Artificial Intelligence

Prof. P. Dasgupta

Department of Computer Science & Engineering

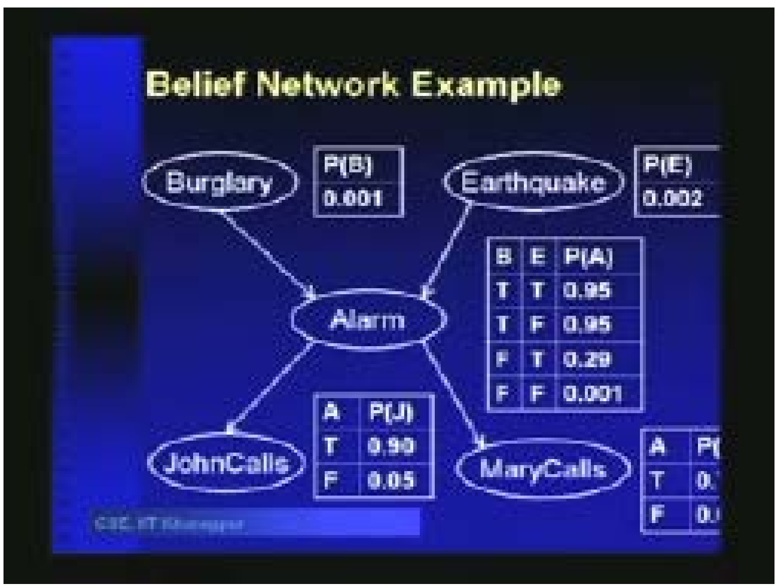
Indian Institute of Technology, Kharagpur

Lecture No - 24

Reasoning with Bayes Networks (Contd.)

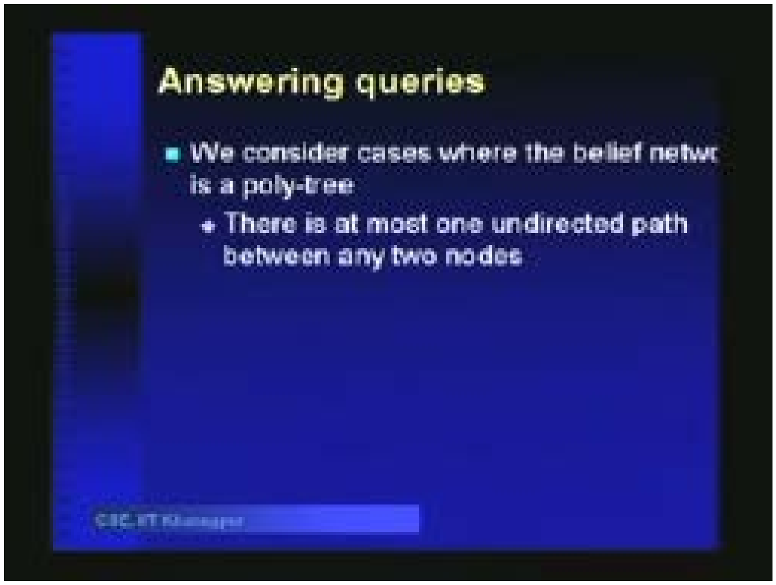
In this lecture, we will continue with our analysis of the Bayes networks.

(Refer Slide Time: 01:00)



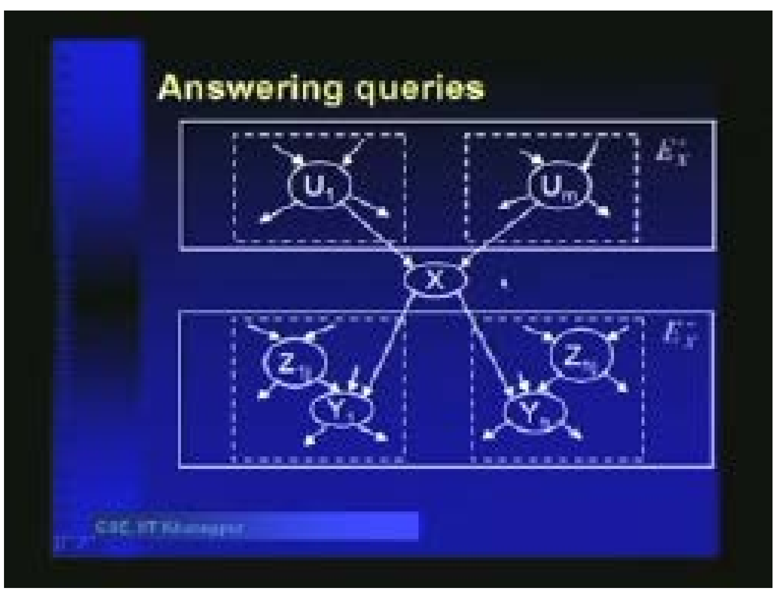
Recall that in the last lecture, we considered belief networks, which are in the form of a poly-tree, which means that between any pair of nodes in the belief network, we have exactly 1 undirected path.

(Refer Slide Time: 01:14)



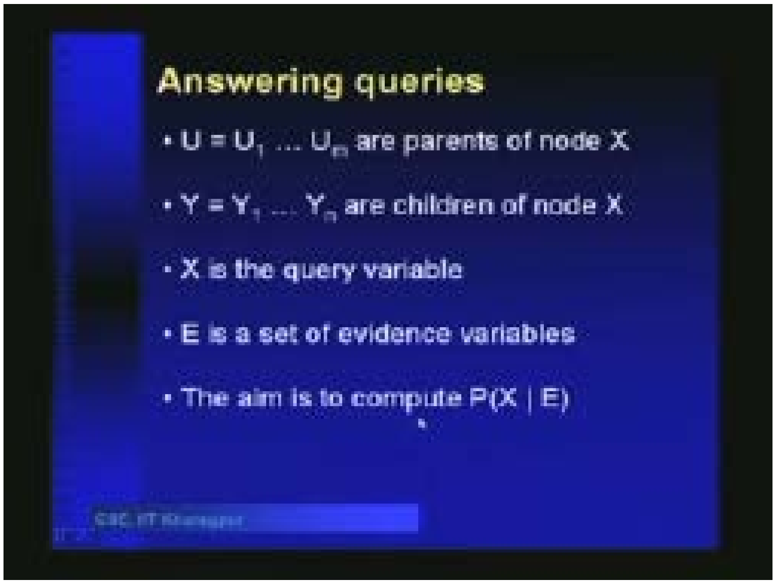
Coming back to the definitions, our objective is to compute the generic query P of X given E, where E is a set of evidence, which means that it is the values of a set of variables in the belief network and our objective is to compute PX given E.

