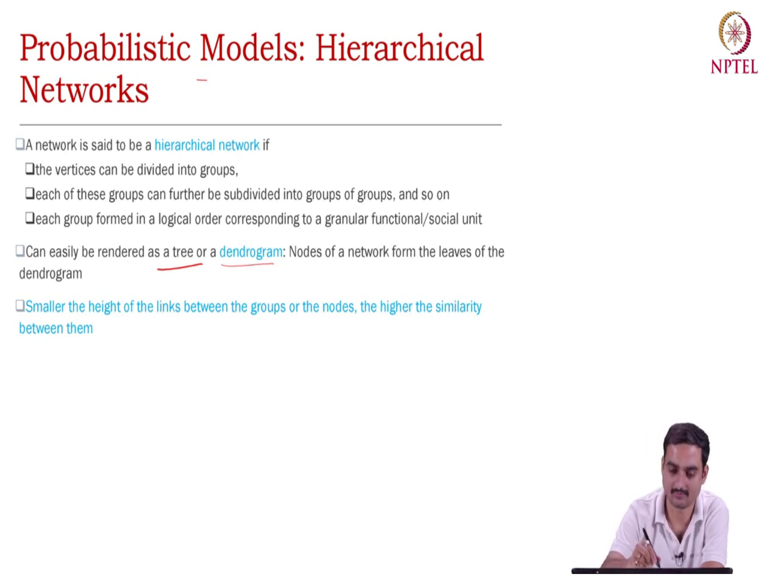
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 06
Lecture - 04

Now, we will discuss one of the probabilistic models and this model is based on the idea that the network actually follows a network is constructed based on some sort of hierarchical structure ok.

(Refer Slide Time: 00:37)



A network is said to be a an a network is said to be a hierarchical network if nodes can be divided into groups. Each of these groups can further be divided into subgroups and so on. Each group formed in a logical order corresponding to a granular functions or social unit right.

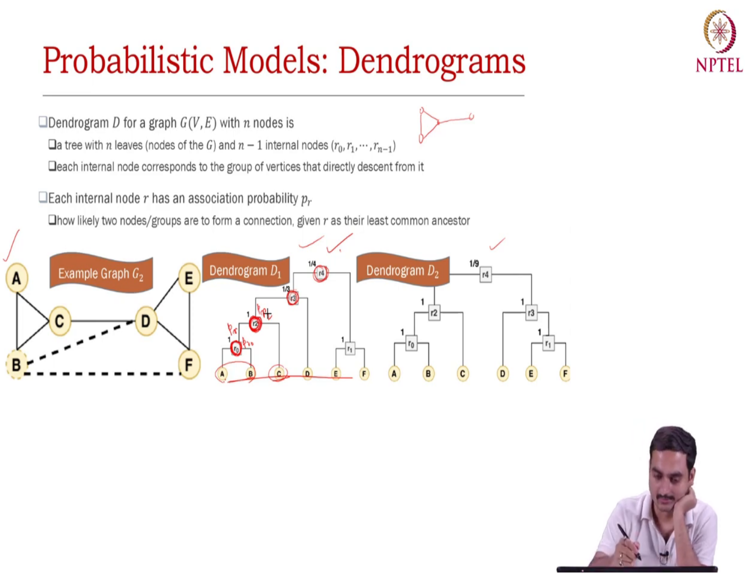
And this is well accepted at least you know from the last chapter on community analysis you must have understood that you know network actually contains some inherent hierarchical structure, we have you know big bigger community then smaller then further smaller and so on and so forth.

And the notion of hierarchical community and the formation of a network these are interrelated. So, here the idea is that can we given a network, can we unfold the underlying hierarchical structure based on which a network is formed ok. So, basically the we will try to come up with such an hierarchy right which would tell us, how this network was formed ok.

