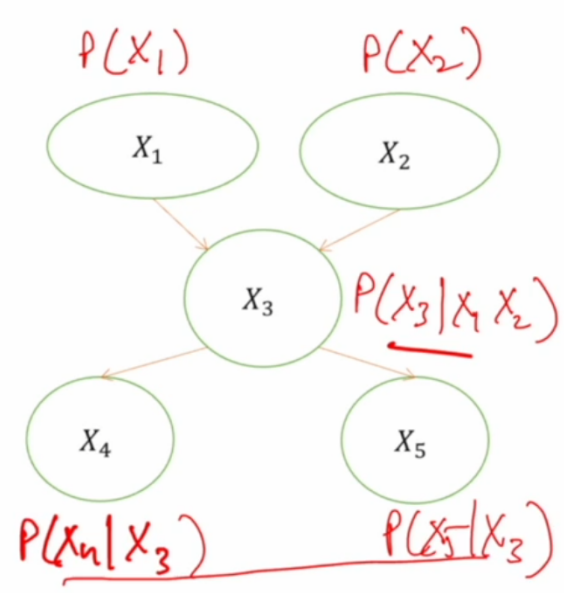
Lecture 11 - Bayesian Networks

Bayesian Network: Directed acyclic graph.

- Each node corresponds to a random variable
- Arrows connect pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y.
- Each node X_i has associated probability information $P(X_i|\,\text{parents}(X_i))$ that quantifies the effect of the parent on the node.

Visually:



What would be the joint probability of this network?

$$P(X_1, X_2, \ldots, X_5) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_3)P(X_5|X_3)$$

This is very similar to the Naive Bayes approach. What about the posterior?

$$P(Y\,|X_1X_2) = \alpha(\text{joint})$$
$$P(X_1|X_4, X_5) = \alpha(\text{joint})$$

To generalize:

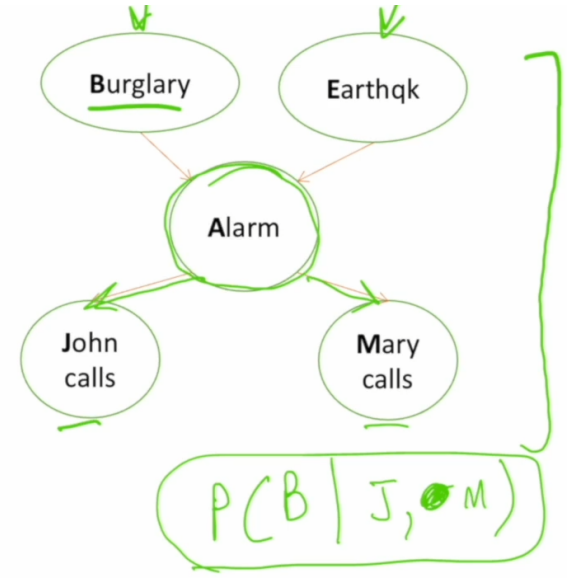
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\,\text{parent}(X_i))$$

Let's write the full joint distribution using the chain rule:

$$\begin{aligned} &P(X_1, X_2, \ldots, X_5) \\ =&P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)P(X_5|X_1, X_2, X_3, X_4) \end{aligned}$$

Let's make an assumption: A variable is independent of its non-descendents, given the parent. How can we simplify the above then? X_4 only has one parent, so we can remove X_1 and X_2 from the term. $P(X_4|\,X_3)$. So on and so forth... that's essentially how we get the above simpler version. We need to use the conditional independence assumption.

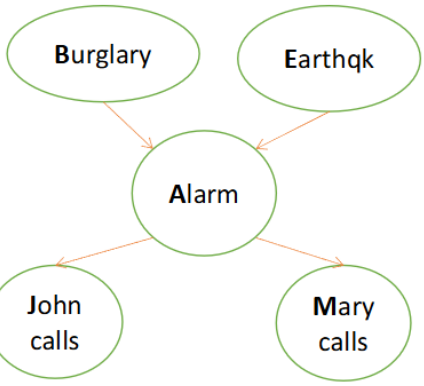
Let's look at an example:



What is the probability that there is a burglary given that both John and Mary call? i.e. $P(B|J, M)$. We need to know the following:

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

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

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

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

Consider the example in the slides: $P(+b, -e, +a, -j, +m)$.