(Refer Slide Time: 01:33)



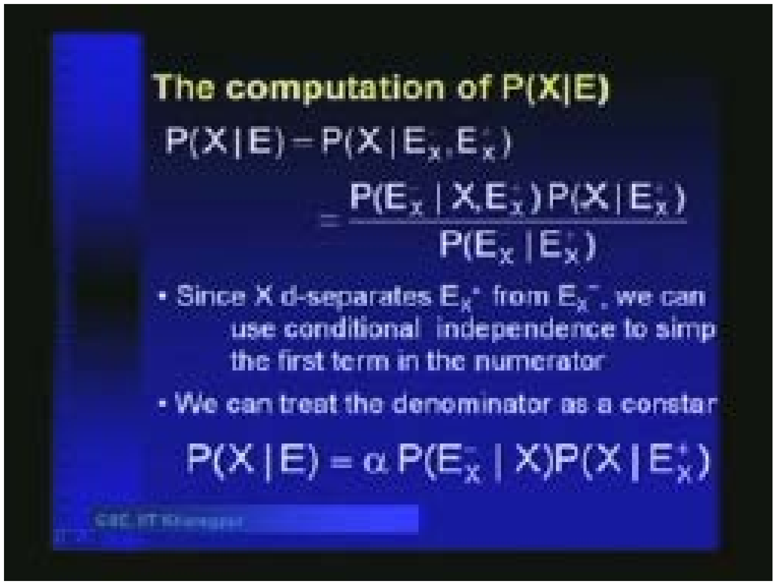
The evidence nodes can be above X in the belief network and can be also below X in the belief network, so, we partition E into 2 parts, namely EX plus and EX minus. EX plus consists of those evidence nodes which are predecessors of X and EX minus consists of those evidence nodes which are descendants of X, so, they may not be immediate parents or children, but they are ancestors or descendants. Then, we had the following definitions: namely, that for the node X, U1 through Um are the parents; Y1 through Yn are the children, X is of course the query node and E is the set of evidence variables, and we want to compute PX given E.

(Refer Slide Time: 02:29)



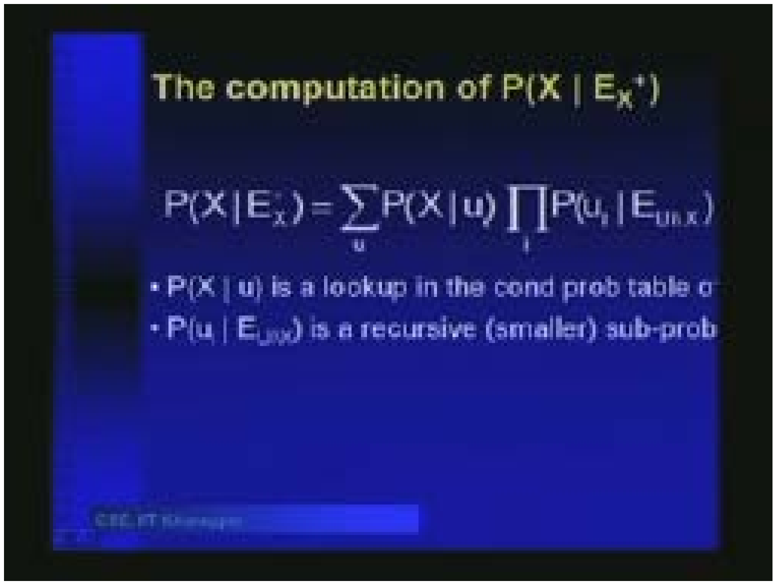
We had initially formulated this as follows: that PX given E is PX given EX minus and EX plus so E is just partition into the 2 sets, EX minus and EX plus, and then, we rewrite this using Bayes rule, as P of EX minus given X EX plus and P of X given EX plus.

(Refer Slide Time: 03:06)



This was written in this way, so that we always have the effects on the left hand side and the causes on the right hand side. We write it in the form in which the Bayes network itself is given and then, we showed that since this is a constant, so, we can replace it by alpha and the first term reduces to P EX minus given X. Because the node X separates EX minus from EX plus and then, again, we have PX given EX plus. We analyze the first term in the last lecture and then, showed that the first term can be represented as sigma over all the parents of X and P of X given U and then, the product of all these independent terms, namely, probability of each UI, given the evidence, connected to the UI except through X.

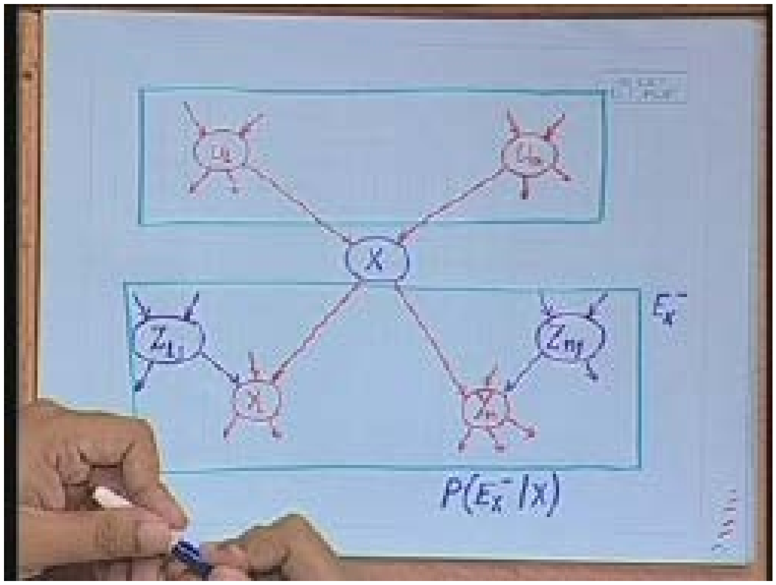
(Refer Slide Time: 04:21)



We had seen that these entries can be found out from the conditional probability table for X and these terms are can be recursively computed, because they are, again, the parents of X and because we are going from X to its parents and from its parents to its parents and so on. This term will recursively take us right up to the first level nodes and in the first level nodes, they will become independent of the evidence, and so, they will have the values from their probability tables directly. So, this part is solved. Now, what we need to do is still, to show that how do we compute probability of X?

