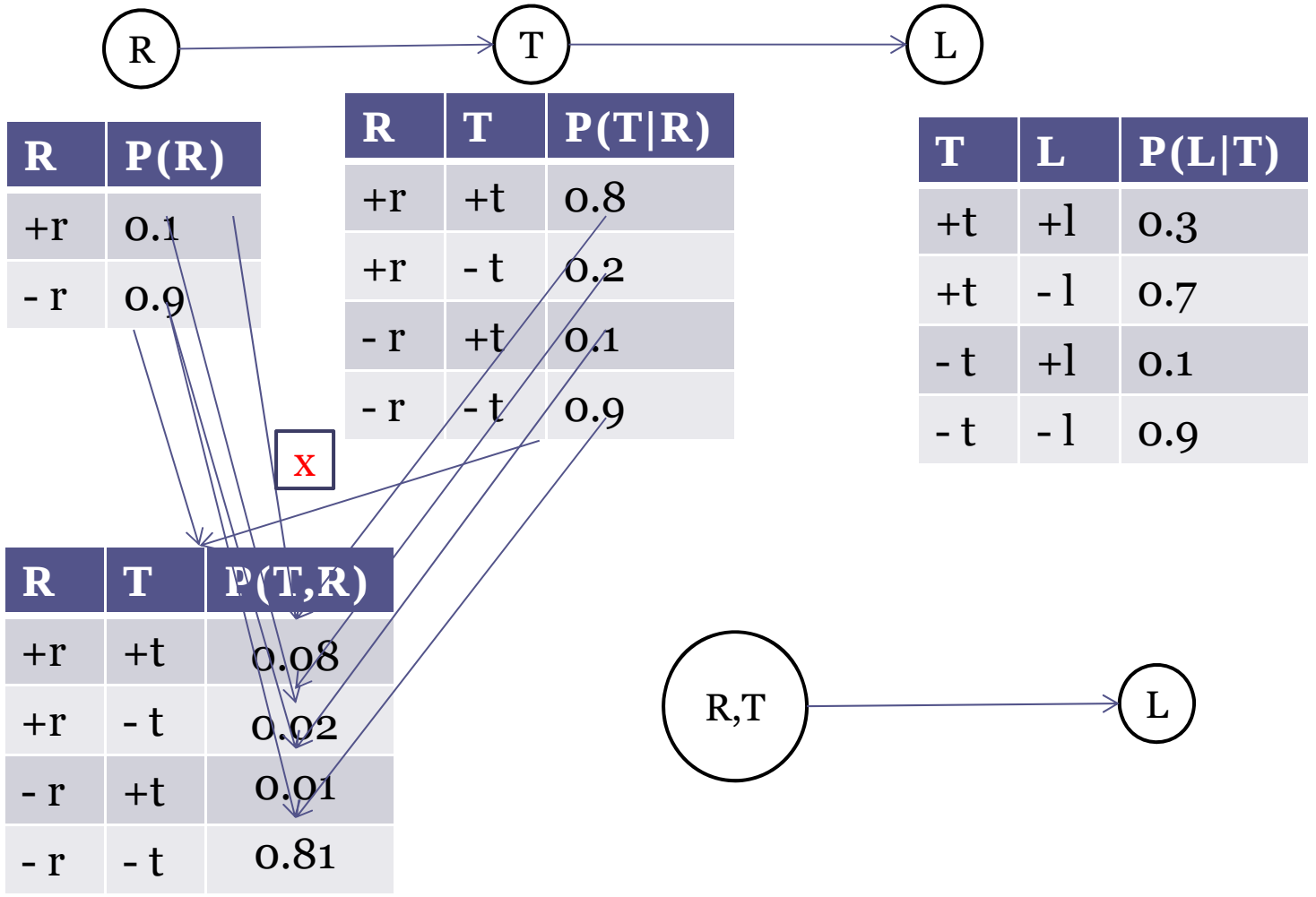
T	L	P(L|T)
+t	+l	0.3
+t	- l	0.7
- t	+l	0.1
- t	- l	0.9

- Am I going to be late P(+l) =??
  - $= \sum_r \sum_t P(R) \,.\, P(T|R).\, P(+l|T)$ With enumeration technique
  - How ever this going to be NB-complete for other examples.