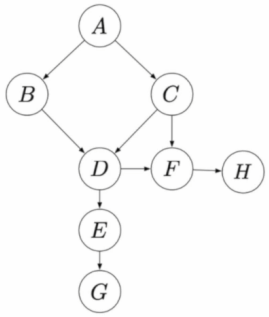
$$\begin{aligned} &= P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) \\ &= 0.001 \times 0.998 \times 0.940 \times 0.1 \times 0.7 \end{aligned}$$

If we were to use the full joint probability, it would contain $2^5 = 32$ values. How many do we need here? One for each of $P(B)$ and $P(E)$, two for John and Mary each, and 4 more for the Alarm. So 10 in total.

In general, let's say we have n random variables, each of which has k parents. Then, we would have $n2^k$ values in total. That's the space requirement. This is much better than 2^n for $k < n$.

We discussed the conditional independence. The nodes are independent of the non-descendents given the parent. If the network is complex, we need to look at an algo. to determine which nodes / variables are independent in a Bayes Net.

Let's start with an example.



$B \perp\!\!\!\perp C$	Guaranteed true	Guaranteed false	Cannot be determined
$B \perp\!\!\!\perp C \mid G$	Guaranteed true	Guaranteed false	Cannot be determined
$B \perp\!\!\!\perp C \mid H$	Guaranteed true	Guaranteed false	Cannot be determined
$A \perp\!\!\!\perp D \mid G$	Guaranteed true	Guaranteed false	Cannot be determined
$A \perp\!\!\!\perp D \mid H$	Guaranteed true	Guaranteed false	Cannot be determined
$B \perp\!\!\!\perp C \mid A, F$	Guaranteed true	Guaranteed false	Cannot be determined
$F \perp\!\!\!\perp B \mid D, A$	Guaranteed true	Guaranteed false	Cannot be determined
$F \perp\!\!\!\perp B \mid D, C$	Guaranteed true	Guaranteed false	Cannot be determined

B is indepedent of C or not, B is independent of C given G, etc… Let's look at $F \perp B \mid D, A$ (F is independent of B given D and A. In this network, is that true or not?

The algorithm D separation:

1. Draw the ancestor graph of all variables mentioned in the probability expression, consisting only of the vars mentioned and their ancestors.

   Here, we draw C as well, because C is an ancestor of F. We have to draw all of its edges as well.

2. Moralize: For each pair of variables with a common child, draw an undirected line (edge) between them. B and C have the same child, D. So we will draw a line between them. But also D and C have the same child, F.

3. Disorient the graph: remove the arrows and replace them with edges (no arrows).

4. Delete the givens and their edges.

5. If the variables are disconnected in this new graph, then they are guaranteed to be independent. Otherwise, Bayes net does not require conditional independence.

In other case, we cannot say Guaranteed true. We also cannot say Guaranteed false. D separation it can only guarantee a conditional independence assumption as true but not as false. Therefore, we have to say that it cannot be determined. If there is a path, then we say cannot be determined. If there is no path, then we say guaranteed true. The middle column (guaranteed false) should not be selected in an exam.

Let's repeat this process by hand later during exam study to make sure that we understood everything.