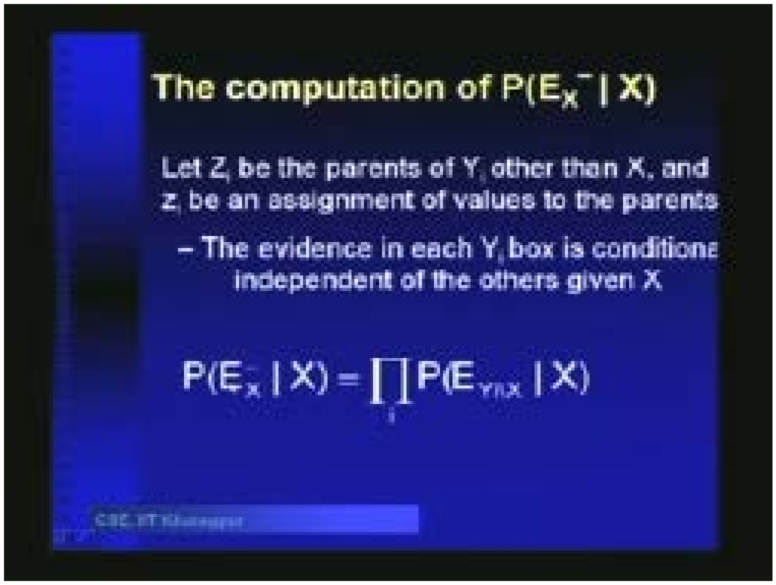
Probability of EX minus given X? This first term out here is something which we have not yet addressed. Today, we will first look at the computation of P of EX minus given X. Now, coming to the picture that we had- text please- let us continue with this. We have- yes, for each YI, we have a parent; 1 or more parents, which we will call, say, for Y1, we will call it Z1 to Z1 j through Z1 k. Then, for Yn, it will be Zn 1, Zn 2, Zn 3 and so on. What we are going to do is, when we- so, this is the set of evidence nodes in this part is what we are calling as EX minus.

(Refer Slide Time: 07:32)



Now, what we are interested right now, is in computing probability of EX minus given X. So, we will try to decompose this term in terms of the probabilities of these children, given the evidence in their predecessors. That is why we are interested in just considering the immediate parents of these Yi nodes- Y1 through YN are the children of X and we are looking just at the immediate parents of these. The evidence in each Yi box is conditionally independent of the others given X. See, if X is given, then, the probability of Y1 is independent of the probability of Y2 and the independent of the probability of Yn, because this node X will d-separate them provided that X is given, so, we can rewrite P EX minus given X as the probability over EYI except X given X and take their product.

(Refer Slide Time: 08:06)

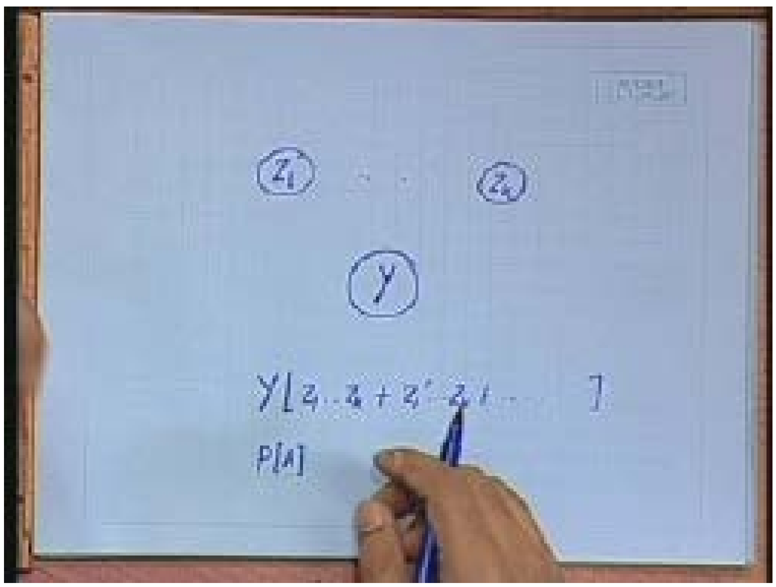


Is that understood? No? Okay, what we are trying to compute here is that the probability of the set of evidence nodes that are below X given X. Each of these EX minus consists of- what it consists of? Text please. It consist of some of the evidence nodes around Y1, some evidence nodes around Y2, some evidence nodes around Y three, and so on. Now, our claim is that the if X is given, then, the set of evidence nodes around Y1- their probability is independent of the set of evidence nodes around Y2 and their probability is independent of the set of evidence nodes around Y three and so on.

Therefore, EX minus can be partitioned into the set of evidence nodes around Y1, around this evidence nodes around Y2, and so on. So, we partition the evidence EX minus like that and that gives us each of this EYI except X. Recall that the definition of EYI except X is the set of evidence nodes connected to Yi except through X. After partitioning that out, because these evidence nodes are independent, so, we can split them up into a product of their probabilities. Then, we average over Yi and Zi. Now, let us understand what is this averaging.

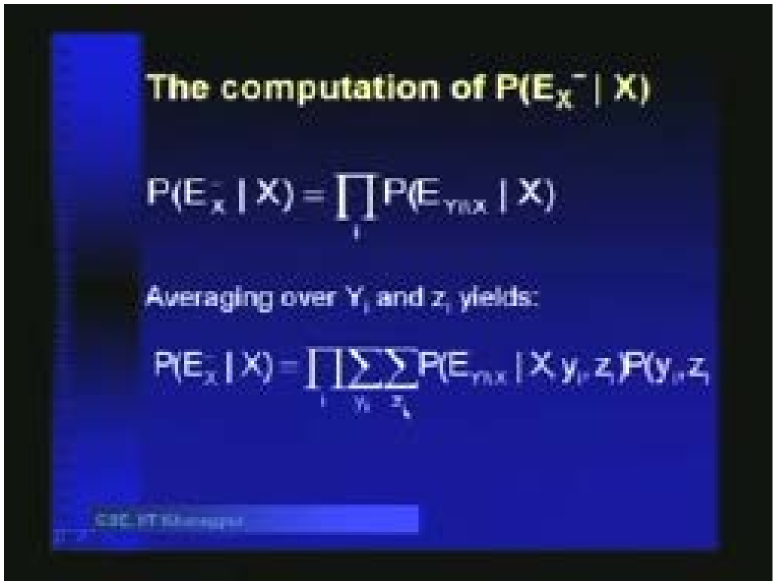
Now, when we look at the probability- text please- when we look at the probability of a particular event here, we can average out over its parents. Remember that the probability of YI- we can write that as probability of Y1 and Z1 one plus probability of Y1 and you know if you look at the parents, let me suppose we have a node Y, and we look at its parents, say Z1 through Zk. We can split up Y as the probability of Y times say Z1, Zk plus Z1 dash Zk and all to the power of k combinations like this and we can distribute it out, average it out, over that. Is that okay? Yes or no?