- Variable elimination is a practical solution by reducing the number of nodes within the Bayes network by:
  - Joining factors
  - Node elimination (summing out or marginalization)

Variable elimination cont.1

- Joining factors (Multiply)



R	P(R)
+r	0.1
- r	0.9

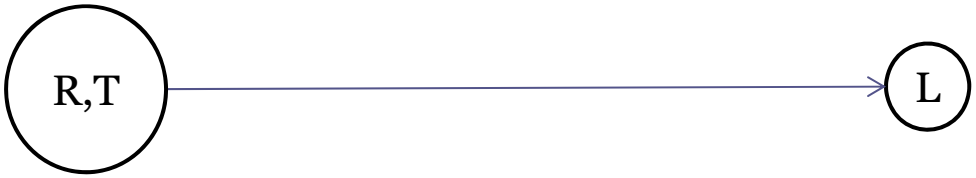
R	T	P(T|R)
+r	+t	0.8
+r	- t	0.2
- r	+t	0.1
- r	- t	0.9

T	L	P(L|T)
+t	+l	0.3
+t	- l	0.7
- t	+l	0.1
- t	- l	0.9

R	T	P(T,R)
+r	+t	0.08
+r	- t	0.02
- r	+t	0.01
- r	- t	0.81

Variable elimination cont.2

- Elimination (Summing out or marginalization)



R	T	P(R,T)
+r	+t	0.08
+r	- t	0.02
- r	+t	0.09
- r	- t	0.81

T	L	P(L|T)
+t	+l	0.3
+t	- l	0.7
- t	+l	0.1
- t	- l	0.9

+
+

T	P(T)
+t	0.17
- t	0.83

Variable elimination cont.3

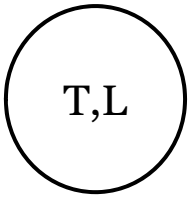
T	P(T)
+t	0.17
- t	0.83

T	L	P(L|T)
+t	+l	0.3
+t	- l	0.7
- t	+l	0.1
- t	- l	0.9

x

T	L	P(T,L)
+t	+l	0.051
+t	- l	0.119
- t	+l	0.083
- t	- l	0.747

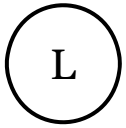
Joining



+

L	P(L)
+l	0.134
- l	0.866

Eliminate



Variable elimination cont.4

- Variable elimination is a continued process of joining together factors to form a large factor, and then eliminating variables by summing out.

- The problem is to make a good choice of the order in which we apply these operations.

Approximate Inference

Second method for probabilistic inference

Approximate Inference

- Instead of using probability rules we can simulate the experiment and count the samples to calculate approximate probability.

1 £	0.5£	Count
H	H	𝍸 //
H	T	𝍸 ///
T	H	𝍸 //
T	T	𝍸 /

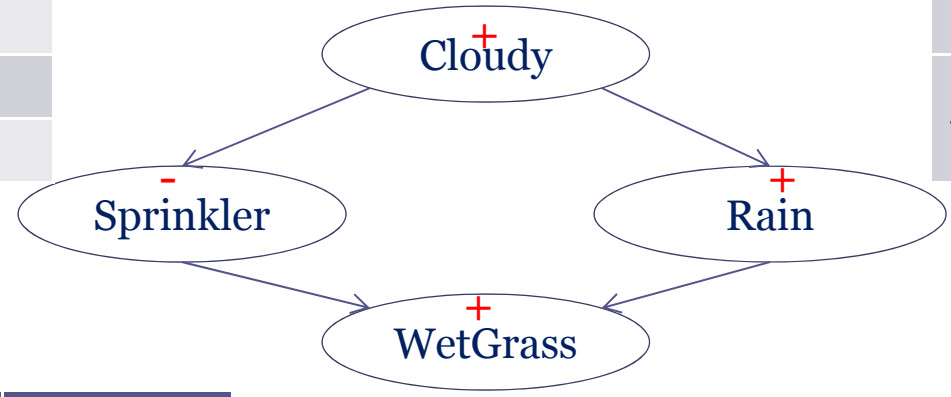
- Disadvantage
  - Accurateness depends the number of samples.
- Advantage
  - Avoid rules and enumeration complexity.
  - Still work even if conditional probability tables are unknown.