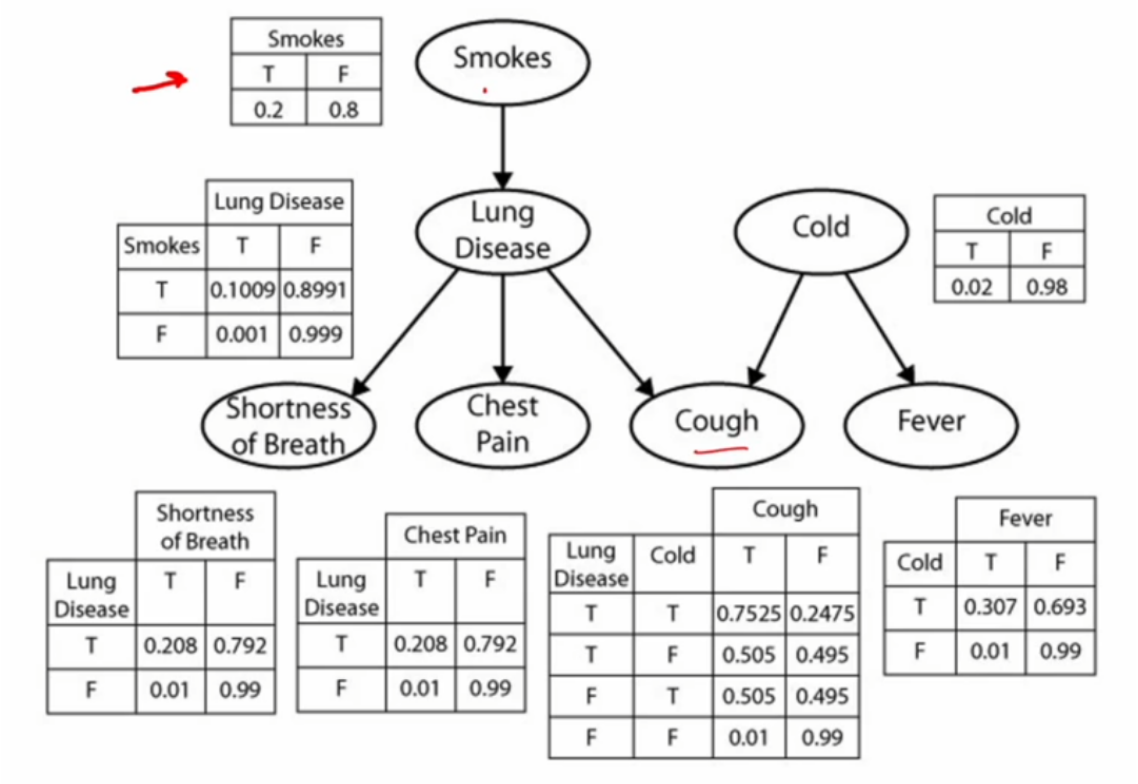
Inference by Enumeration: Let's look back at the same burglary case. We want:

$$P(B|J,M) = \alpha P(B,J,M)$$

We have two missing variables, so we sum over those two.

Time complexity of inference by enumeration is $O(2^n)$, where n is the number of non-evidence variables. What does this mean? It takes a lot of time. In the above, two variables are missing, so $2^2 = 4$ configurations that we need to look into.

Variable Elimination Algorithm: Gets rid of the repeated calculations. Let's look at an example:



Smokes	
T	F
0.2	0.8

Smokes	Lung Disease	
	T	F
T	0.1009	0.8991
F	0.001	0.999

Cold	
T	F
0.02	0.98

Lung Disease	Shortness of Breath	
	T	F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Chest Pain	
	T	F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Cold	Cough	
		T	F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

Cold	Fever	
	T	F
T	0.307	0.693
F	0.01	0.99

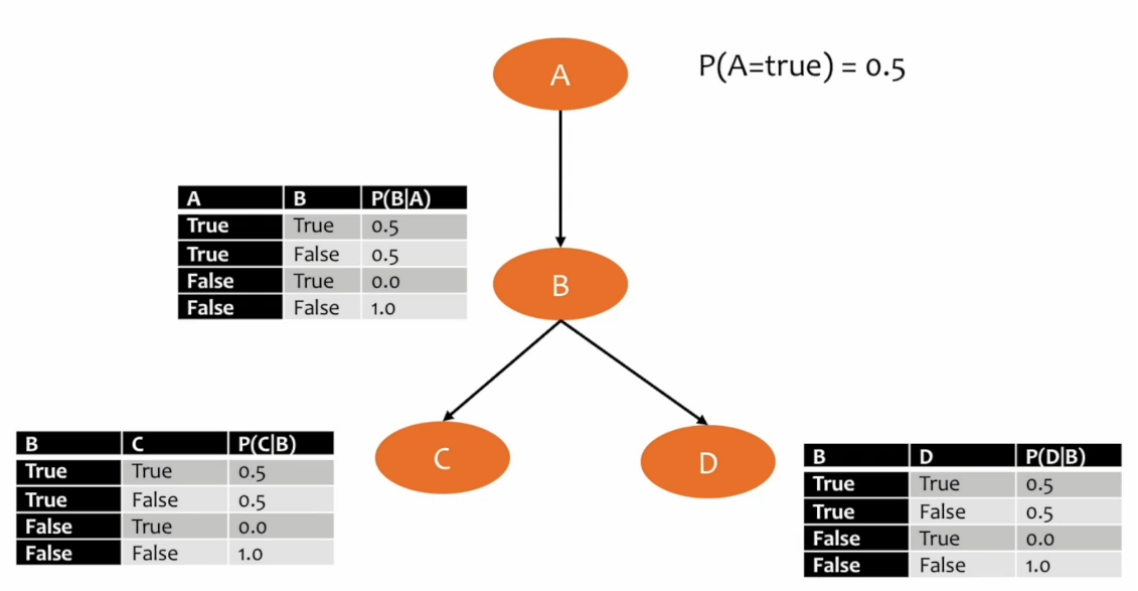
Given this network, let's try to answer some questions. How can you write the full joint probability using local conditional probabilities?