(Refer Slide Time: 11:21)



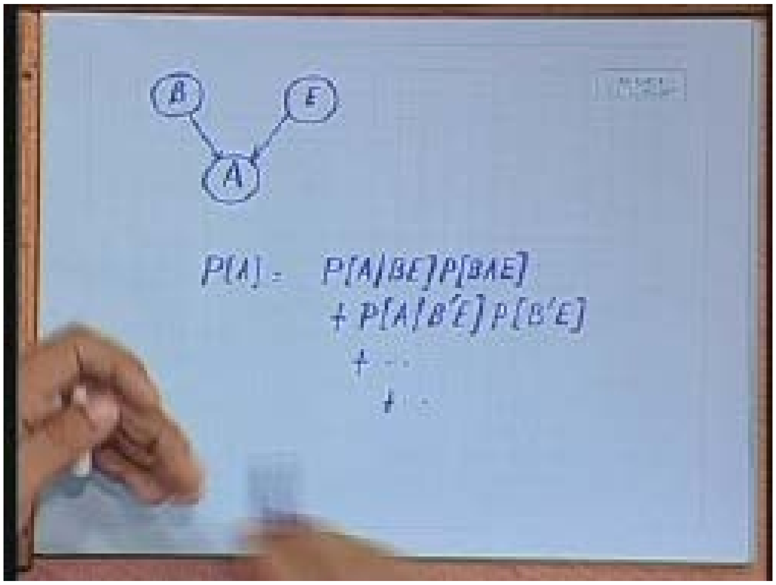
(Students speaking). Yes, like, probability of alarm, you can write this as probability of alarm, burglary, earthquake; probability of alarm, burglary, not earthquake; probability of alarm, not burglary, earthquake and probability of alarm, not burglary, not earthquake. (Students speaking). Yes, average it out, because we are talking about probabilities, so, averaging means taking up the sum over the- taking the weighted sum, actually.

(Refer Slide Time: 12:09)



If we do that for this term PEYI except X over X, then, what we have is, if we take it over Yi and Zi, then, we have this Xi, Yi, Zi here and summation over all Yi and all Zi, and we also have the term PYi Zi given X. Just have a look at this. Not clear? (Students speaking). Recall our burglary alarm example, so, we had alarm and we had burglary and we had earthquake. So, this was what we had. If you look at probability of alarm, then, we can write this as probability of alarm given burglary, earthquake times probability of burglary and earthquake plus probability of alarm given not burglary, earthquake times probability of not burglary, earthquake plus- got it? This is what we are doing here.

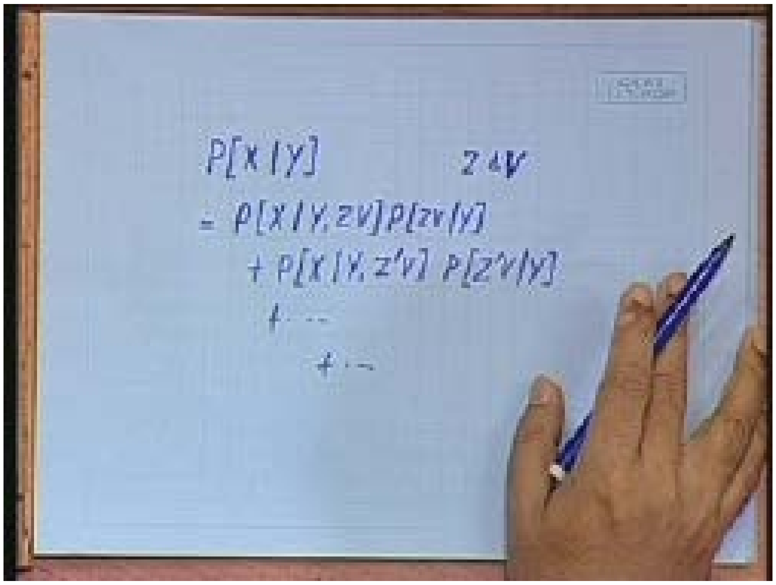
(Refer Slide Time: 14:01)



If you look at this term now, you see, it is exactly the same thing that is happening. We have all these: burglary earthquake kind of pairs here and then, we have the probability of this burglary, earthquakes, etc., and because this was the conditional probability with X, so, we have the conditional X here as well. Is it clear? It is the same thing that we are doing here; we are just averaging out over the parents of Yi or over the Yi and the Zi. (Students speaking). You can do this regardless of what Yi and Zi are- you can always do this averaging, you take any 2 variables and you can always write this average and it is a correct probability expression, is it not?

(Students speaking). Okay, why I am doing that- that will be become clear later on, but the question is that if you just have any 2 variables- additional variables- 2 or more variables, can you or can you not do an averaging out like this? The question is that if I have any variable, say P of X given Y and I have variables, say Z and V, then, can I write this as probability of X given Y ZV times probability of ZV given Y plus probability of X given YZ dash V times probability of Z dash V given Y, and so on. The question is, can I write this regardless of what Z and V are? I can do that, so, that is what we are saying, that for the time being, what we are doing is, we are averaging out over Yi and Zi.

(Refer Slide Time: 16:15)

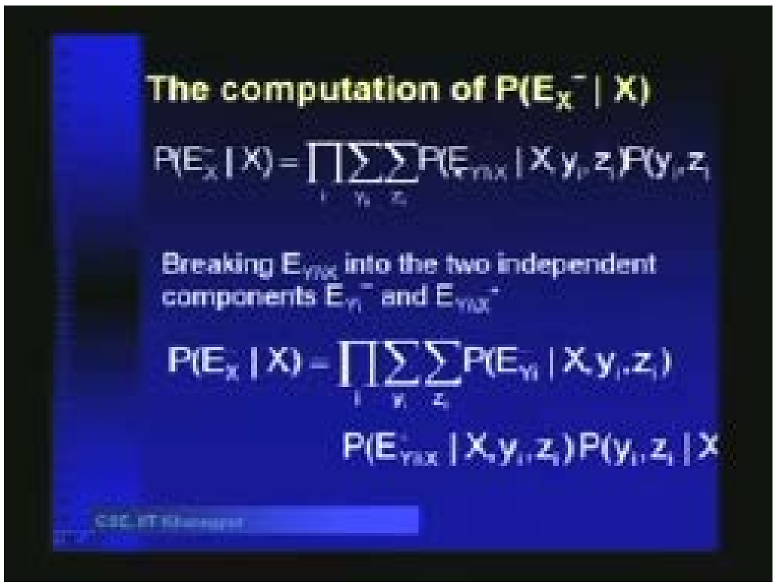


Why we are doing that? That will become clear as the analysis progresses, but the fact is that we can always do this. There is no- it is not incorrect to do this right now, whether it is useful or not, that we are going to see. Now, averaging this out, then, we do the following thing: we break this EYI except X into 2 independent components: EYI minus and EYI except X plus. Let us see what is meant by this. If you go back to our picture here of the Bayes network, then, we have this YI, so, we have some evidence above the Yi and some evidence below the YI.

The evidence above the Yi will be called EYI except X plus, so, all the evidence which is connected above YI, except those which goes through X- we are not considering these ones that will be called EYI except X plus and whatever we have below the Yi is simply E of Yi minus. Here, we need not consider- we do not have to specify except X, because whatever is below Yi does not have anything to do with X at all. Having broken up this EYI except X into those 2 parts, so, we have this split up as EYI minus given XYI Zi and P EYI except X plus this thing.

