

(Refer Slide Time: 02:25)

• A distribution  $P$  factorizes over a Bayesian Network  $G$  if  $P$  can be expressed as

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa_{X_i})$$

• A distribution factorizes over a Markov Network  $H$  if  $P$  can be expressed as

$$P(X_1, \dots, X_n) = \prod_{i=1}^m \phi(D_i)$$

where each  $D_i$  is a complete sub-graph (maximal clique) in  $H$

A distribution is a Gibbs distribution parametrized by a set of factors  $\Phi = \{\phi_1(D_1)\}$  if it is defined as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{i=1}^m \phi_i(D_i)$$

So I'll just quickly recap what we've been discussing it, so we've been talking about graphical modules both directed and undirected, and driven by this primary goal that we are interested in joint distributions of a large number of random variables, and we're just considering the discrete case, and even in the discrete case we see that this is intractable because you end up with exponential number of parameters in your distribution, and it's impossible to specify those.

So the basic idea that we have been going towards is that given this large distribution, how do you factorize into small factors which we can deal with, and as you factorize a graph, as a factorize a distribution properly, what will happen is the number of parameters that you need to learn to fully specify the distribution is going to decrease, yeah yeah so we want to factorize it and this factorization essentially reduces the number of parameters that we have in the joint distribution, and that's what our goal has been, and then we saw that graph is a good way of representing this, and the nodes in the graph essentially are associated or with each node in the graph we have an association, associated conditional probability distribution in the directed case, and these conditional probability distributions are the factors in our graph, right.

And we also saw for some twelve examples how the number of parameters drastically decreases for these conditional parameters or conditional probability distributions as compared to specifying the full joint distribution.

And not only that it's also more compact, it's more modular, if you want to add new variables it becomes easier, and it's also more tractable computationally less storage statistically, lesser amount of data required because you have to learn lesser number of parameters and cognitively if you were to ask human to give you values for some of these tables, okay can you at least tell me what is the priority of salinity given pressure that is something that expert could probably tell you, he just has to give you these four values and that's easy and more tractable as asking him or her to specify the full joint distribution, I mean that's what our goal is we are always interested in reducing the number of parameters so that our learning eventually becomes easier.

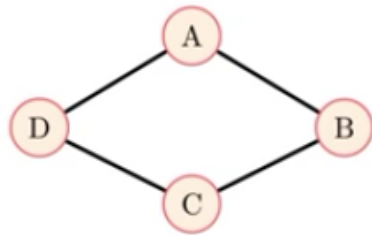
So far for all our discussion we have assume that someone is going to give us this factors, but eventually we'll head to a state where we'll try to learn these factors, okay.

And from the directed case we move on to the undirected case, because we came up with a very simple example and where we saw that having directions does not make sense, because there is no hierarchy in some cases and it's like both the factors or both the random variables that you are considering interact with each other, rather than depend on each other, right, I mean they do depend on each other but they dependence is symmetric instances, right, it's like an interaction both are equal contributors, and that's what happened in the study group example.

And from there we came up to undirected graphical models and where we argued that the factors in the undirected graphical models should correspond to these maximal cliques, okay. And that was not sacrosanct, you could also have cliques instead of maximal cliques, but it just that maximal cliques again gives you a minimum possible set of parameters and it captures what you actually want to capture, you want to capture the interactions between all elements of the study group, so why not you just one clique to represent that study group, okay.

So that was the factorizations on the left hand side you see the factorization for the Bayesian network which was factors were conditional probability distribution. On the right hand side you also see a factorization with the difference that the factors here are not probability distribution, they are just known as clique potentials, these did not take values between 0 to 1, these could be arbitrary things then we actually saw some examples of these arbitrary factor values, I just probably go to that quickly if it's nearby, it's not nearby, not nearby yeah something like this right, so it looks,

(Refer Slide Time: 04:05)



$\phi_1(A, B)$			$\phi_2(B, C)$			$\phi_3(C, D)$			$\phi_4(D, A)$		
$a^0$	$b^0$	30	$a^0$	$b^0$	100	$a^0$	$b^0$	1	$a^0$	$b^0$	100
$a^0$	$b^1$	5	$a^0$	$b^1$	1	$a^0$	$b^1$	100	$a^0$	$b^1$	1
$a^1$	$b^0$	1	$a^1$	$b^0$	1	$a^1$	$b^1$	100	$a^1$	$b^0$	1
$a^1$	$a^1$	10	$a^1$	$b^1$	100	$a^1$	$b^1$	1	$a^1$	$b^1$	100