Sampling

C	P(C)
+c	0.5
- c	0.5

C	S	P(S|C)
+c	+s	0.1
	- s	0.9
- c	+s	0.5
	- s	0.5

C	R	P(R|C)
+c	+r	0.8
	- r	0.2
- c	+r	0.2
	- r	0.8



S	R	W	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

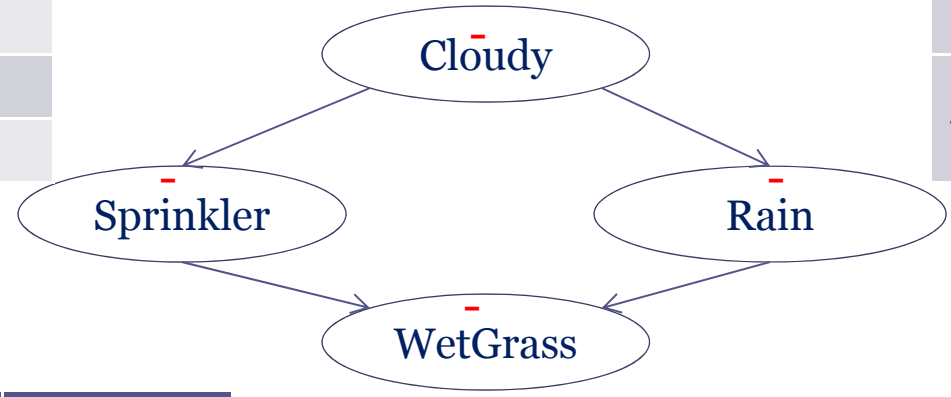
- Samples (with random generator):
  - 1st sample : +c, -s, +r, +w
  - repeat this again and count similar samples
- If we have infinite number of samples then this method is consistent and we can calculate:
  - P(C,S,R,W)
  - P(any individual variable) .e.g. P(W)
- How to compute probability of conditional variable. E.g. P(W|-c) ?

Rejection sampling

C	P(C)
+c	0.5
- c	0.5

C	S	P(S|C)
+c	+s	0.1
	- s	0.9
- c	+s	0.5
	- s	0.5

C	R	P(R|C)
+c	+r	0.8
	- r	0.2
- c	+r	0.2
	- r	0.8



S	R	W	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

- How to compute probability of conditional variable. E.g. P(W|-c) ?
  - In this case we interested in –c only.
- Rejection sampling:
  - Is a technique to reject samples not related to the required conditional probability. E.g (+w,+c)
  - This method still consistent

Rejection sampling cont.

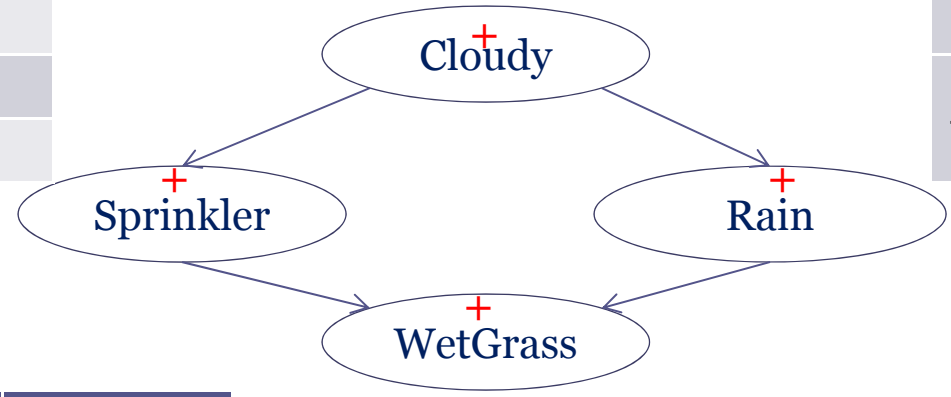
- Back to earthquake example, suppose we need P(B|+a).
  - If we apply sampling we would receive a lot of P(-b | -a) and reject it. Because in this example Burglary is less likely.
- To avoid rejected samples at all, We fix A to be +a at sampling time so when performing sampling we only receive. (+b, +a) , (-b, +a).
  - However the result set of samples is inconsistent.
- We can solve the inconsistency problem by assigning a probability to each sample and weight them correctly.