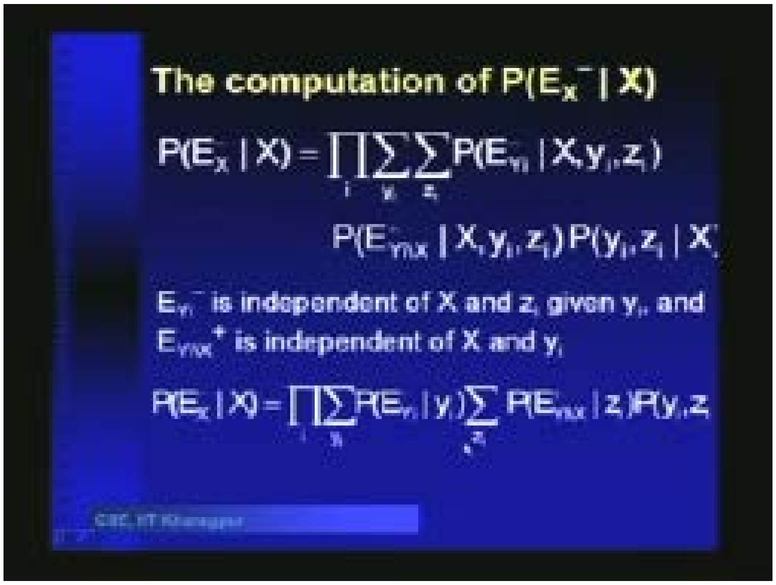
Now, we are saying that this probability term can be split up as the product of these 2 terms, because these 2 are independent. And why are they independent? Because we have Yi here. YI is given, so, if Yi is given, then, Yi d-separates out: the EYI minus and the EYI plus. Therefore, these 2 are independent and we can write them as the product of 2 probability terms.

(Refer Slide Time: 16:49)



This is where we are right now, that we have these 2 probability terms: EYI minus- here, is independent of X and Zi given YI, because Yi separates them. If you look at the picture, see EYI minus is all the evidence that is here, so, that is independent of X and Zi, because it has to go through Yi and Yi is given. Therefore, we can simplify this term: the first term, as P EYI minus given Yi the second term in the second term EYI except X plus is independent of X and Yi because- why is it independent? Because again, Yi separates it.

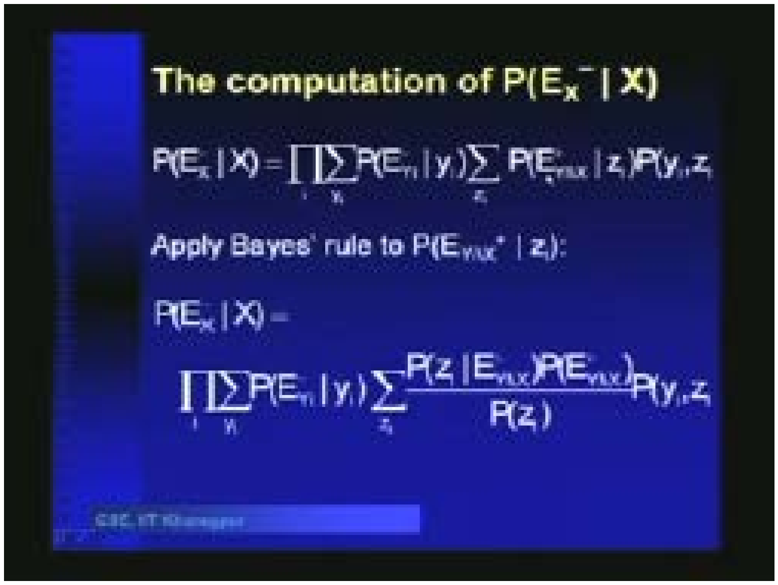
(Refer Slide Time: 19:56)



Therefore, we can simply write this as P EYI except X given Zi and then, we have this term which we originally had. See, what we are trying to do is, if you look at the picture, what we are trying to do is, we are trying to split up the probability terms in terms of the parents of Yi and the children of Yi. That recursively, from the term for X, we will be able to recursively bring it down to the terms of its children, right? That is the idea that we are trying to do.

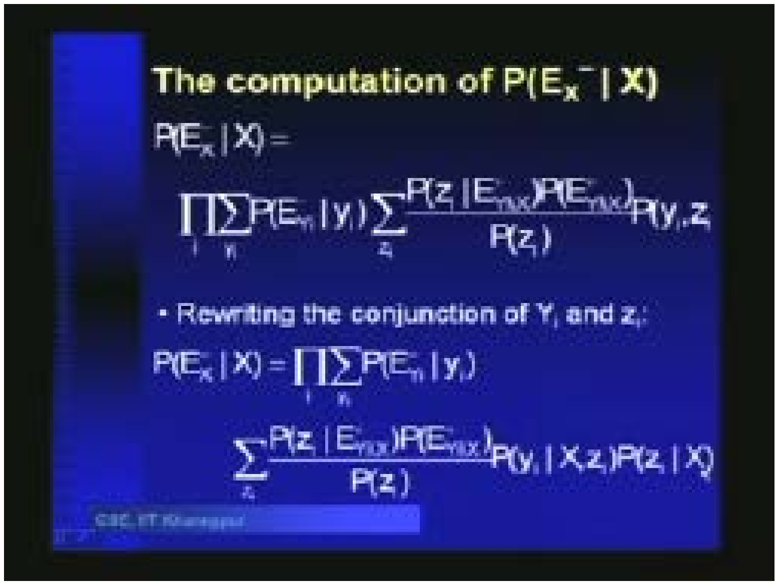
Then, we apply Bayes rule to the term PYI except X plus, and that gives us this term, that PZi except Yi this thing and P EYI except X plus divided by PZi, so, this simply- it is this term which gets replaced by this term using Bayes rule. The reason we are doing this is that this is again in the wrong direction. This is in the wrong direction- this is the cause and this is the effect, so, we need to turn it around. Whenever we need to turn it around, we use Bayes rule. After using Bayes rule here, you can see that we have it in the right order, so, Zi given PYI except X plus.

(Refer Slide Time: 21:48)



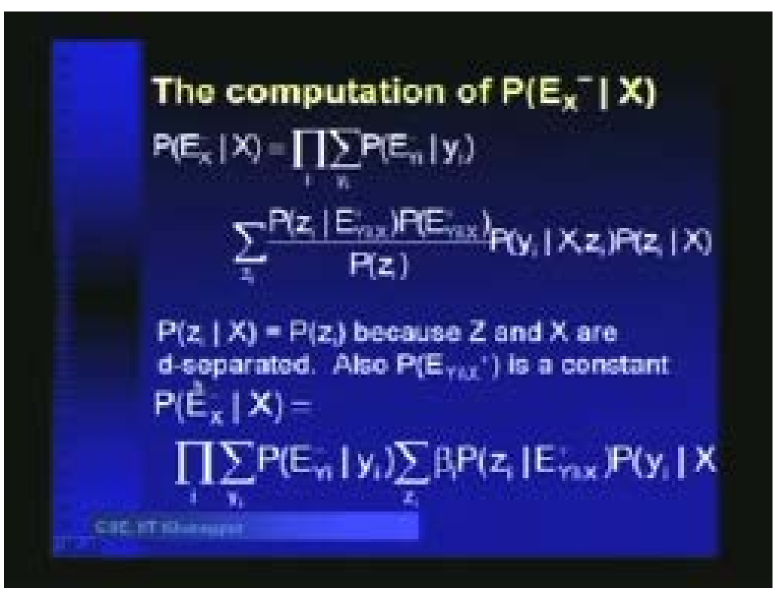
Then, after rewriting the conjunction of Yi and Zi, we get this. See this Y, this PYi Zi given X is has been broken up into PYI given XZ and PZi given X. Again, standard way of writing this using Bayes rule. After having done this, then, we note that PZi given X is equal to PZi, because Z and X are d-separated, if you look at the figure again: this and this are d-separated by this.