$$P(\text{all the variables}) =$$
$$P(S)P(\text{Cold})P(L|S)P(\text{Short}|L)P(\text{Ch}|L)P(\text{Cough}|L,\text{Cold})P(F|\text{Cold})$$

This is the full joint. Now, compute the following:

$$P(\text{smokes} = T, \text{Lung} = T, \text{Cold} = T, \text{Short} = T, \text{Chest} = T, \text{Cough} = T, \text{Fever} = T)$$

Just plug it in based on the table that we have. Usually, we are not interested in this. We are interested in the posterior. $P(\text{LungDisease} = T \mid \text{Cough} = T, \text{Fever} = F)$. Let's look at another example:



A	B	P(B|A)
True	True	0.5
True	False	0.5
False	True	0.0
False	False	1.0

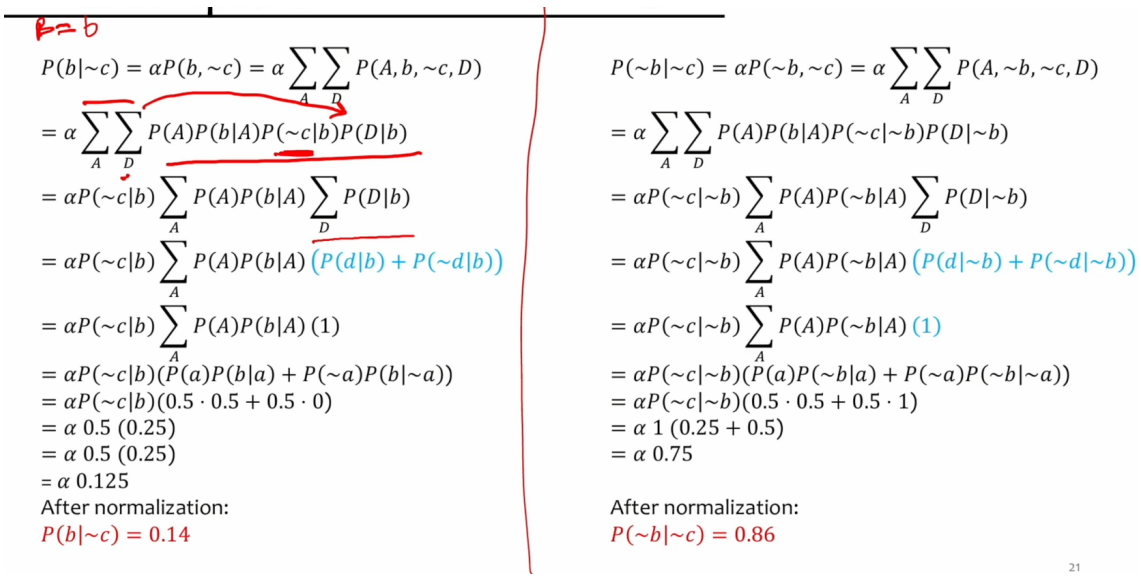
B	C	P(C|B)
True	True	0.5
True	False	0.5
False	True	0.0
False	False	1.0

B	D	P(D|B)
True	True	0.5
True	False	0.5
False	True	0.0
False	False	1.0

We need to find $P(B \mid \sim C)$.

$$\begin{aligned} P(B \mid \sim C) &= \alpha P(B, \sim C) \\ &= \alpha \sum_A \sum_D P(B, \sim C, A, D) \\ &= \alpha \sum_A \sum_D P(A) P(B \mid A) P(\sim C \mid B) P(D \mid B) \end{aligned}$$

Now, we need to move the summation signs, and some other stuff happens. See below:



Basically, after doing all this, we see that we can just kind of ignore the missing variables. In this case, D is not there, so it's useless. But A is an ancestor, so we need it.