- But eventually we are interested in probability distributions
- In the directed case going from factors to a joint probability distribution was easy as the factors were themselves conditional probability distributions
- We could just write the joint probability distribution as the product of the factors (without violating the axioms of probability)



the table actually looks very much similar to the probability distribution except that it's not a probability distribution, it's just some values which capture the affinity between different value, different possible assignments of the random variable, so 0 0 is more likely as compared to 1 1 or 0 1 or 1 0, that's what this table essentially captures, right.

And again I repeat so far we have assume that someone has given us these tables and we are talking about things that once these tables are given, what are the kinds of reasoning that we could do, so we saw some reasoning's like causal reasoning, evidential reasoning, explaining away and so on right, and the case of the directed graph at once, right.

And coming back to the undirected case, even though these factors are not probability distributions, they are not too worried about it that because we know that given any kind of real numbered values, we can always do this normalization so that the resulting quantity ends up being a probability distribution for that the values lie between 0 to 1, right, and that actually is one concern that we'll have to deal with going forward, that this partitioning function or the value Z which make sure that these factors eventually give us a probability distribution that is interactable, because it has to sum over all possible values of all the random variables, right, that's what Z tells us, it's over the entire universal set, it gives you the assignments to all possibilities in the universal set, and that's the same as, it's just the fancy we're saying what we already know right, we always divided by all possible outcomes in the set, that's what probability tells us right, your interest, events of interest divided by all possible events at the center, and that's what Z actually does.

And those number of outcomes are very very large they are exponentially actually, right, because we need to consider all possible assignments to all these values which even in the binary cases do rise to them, okay, so this is Z is going to be a problem going forward and at some point we will have to deal with that, okay.

So that's the summary of what we did before the summer vacation, and now we'll continue from that point, and we are again interested in this question of what are the independencies encoded by a Markov network, so let  $U$  be a set of all the random variables in our joint distribution right, so  $X_1$  to  $X_N$  is that set  $U$ .

And now let  $X, Y, Z$  they're some overloading of variables, now these  $X$  is different from the  $X_1$  to  $X_N$  that we've been considering so far, let  $X, Y, Z$  be some distinct subsets of  $U$  right, so say the first  $K$  random variables is  $X$ , the next  $K$  random variables is  $Y$ , and the remaining random variables are there, some distinct subsets so it doesn't matter in what order you have taken or whatever, okay.

(Refer Slide Time: 06:40)

- Let  $U$  be the set of all random variables in our joint distribution
- Let  $X, Y, Z$  be some distinct subsets of  $U$
- A distribution  $P$  over these RVs would imply  $X \perp Y | Z$  if and only if we can write

$$P(X, Y, Z) = \phi_1(X, Z)\phi_2(Y, Z)$$

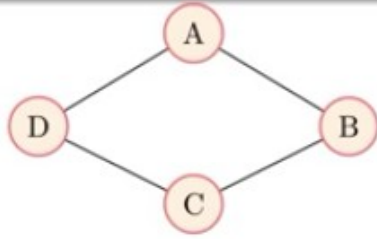


Now a distribution  $P$  over these random variables would imply that  $X$  is independent of  $Y$  given  $Z$ , if and only if we can write the joint distribution as a product of the following factors.

So what is so unique about these factors?  $X$  and  $Y$  do not appear in the same factor, right, and they do not appear in the same factor that means they are not connected, they are not part of any clique, right.

And  $X$  and  $Z$ , and  $Y$  and  $Z$  can appear in the same factor, that's fine, so given  $Z$ ,  $X$  is independent of  $Y$ , that's what this means, so if that condition if the distribution can be factorize like this, then it means that  $X$  is independent of  $Y$  given  $Z$ , right, this is again define the semantics of a Markov network just as we had define the semantics of a Bayesian network, okay.