(Refer Slide Time: 23:12)



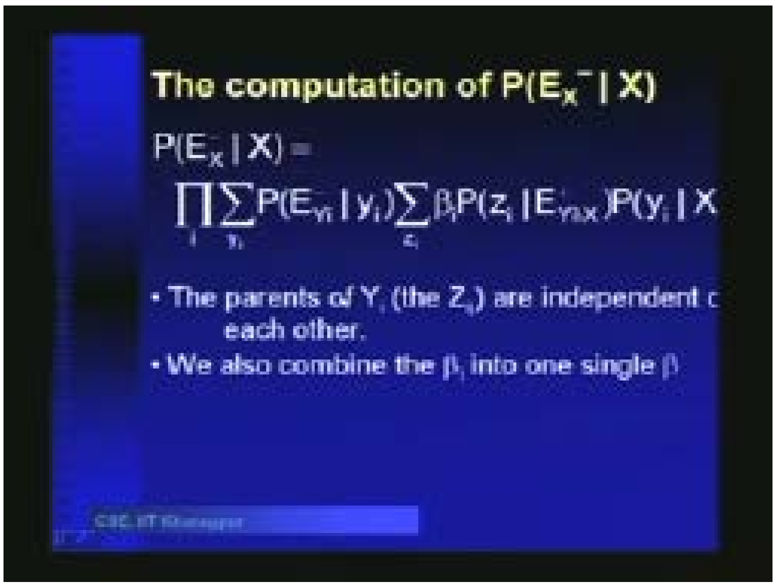
If there is no evidence about this, then, this and this are independent and in this particular probability term, we are looking at PZi given X, so, Yi is not given, so therefore, Zi and X are independent, so therefore, we can write PZi given X as simply PZi and also PYi except X plus is a constant, because this does not have any XY, etc.

(Refer Slide Time: 23:03)



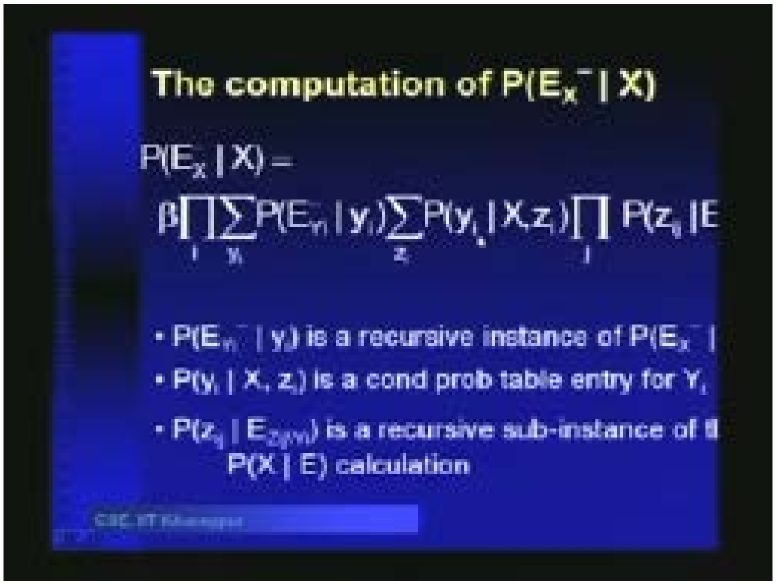
That gives us this term, where we replace this constant term by beta i, and we also replace PZi given X by PZi. Then, what do we have? Then, this is the reduced expression that we have. Now, the parents of YI, namely the Zij, are independent of each other. And why are they independent of each other? Because Yi is not given in this term, because Yi is not given, so, these Zi are all independent of each other.

(Refer Slide Time: 24:37)



Therefore, we combine all the beta i's into 1 single beta that gives us finally this expression. Now, if we look at each of the individual terms, then, this is a recursive instance of P EX minus given X, because it talks about PYI minus given YI, but it is a recursively smaller instance, because we have moved 1 level down from X. We have moved into YI, which is a child of X. Then PYI given X and Zi is a conditional probability table entry for YI, so, we get that directly and P of Zij given E of Zij except Yi is again a recursive sub-instance of the PX given E calculation.

(Refer Slide Time: 25:33)



But because we have this except Yi, so therefore, the evidence set has reduced. See, in order to guarantee termination, we have to show that the recursive sub-instances that you create are smaller than the original one. How is this smaller than the PX given E calculation? It is smaller because the set of evidence nodes is now reduced, because the evidence nodes which are through Yi are not there anymore in this computation, so, then, again, recursively, we have to apply the same procedure. What was the objective of going through all these computations?

The objective of going through all this computations was to show that if you use Bayesian analysis, then, things can become pretty complex. If you have Bayes network which are not small enough. Otherwise, if you have to find out the probability of any particular kind of event or any conditional event, then, it can become really complex to compute this. Moreover, what we have seen so far is only tree like instances; we have looked at only tree like instances. But in general, the Bayes network need not be tree like; it can be a complete graph also. If it is a graph, then, we can have further combinatorial blowup.

What we are now going to see in the remainder of this chapter is, what alternative does 1 have to Bayesian reasoning? Probabilistic reasoning is very important, because as we had discussed at the beginning of the lecture, of this chapter, that there are cases where you do not know of the exact cause-effect relationships, but we derive that statistically. Therefore, it is easier to think in probability terms than to be able to comprehensively define all kinds of causes for a given effect.

That is why people do use probabilistic reasoning and the problem that we see here is, if you use strict Bayesian reasoning, then, it can become pretty complex, so, people have thought of other ways doing this, which required less computational overhead. We will look at some of the other theories in brief and see how our computation goes there, but before we move into that, I would just like to touch upon what happens if we have Bayes networks which are not tree like. What approaches can we have then? They are called multiply connected belief networks, belief networks which are not singly connected.

Singly connected is the ones that we were talking about; between a pair of nodes, we have only 1 undirected. What could be an example of a multiply connected belief network? Let us look at an example which is as follows- that we have the event cloudy, and then, we have a sprinkler and we have rain and we have wet grass. And the sprinkler should normally not go off when it is cloudy, because it expects rain if it is cloudy; then, we can have rain and both of these has an effect wet grass.