Approximate Inference: Draw N samples from a sampling distribution S. Compute and approximate posterior prbability $\hat{P}$. Show convergence to the true probability.

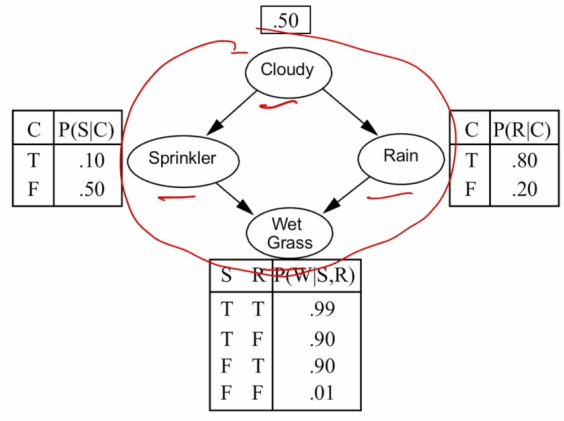
Suppose we can generate random numbers from a uniform distribution over the interval $[0, 1]$. How can we transform this to sample from any desired distribution? For example, sampling a coin flip. We can threshold, if it's between 0 and 0.5, heads, then 0.5 and 1 is tails. It's just thresholding between $[0, 1]$. That's how we use the uniform random distribution.

Sampling from Bayes Network:

- Consider a Bayes Net that has no evidence associated with it
- Sample each variable in topological order
- Sample from the variables' conditional probability distribution
- Distribution is conditioned on the values already assigned to the variable's parents
- Parents are guaranteed to have values due to topological ordering.

What does any of this mean?

At the beginning, none of the random variables are evidence. Otherwise, it wouldn't have any probability associated with it.



So in the above, at first all the different things are random. We need to sample each one of them in topological order (Cloudy $\longrightarrow$ Sprinkler $\longrightarrow$ Rain $\longrightarrow$ Wet Grass).

Sample from the variables' conditional probability distribution: If we are sampling from Sprinkler, we will use the table:

$$\begin{bmatrix} C & P(S|C) \\ T & 0.1 \\ F & 0.5 \end{bmatrix}$$

It depends what the sample for Cloudy was. That's why we need to sample each variable in topological order. Let's do all of it. Start with Cloudy:

- Sample from $P(\text{Cloudy}) = [0.5, 0.5]$, value is true. For Sprinkler and Rain, we use 0.1 and 0.8.
- Sample from $P(\text{Sprinkler}|\,\text{Cloudy} = \text{True}) = [0.1, 0.9]$, value is false.

- Sample from $P(\text{Rain}\,|\text{Cloudy} = \text{True}) = [0.8, 0.2]$, value is true
- Sample from $P(\text{Wet Grass}|\,\text{Sprinkler} = \text{False}, \text{Rain} = \text{True}) = [0.9, 0.1]$, value is true

$$P(C, \sim S, R, W)$$

Now, we have apparently sampled. We will need to look at multiple samples? Let's say the question is what is $P(\text{Cloudy} = \text{true})$? How do we use the samples here? Count all the samples. Let's say 100. 45 true, 55 not true, then $P(\text{Cloudy} = \text{true}) = 0.45$. Literally just counting. How do we know that this will give us the right probability?

$$S_{PS}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\,\text{parents}(X_i)) = P(x_1, x_2, \ldots, x_n)$$

S_{PS} is the probability of choosing the event $(x_1, x_2, \ldots, x_n)$. In the prior example, $S_{PS}(C, \sim S, R, W) = 0.5 \times 0.9 \times 0.8 \times 0.9$. This is a different expression to $P(C, \sim S, R, W)$, but we would get the same values. Using the sampling procedure is the same as the joint probability for the Bayes network.