Likelihood weighting

C	P(C)
+c	0.5
- c	0.5

C	S	P(S|C)
+c	+s	0.1
	- s	0.9
- c	+s	0.5
	- s	0.5

C	R	P(R|C)
+c	+r	0.8
	- r	0.2
- c	+r	0.2
	- r	0.8



S	R	W	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

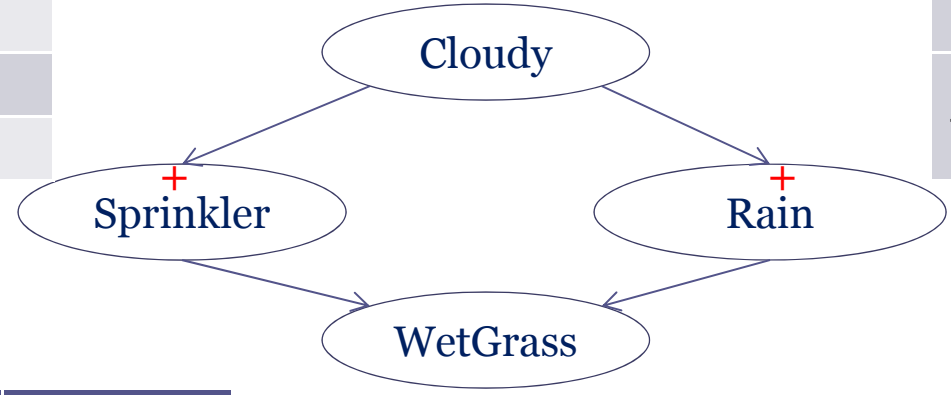
- P(R|+s, +w)?
  - Do samples like before but add weights
  - +c, +s (0.1), +r, +w(0.99)
- To calculate the weight for this sample we multiply the weights of the variables.
  - (+c,+s,+r,+w) has a weight 0.1*0.99=0.099
- Now when we count the matched sample to calculate the probability, instead of count (+c,+s,+r,+w) as 1 sample, we count it as 0.099

Likelihood weighting cont.

C	P(C)
+c	0.5
- c	0.5

C	S	P(S|C)
+c	+s	0.1
	- s	0.9
- c	+s	0.5
	- s	0.5

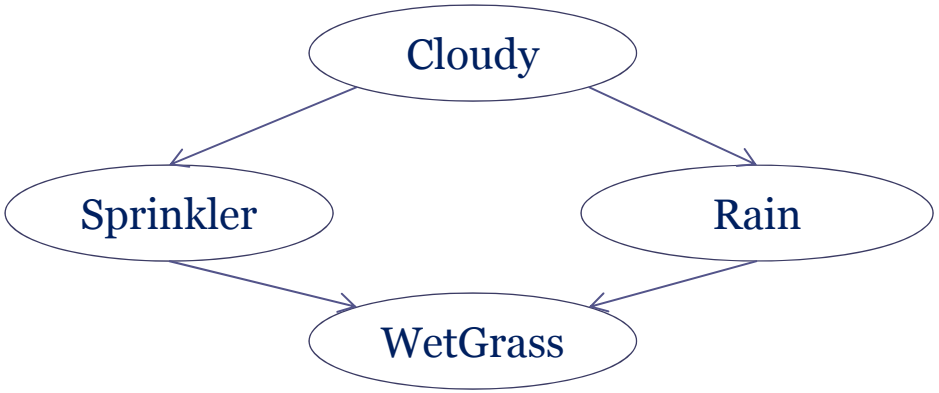
C	R	P(R|C)
+c	+r	0.8
	- r	0.2
- c	+r	0.2
	- r	0.8



S	R	W	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

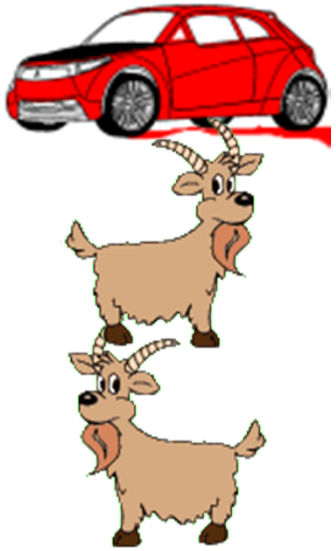
- What about P(C|+s, +r)
  - The generator in this case will generate +w most of the time which is not a good evidence for the Cloudy variable
- So likelihood weighting is not adequate for some conditional probability.