In that as in that aspect this is also a generative model because based on the hierarchy let us assume that we know the hierarchy, based on the hierarchy we will generate the network and we will see that the network generated by the hierarchy at all matches with the network that is already given to us ok, but there are many questions that you can understand that how do you generate the hierarchy from hierarchy how do you generate the network and so on and so forth ok.

So, in general when we look at this hierarchy we call this hierarchy as dendrogram and dendrogram is a very famous I mean well known term which generally is used in the context of clustering right. We have we you may have heard about you know this hierarchical clustering when we got single link partition you know all these link partitioning methods and so on alright.

(Refer Slide Time: 02:53)



So, again what is the target? The target is that given a graph G. The target is to come up with a decent dendrogram decent hierarchy that explains the network generation process ok, and the hope is that when we get the exact dendrogram right we should be able to know that what is the probability that two nodes will be connected, because this dendrogram itself is responsible for creating a graph right.

Now, then of course, you will be able to say that given a particular pair of nodes, what is the probability that they will be connected in the future ok. So, what do you mean by a dendrogram here. Let us assume that this is the graph ok, now from the graph you can actually create infinite number of dendrograms, exponential not infinite at least exponential number of dendrograms ok and each of these dendrograms would look like this.

So, this is one dendrogram dendrogram D 1, hierarchy D 1 right where, leaf nodes are the original nodes present in the graph right. So, there are 6 nodes right and according to this dendrogram A and B are connected first, then C then C is connected then D is connected right with the process right, I mean what I am what; I am trying to say is that first A and B are connected then C also got connected with A and B, then D got connected with C and so on and so forth right.

And, each of these internal nodes it sounds little bit tricky to understand, but you just follow what I am saying at the end of the lecture, you will understand what does it mean ok. So, each of these internal nodes right indicates the probability the association probability how likely two nodes or two groups right when I say that what about this node because this node connects a group to a node right.

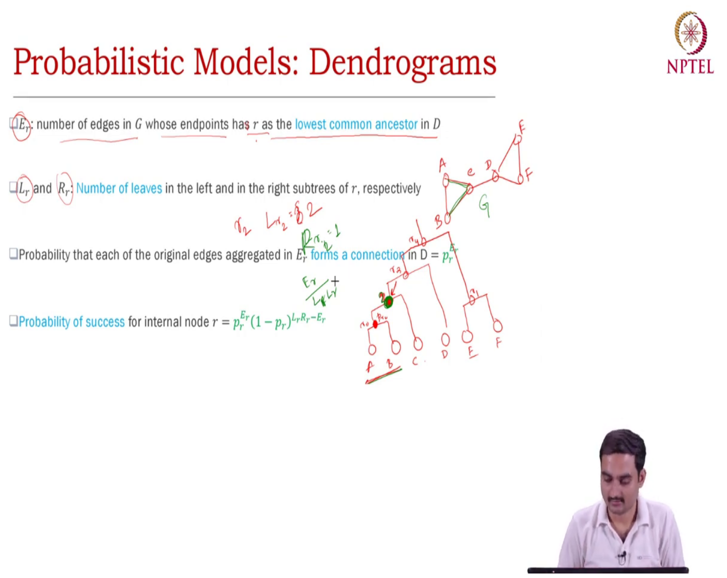
What is the, what is the association probability. What is the likelihood that two nodes or two groups are to form a connection ok in the original graph. Given that r, so say let us say this is r ok, r as their least common ancestor ok because if you look at it carefully A and B and C they may also have this as a common node common ancestor node, they also have this as a common ancestor node right.

But which one is the least one. This is the least one, this is the recent one. If you move from bottom to top this is the recent one ok. Now, you may ask that how do you; how do you come up with this hierarchy right, I can I mean I can connect this nodes in a different manner yes that can also be possible let us look at this dendrogram here if you see it carefully nodes are connected in different manners and they say this connection strategy is different from D 1.