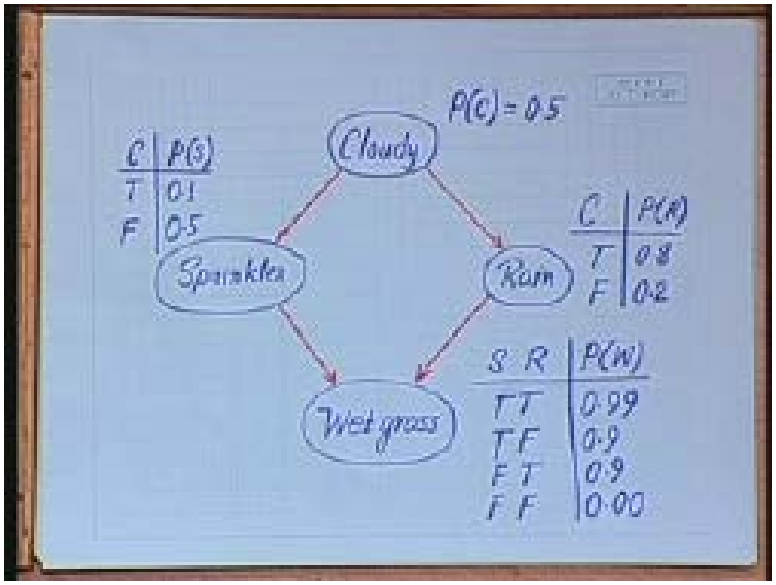
This is an example of a network where we have multiple connections. Wet grass can be result of sprinkler; it can also be the result of rain and both are conditionally dependent on whether it is cloudy or not. Now, how do we analyze this kind of networks? The problem in our previous approach will be that you can have- if it is a multiply connected network, like this, say, what was happening there? We were recursively decomposing the problem into smaller and smaller problem instances and as we were going out from the center of the tree towards the leaves, our problem size was diminishing.

And when we reach the leaf nodes, we are solved. But here, what may happen is that different paths in the network can create similar sub-problems along different paths. It is not just that you are propagating from the center outwards; you are, but what may happen is, through some part, you generate a sub-problem instance, through another path also, you generate another sub-problem instance and depending on the number of paths that you have in the network, the number of sub-problem instances can be growing.

Again, it boils down to things like dynamic programming and what ways to compute those. For example, if you have similar sub-problem instances, it might make sense to not compute them in a depth first manner or if we do compute it in a depth first manner, then, to memorize it in some way. People have also thought of alternative approaches for handing multiply connected belief networks and we are going to look at a couple of these approaches. Firstly, let me add on some probability value here- just a minute- so, let us add some probability values here. So, in this table, we will have cloudy versus probability of sprinkler, so, in this table, let us say for true, we have 0.1, for false, we have 0.5. If it is not cloudy, then, the probability of the sprinkler going off is 50 percent and otherwise, it is just 0.1.

And let us say probability of cloudy is simply 0.5 and then here, we have cloudy and probability of rain, and let us say that if it is true, then, the probability of rain is 0.8; if it is false, then, the probability of rain is 0.2. And then, for this one, we have sprinkler rain and probability of wet grass, so, we will have four possibilities here: namely, true, true, and then, we have a probability of 0.99. Let the grass is wet, then, true-false probability of 0.9, false-true probability of 0.9, and false-false: then, we have the probability of 0.00 nearly, which means that it is almost zero.

(Refer Slide Time: 35:37)



What the first method attempts to do are the clustering kind of method attempts to do is to combine these 2 events into 1 event. Slides please- transform the net into a probabilistically equivalent poly-tree by merging offending nodes. What we are going to do here is, we will try to combine these 2 into 1 single node, but then, the event that we have on the right hand side; the probability of the event that is not going to be just sprinkler or just rain anymore; it is going to be sprinkler plus rain, kind of stuff. What will happen is, we will reduce this to an equivalent network where we have cloudy, we have sprinkler plus rain and we have wet grass; the probability of cloudy remains as 0.5.

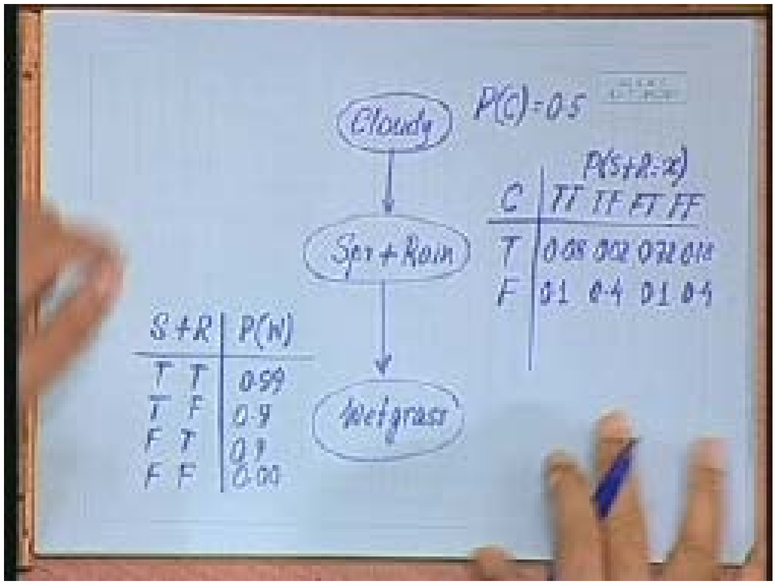
The interesting thing will be the probability table for this one. We are going to have c which is cloudy as before and we will have this probability of s plus r, which we will call as the new random variable x, so, it is just the sum of these 2 random variables. Again, we can have true true true false false true and false false, and the conditional probability of this thing is going to look like- if it is true, then, this is 0.08, then, this is 0.02, then, this is 0.72, and this is 0.18; where are we getting these probabilities from? If you look at this thing that we had for the sprinkler- 0.1 here and 0.8 here- when it was cloudy. The probability that we have both is 0.08, we had when it was cloudy.

This was what we had for true. How do we get the true false one? This times 1 minus 0.8. Are you getting it? This is the probability of s given c and this is the probability of r given c, when we are looking at true false, it is probability of s and probability of not r and probability of not r given c is 0.2. times 0.1 gives us 0.02, so, in this way, again, when we look at false-true, so, this is going to be 0.9 and 0.8, so, that gives us 0.72 and false false is 0.9 times 0.2, 0.1. Similarly, we have also the entry for c as false, in which case we will look at these entries in the table. (Students speaking). For the sprinkler given, yes, so, this is 0.5. (Students speaking).