Gibbs sampling



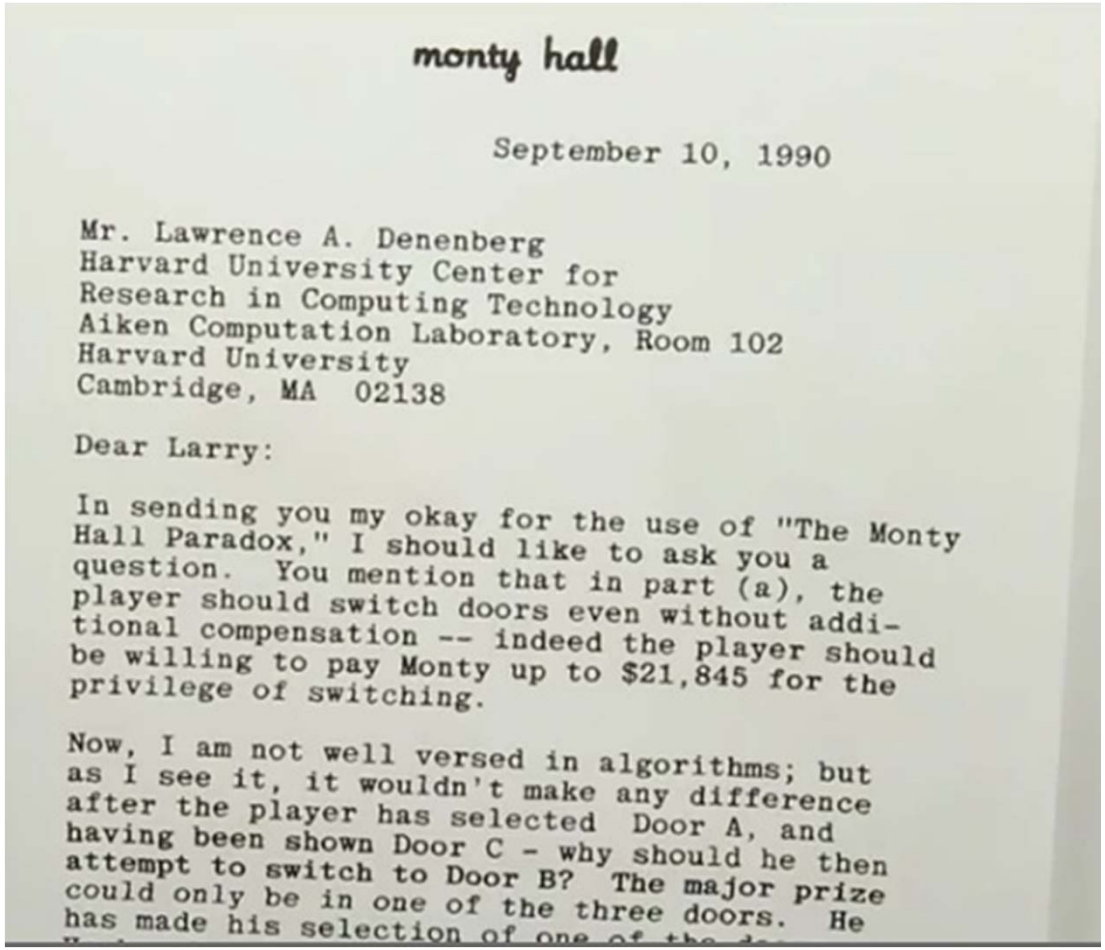
- P(R|+s, +w)?
- Instead of sampling all non-evidence variables (C, R in this example) at the same time, we resample with respect to only variable at a time conditioned on all the others.
  - We start with initial state +c, +s, +r, +w keeping +s, +r (evidence) fixed
  - We select C to resample so we could have
    - –c, +s, +r, +w (we could have +c, +s, +r, +w again)
  - Then we select W to resample last output so
    - -c, +s, +r, -w
  - This method called Markove Chain Monte Carlo (MCMC).
- In this way we take all he evidence into account not just the upstream evidence
- In contrast to Rejection and Likelihood techniques, each sample here is dependent on the other (one change between each sample).
- This technique still consistent.

Monty Hall problem



- You select a door (still closed)
- The host (Monty Hall) open 1 of the other doors
- Then you have the choice to stick with your first choice or switch to the reset door.
  - Probability of your door 1/3
    - At your choice time it was 1/3, and opining one door will not change this fact.
  - Probability of the reset door 2/3
    - It was 1/3 but when the host open one door, absolutely he won't open the door with the car because he knows in advance where it's which increase the probability of the reset door.

Monty Hall problem cont.



monty hall

September 10, 1990

Mr. Lawrence A. Denenberg
Harvard University Center for
Research in Computing Technology
Aiken Computation Laboratory, Room 102
Harvard University
Cambridge, MA 02138

Dear Larry:

In sending you my okay for the use of "The Monty Hall Paradox," I should like to ask you a question. You mention that in part (a), the player should switch doors even without additional compensation -- indeed the player should be willing to pay Monty up to $21,845 for the privilege of switching.

Now, I am not well versed in algorithms; but as I see it, it wouldn't make any difference after the player has selected Door A, and having been shown Door C - why should he then attempt to switch to Door B? The major prize could only be in one of the three doors. He has made his selection of one of the