In fact, as I mentioned there can be exponential number of such dendrograms that you can compute ok. So, again recap each of this internal nodes. So, all these leaf nodes are the nodes that are present in the graph, but internal nodes indicate. So, internal nodes are virtual nodes these nodes; these nodes are not present in the graph ok. These nodes kind of indicate the association between pair of nodes or a pair of graph; a pair of groups ok.

So, and another thing to note is that every internal node is also associated with the probability p r every internal node r is associated with the probability p r. How do you compute p r we will discuss ok. So, for r 0 there is p r 0 with r 2 there is p r 2 and so on and so forth ok.

(Refer Slide Time: 07:18)



Now, let us first discuss that how to compute p r ok. So, p r let me draw the diagram again right, one of the at least one of the dendrograms again say this is the original graph A B C D E F and this is one of the dendrograms A B C E F ok. This is one internal node this is another internal node right say this is r 0, this is r 1 right similarly r say r 0 does not matter say this is r 1 right this is r 2 ok this is r 3 and this is r 4 ok.

So, how to how do you compute p r 0. To compute p r 0 we need 3 quantities one is E r ok, for a internal node r you need E r, you need L r and you need R r ok. So, what is E r? So, if you look at this one right say this is r ok. What is E r? E r is the number of edges in the graph G in the original graph G whose endpoints has r whose endpoints have r as the lowest common ancestor in D ok. What does it mean?

It means that let us take this internal node right you will it will be easy to explain. Let us take this one ok. So, this internal node it has two left children A and B and one right child C right. We also quantify L r and R r. L r is the left child R r is the right side left leaf and right leaf. So, for this node r 2, L r L r 2 would be 2 number of left leaf nodes A v this is 2 ok, this is 2.

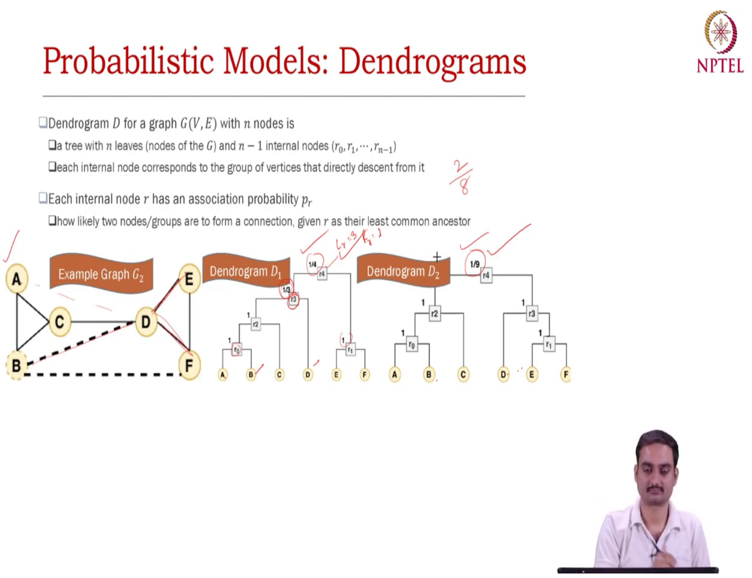
So, there are two leaf nodes there are two left leaf nodes and there is one right leaf node ok. So, you have two groups A B and C. So, how many possible combinations are how many possible combinations you can think of two times one A B A C and B C ok, you have two groups A B and C right. How many possible combinations can be there A C and B C.

Now, out of this A C and B C pairs ok, how many of; how many of such pairs actually are materialized in the form of edges, means how many such pairs are actually forming edges in the original graph ok. So, A C and B C you see that both A C and B C form edges in the original graph and that is your E r. So, E r is a number of edges in the original graph G, whose end points A B C whose end points whose end points have r. So, this is r, r 2 in this case r as the lowest common ancestor in D ok.

So, we take an internal node for which we want to measure p r we look at its left node left leaves you also look at its right leaves and you see how many pairs are possible right. Now, L r is the number of left nodes R r is the number of right nodes. So, L r times R r is the number of is the number of pairs right that is it, that is possible.