Rejection Sampling: Used to compute the conditional probabilities $P(X\,|e)$. Generates samples from the prior distribution using the prior sample.

- Reject all samples that do not match the evidence
- Estimate $P(X = x|e)$ y counting how often $X = x$ occurs in the remaining samples.

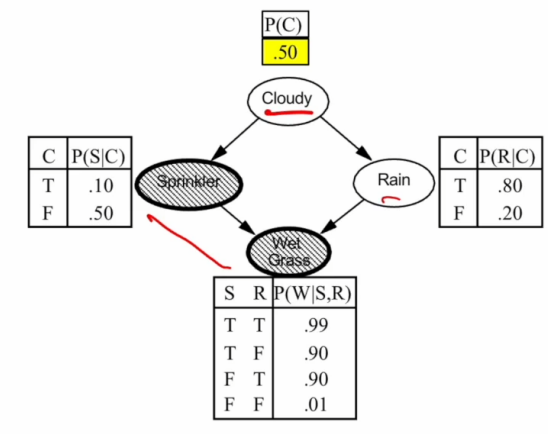
For eaxmple, estimate $P(\text{Rain}|\,\text{Sprinkler} = \text{True})$ using 100 samples, 27 samples have Sprinkler = True. Of these, 8 have Rain = true and 19 have Rain = false.

$$P(\text{Rain}|\,\text{Sprinkler} = \text{True}) = \frac{8}{27} \text{ and } \frac{19}{27}$$

The only problem is that we reject a lot of samples. Especially if we have multiple evidence. The next method is:

Likelihood Weighting:

Idea: Fix the evidence variables, sample only non-evidence variables and weight each sample by the likelihood it accords in the evidence. Example:



Let's say Sprinkler and Wet Grass are evidence. So we only need to sample Cloudy and Rain. Start with Cloudy topologically. We will not use Sprinkler, but we will use the probability for the weighing.

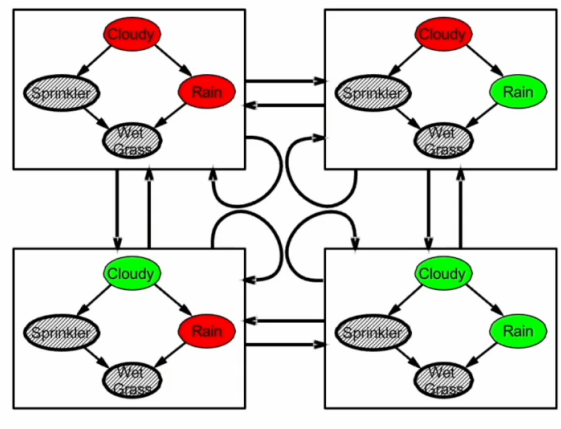
Cloudy = True $\longrightarrow P(S|C) = 0.1$, $P(R|C) = 0.8$. Let's say Rain is true, then go to Wet Grass. Both true so 0.99.

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

What is he saying here? No clue. Look at problems later.

Inference by Markov Chain Monte Carlo: We do not start from scratch. We start from an arbitrary state with evidence variables fixed. The state is the current assignment of all variables.

- Generate next state by sampling one variable given the Markov Blanket
- Sample each variable in turn, keeping evidence fixed.
- Repeat



There are only 4 possible things in this chain, because we have two non-evidence only.

Theorem: Chain approaches stationary distribution. Long run fraction of time spent in each state is exactly proportional to its posterior probability.

Markov Blanket: Parents + Children + Children's parents.

For cloudy, the MB is Sprinkler and Rain.

For Rain: Cloudy (Parents), Wet Grass (Children), Sprinkler (Children's parents). Basic stuff I think.

$$P(x_i \,|\, \text{MB}(X_i)) = \alpha P(x_i \,|\, \text{parents}(X_i)) \prod_{Y_j \in \text{Children}(X_i)} P(y_j \,|\, \text{parents}(Y_j))$$

Main computational problems:

1. Difficult to tell if convergence has been achieved.
2. Can be wasteful if Markov blanket is large.