It will be used now. See, when we look at cloudy not cloudy, then, what is the probability that we have the sprinkler and the rain? That is going to be 0.5 into 0.2, so, 0.1. Then, when we look at cloudy, but we have sprinkler, not rain, so, that will be 0.5 into 0.8. That is going to be 0.4 then, not cloudy but we have rain, so, that is going to be 0.5 into 0.2, so, 0.1 and then, not cloudy and we do not have rain and we have the sprinkler, so, that is going to be 0.5 into 0.8, so, 0.4. That is how we get the conditional probability table for this. What about the conditional probability table for this? For this, we will again have s plus r and we will have p w and then, this is going to be just as before.

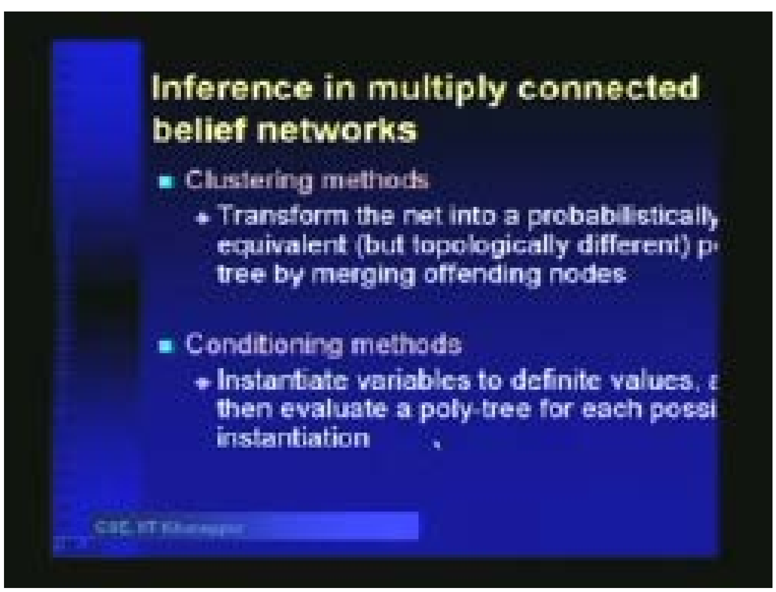
It is going to be true true true false false true and false false and we have exactly the probability values that we had previously, namely, these values: 0.99, 0.9, 0.9. Now, having done this, we have converted the multiply connected belief network into a single connected network. Then, we use the same algorithm as we have done through.

(Refer Slide Time: 42:16)



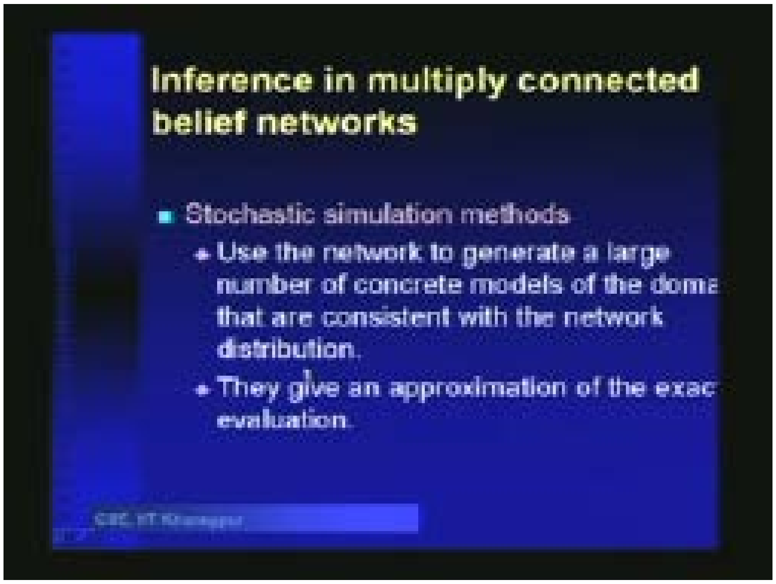
That is 1 approach to handling multiple connected belief networks. Another approach is instant conditioning methods, where we instantiate variables to definite values and then, evaluate a poly-tree for each possible instantiation.

(Refer Slide Time: 43:00)



Then, the idea is that we have this multiply connected belief network. What we are going to do is, we will instantiate 1 of these variables to some definite value. That is going to take off that part and then, we evaluate the remaining network which is a poly-tree, but 1 of the most popular ways of analyzing belief networks is stochastic simulation methods. What we do is, we use the network to generate a large number of concrete models of the domain that are consistent with the network distribution, and they give an approximation of the exact evaluation.

(Refer Slide Time: 43:37)



Let us understand what happens here. Here, we start from the top most node, from the start node, and we progressively go downwards. Let us see what we are going to do; we will start by choosing the value of cloudy with the probability of 0.5. Now, what do we require in order to do this? We require to be able to simulate a distribution which gives us samples of the value of cloudy with 0.5 probability. Suppose we have a generator which generates the value of cloudy and we want that generator to be such that half of the times it gives yes; half of the times it gives no.

That is what we mean by simulating the probability distribution. So, we pick up the value of cloudy by this random choice. Having picked up the value of cloudy, then, suppose cloudy is true; if cloudy is true, then, we will pick up the value of sprinkler using a distribution which says yes 10 percent of the time and no 90 percent of the time. Understood why we are choosing that? Because we know that if cloudy is true, then, the probability of sprinkler is 0.1. Then, similarly, if cloudy is true, then, we pick up the value of rain with a probability generator which will say yes 0.8 number of times and no 0.2 number of times, and then, having chosen the values of sprinkler and rain, then, suppose we have chosen sprinkler to be true and rain to be false.

Then, with 0.9 probability, we will choose wet grass to be true. It is all in simulating the probability distribution and generating this. Then, we will note down that whether we got wet grass or not. Again, repeat this experiment large number of times and then, find out what is the probability of the event that you wanted to determine. This is going to simulate the scenarios which the belief network captures with the biased by the probability distribution of this belief network and then, eventually, after doing a large number of experiments, then, you see what is the probability or what is the percentage of times where the event whose probability you wanted to compute- how many times did that occur?

Suppose we wanted to compute the probability of cloudy, given wet grass. We do this simulation and determine on how many occasions did we have cloudy and wet grass, and on how many occasions did we have cloudy and not wet grass. We do that, and then, from that fraction, we conclude the probability value and there has been a large number of different kinds of ways of doing this simulation. Some of the most popular ones are Monte Carlo simulation. You must have heard of Monte Carlo simulation: they actually do this kind of simulation. Then, there are more recent methods like Gibbs sampling.

You can read up the book and some other related references that are given here, to know more about the different kinds of stochastic simulation methods. In fact, the simulated annealing method that we had studied is also a kind of stochastic simulation. In the next class, what we will do is, we will look at some of the other kinds of reasoning that people do- probabilistic reasoning that people do- like fuzzy logic and other kinds of things, and see what are their relative merits with respect to Bayesian analysis.