And, out of L r times R r out of L r times R r possibilities right. How many such pairs actually have been materialized in the form of edges and that is E r ok. So, that is the idea. So, and we will show later theoretically that the that optimal probability that the probability p r would be E r divided by E r divided by L r times R r, meaning out of all the pairs how many pairs are there in the graph as edges ok.

(Refer Slide Time: 13:19)



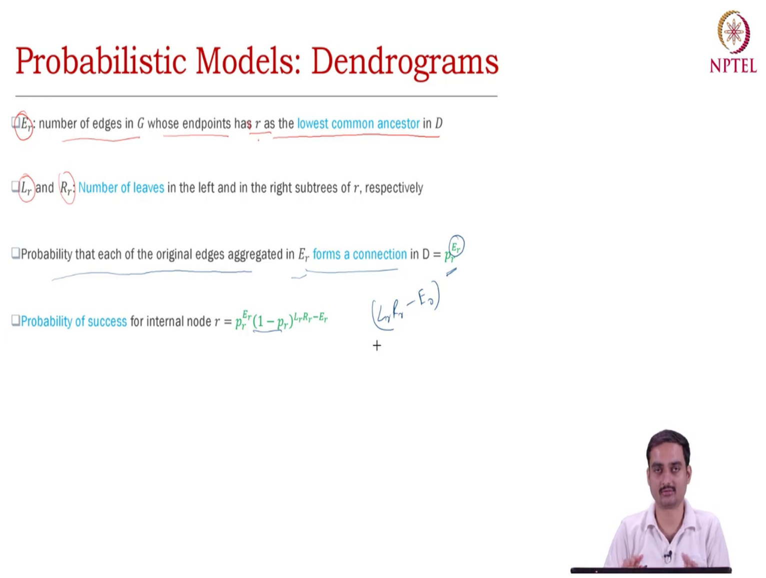
So, now let us look at an example let us look at; let us look at this one let me erase it ok. So, let us look at this internal node. So, it has L r equals to 1, R r equals to 1. So, denominator is 1 and the numerator. So, the number of such pairs is AB right and AB also exist in this one. So, the this is also 1. So, 1 probability is 1. What about r 2, how many pairs are possible A C and B C.

So, denominator is 2, 2 times 1 and both A C and B C are there therefore, 2 by 2 its 1. What about r 3, how many pairs are there. So, in for r 3, L r is 3 and R r is 1 ok. How many pairs are there A D B D and C D. So, A D is not there, A D pair is not there no A D edge is not there, B D edge is also not there, C D edge is there. So, the numerator is 1.

So, this is 1 by 3, 1 by 3 times 1. So, 1 by 3 ok, what about this one. So, how many pairs are possible? 4 times 2 8 and how many of them are actually, how many of them actually exist? A E not there, A F not there, B E not there, B F not there right, C E is not there C F is also not there right. D E is there and D F is also there. So, 2 numerator is 2. So, 1 by 4 similarly this one ok.

Similarly, for dendrogram D 2 you can compute the same things for every internal node you can measure the probability. So, now, we have understood that, given a dendrogram how do we calculate the probabilities of internal nodes ok alright. So, if this is; if this is understood then actually you have understood many things alright. So, now, let us look at the equations again.

(Refer Slide Time: 15:36)



So, what is the probability that each of these original edges aggregated in E r forming a connections in D this is p r. p r is something that we calculate right p r times E to the power r p r times E r. What is E r? E r is the number actual number of edges. So, for every edge that has been materialized for every pair that has been materialized as an edge the probability of the connection is p r right and how many such cases are there E r.

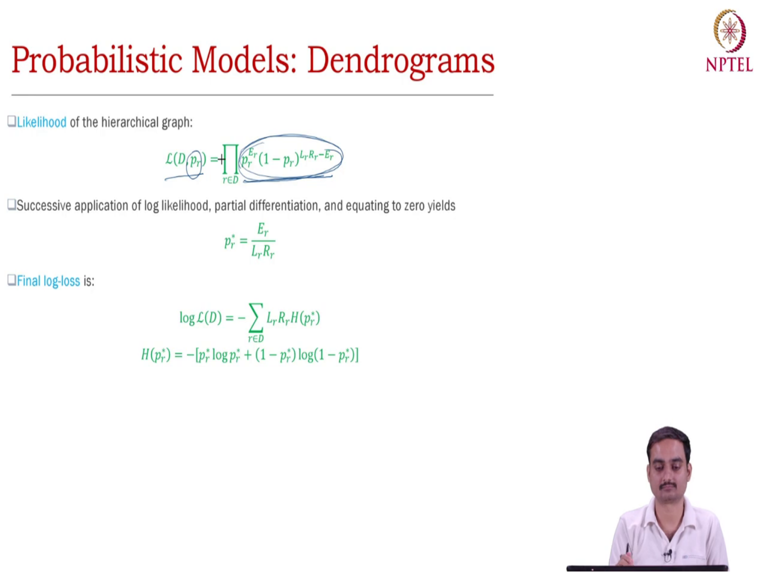
So, the total probability is p r times p r to the power E r binomial distribution right, and how many false cases are there, how many pairs are there which have not been materialized there are total pairs L r times R r. So, L r times R r minus E r this is the number of pairs which have not been materialized in the form of edges right and what is the probability? The probability is 1 minus p r.

So, 1 minus p r to the power L r R r minus E r. So, very same as that and if you remember correctly this is very same as the E r model that we discussed initially right in the network growth model the Erdos Renyi graph the random graph model. The random graph model we derived things in the same way, but there was no concept of p r.

There was only one probability which is p r which is p and we said that take a pair of nodes and connect it with the probability p and do not connect with the probability 1 minus p. Now, I am saying that the edges the formation of edges have different probabilities right and some edges will be formed with probability p r some edges will be formed with probability p r dash and so on and so forth right, that is the difference.

Another similarity is that in case of random graph model, we assume that the graph is formed I mean each edge the formation of edges is independent of each other right. Here also there is no dependency between the formation of edges, it is not the case that the formation of one edge is dependent on the formation of other two edges it is not the case ok.

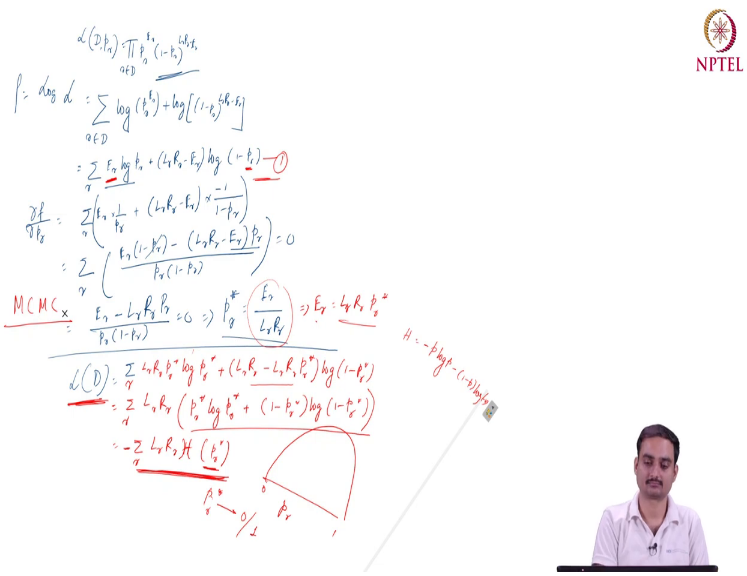
(Refer Slide Time: 18:03)



So, now, we compute what is the likelihood of a dendrogram remember we can have exponential number of dendrograms, we need to understand which dendrograms are meaningful right, for that we need to compute the likelihood ok. So, the likelihood of a dendrogram D given the p r given each such p rs, p r values would be basically product of all the inter I mean product of this one product of this quantity for all the internal nodes present in the dendrogram right.

For every p r, for every p r for every r we have a p r for every p r we can compute this one. So, that is your likelihood right. So, if this is your likelihood then when what is the target? The target is to maximize the likelihood. So, we basically want to get the dendrogram whose dendrogram who’s likelihood is maximum right and we know how to do this? Let us derive this.

(Refer Slide Time: 19:20)



So, we have already seen that the likelihood of a dendrogram is the product over all the internal nodes p r to the power E r 1 minus p r to the power L r R r minus E r ok, and we want to maximize this. So, we generated the log likelihood because this bulky product is difficult to compute we take the log. So, log of this L is. So, it would be sum over all r log of right p r to the power E r plus log of right 1 minus p r to the power L r R r minus E r ok.

And then this would be some overall r this is a power. So, this will come outside log of p r plus L r R r minus E r log of 1 minus p r ok and we want to maximize this. So, we want to maximize with respect to p r. So, how do we do that we take the derivative with respect to p r. So, say this is f right. So, d f d p r would be what, would be sum over log.

So, derivative would be E r times 1 by p r right plus ok plus L r R r minus E r times 1 by 1 minus p r because it is negative. So, it should be minus ok. So, this would be r E r 1 minus p r minus L r R r minus E r times p r by p r into 1 minus p r ok. This would be 0, if it is this 0 then you will see that what would happen. So, you have E r minus p and see this p r E r and this p r E r will cancel out right.

So, we will have L r R r p r divided by p r times 1 minus p r right this is 0. So, it would be p r equals to E r right by L r R r ok, and this is the optimal value of p r. We have already seen that the optimal value p r. So, we started off by saying that optimal value of p r is this one right E r by L r R r, but now we derived that y the optimal value is this one ok.

So, if this is the optimal value then again let us put this p r value here. Let us try to compute the likelihood of a dendrogram D that would be; that would be sum over let me change my pen and this is p star right. So, what we will do, I will actually put it here this is equation 1. So, I will what I will do? I will replace p r here by this one ok so. In fact, what I will do instead of doing this better would be to replace E r right.

Because E r is L r R r p r star ok I replace E r ok by this one. If I do that in equation 1 this would be L r R r p r star log of p r star plus L r, why I am deriving this thing because of beautiful equation will emerge eventually you will see this. E r R r minus L r sorry L r R r minus L r R r p r star log of 1 minus p r star ok this would be r ok.

E L r R r p r star log of p r star plus if you take L r R r L r R r common you will be, it will be this will be 1 minus p r star log of 1 minus p r star. What is this? This is entropy ok of course; negative entropy because entropy is minus p log p minus 1 minus p log of 1 minus p ok. So, this is r L r R r a entropy. So, entropy is H ok H p r star but negative.

So, this is the; this is the likelihood of a dendrogram think about it and we want to maximize this. So, we want to maximize this. So, this is negative therefore, we need to; we need to minimize this one; we need to minimize this one. So, for which value of p r this will minimize when entropy is minimized. So, entropy curve is this one right this is p r and this is 0, this is 1. So, entropy is maximum when p r is half right.

So, when entropy would be minimum, entropy would be minimum if p r tends to 0 or p r tends to 1. So, if p r star tends to 0 or 1, it would be minimum and you will get maximum likelihood. So, what is p r is the probability of an edge of formation of an edge. So, we basically say that higher likelihood dendrogram partitions I mean if you take the dendrogram whose which is highly likely.

So, that dendrogram partitions nodes into groups where connections are either very common, in that case p r tends to 1 or very rare p r tends to 0 because if p if connections are common then that would be connections within the group, and if connections are rare then that would be intergroup connections ok. So, now, we understood that we understood how to compute how to compute the how to compute the p r right.

Now, the question remains is how to generate because since there are exponential number of such dendrograms generated right can we do this thing for all possible dendrograms no right.

So, what you do; what you do? So, we basically sample out we basically sample out a set of dendrograms. So, there is this process there is this interesting process called MCMC this is called Markov Chain Monte Monte Carlo process.

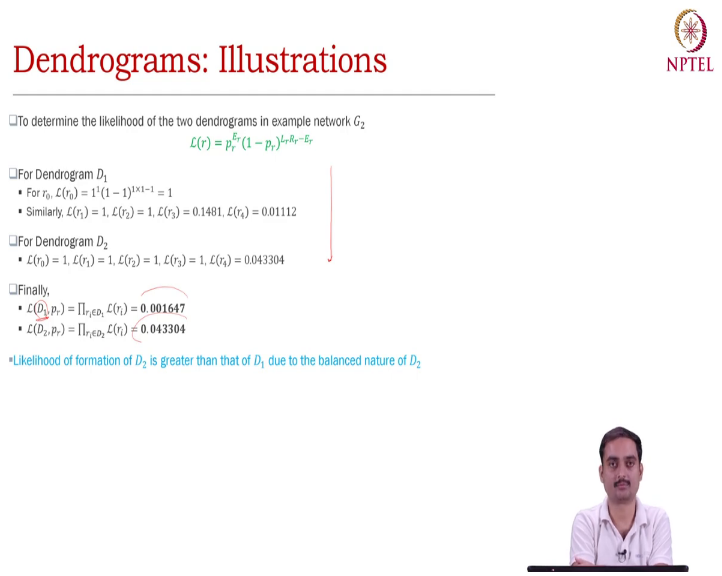
Using Markov Chain Monte Carlo process what we will do? We will sample I am not discussing what is MCMC, I mean you can just google it and get lot of materials. See using MCMC what you will do you will sample some dendrograms right with probability proportional with probability proportional to the likelihood right and let us say you sample 10 such dendrograms with high likelihood ok.

Now, let us assume that you sample these two dendrograms right. Now, how do you then use this 2 dendrograms for link prediction. Now, let us say you want to understand the what is the, you want to predict the link between B and D right. So, what you do? You look at the common ancestor of the least common ancestor of B and D in all the dendrograms.

B and D the least common ancestor is this one and the probability is 1 by 3. So, according to this dendrogram, the probability of formation of this edge is 1 by 3. According to this one B and D 1 by 9. So, the average probability is 1 by 4 plus 1 by 9 by 2. So, you take 10 such dendrograms, 10 such probabilities will come you take the average and that would be the probability of formation of an edge between that pair of nodes right.

You can do these things for all pairs for which you want to come you want to predict links will be formed or not and then we return top n pairs as your predictions ok.

(Refer Slide Time: 30:14)



So, the example I have already given the example if you do the same thing same derivative you will see that the likely of the dendrogram D 1 is this one, D 2 is this one and so on so forth right. So, let us summarize what we have learnt. Given a graph we will create dendrograms we will use MCMC to sample dendrograms based on the likelihood.

How do we get the likelihood? We get the likelihood based on this ps right we compute ps based on E r L r and p r and then we get the dendrograms. Now, given a pair of nodes we look at the corresponding least common ancestor and the corresponding probability and then you return ok.

You take the average and then you return those pairs of those pairs of nodes which will be which will hide which will be which will have a higher probabilities ok. So, that is about this hierarchical right this some sort of maximum likelihood based method using hierarchical structure of a graph ok.

Next class we will discuss another algorithm called the so, it is basically the algorithm will use we will see the algorithm you uses the random work process and this is called supervised random work process right SRW and this is the very famous algorithm for link prediction kind of a supervised method for link prediction, this is not supervised this is unsupervised the one that we discussed today ok, with this I stop here.

Thank you.