(Refer Slide Time: 07:35)



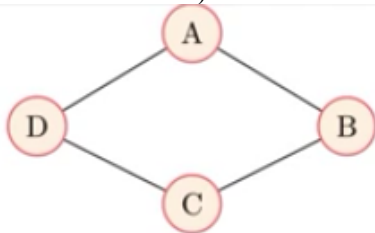
- In this example

$$P(A, B, C, D) = \frac{1}{Z} [\phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)]$$



So let us see this in the context of a original example right, I mean here this doesn't hold right, what was the independence in this example? Do you remember the independences? A is independent of C given B,D and B is independent of D given A,C. So based on the discussion that we just had,

(Refer Slide Time: 07:55)



$$A \perp C \mid \{B, D\}$$

$$B \perp D \mid \{A, C\}$$

- In this example

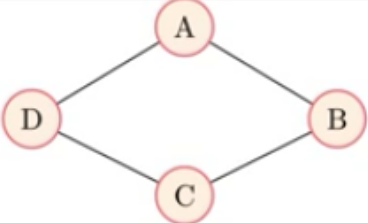
$$P(A, B, C, D) = \frac{1}{Z} [\phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)]$$



this joint distribution should have factorized in a particular way, is it factorizing in that way? So I told you a rule for when is X independent of Y given Z, and remember the X, Y and Z are sets of random variables, they did not be individual random variables, right.

So according to that rule what should the factors have been actually? What is X in the first case? What is Y? And what is Z? B,D so what kind of factors should we actually would have? Phi(X,Z) and phi(C,Z) right, or rather Y,Z, so this is X, this is Z, this is Y, this is Z, do we have factors of that form?

(Refer Slide Time: 09:00)




- In this example


$$P(A, B, C, D) = \frac{1}{Z} [\phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)]$$

$$\frac{A \perp C \mid \{B, D\}}{B \perp D \mid \{A, C\}}$$

$$X \perp Y \mid Z$$

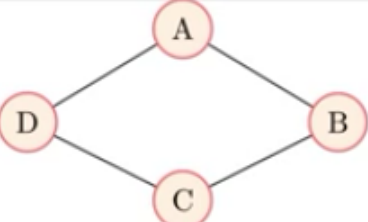
$$\phi(A, \{B, D\}) \phi(C, \{B, D\})$$



 23/26

You were allowed to be a bit creative do we have factors of that form? Okay, we just need to rearrange these terms right, I mean this let's see, we can write it as these two terms together is a larger factor depending on B,A,C,

(Refer Slide Time: 09:17)





- In this example

$$P(A, B, C, D) = \frac{1}{Z} \phi_1(A, B) \phi_2(B, C) \phi_3(C, D) \phi_4(D, A)$$

- We can rewrite this as

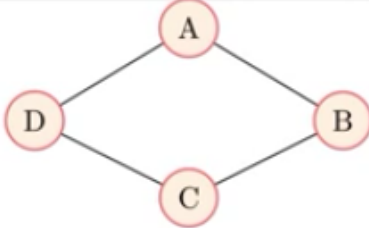
$$P(A, B, C, D) = \frac{1}{Z} \underbrace{[\phi_1(A, B) \phi_2(B, C)]}_{\phi_5(B, \{A, C\})} \underbrace{[\phi_3(C, D) \phi_4(D, A)]}_{\phi_6(D, \{A, C\})}$$



 23/26

and these two terms together is a larger factor depending on D,A,C, right.

So now this is X, this is Z, this is Y, this is Z, so we have the condition that X is independent of Y given Z, right, it's just a matter of rearranging these factors, and nothing changes, right, you still can have the modular factors where you have a phi 1 and phi 2 which operate only on AB and BC, it's just that using that you can always compute phi 5, not a good choice but okay, is that fine? Does that make sense? Okay, so that's the rule for Markov networks, (Refer Slide Time: 09:58)




- In this example


$$P(A, B, C, D) = \frac{1}{Z} [\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)]$$

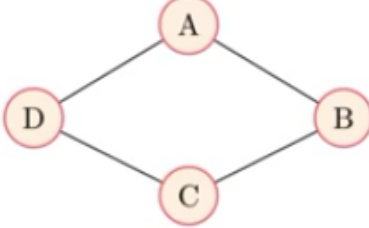
- We can rewrite this as

$$P(A, B, C, D) = \frac{1}{Z} \underbrace{[\phi_1(A, B)\phi_2(B, C)]}_{\phi_5(B, \{A, C\})} \underbrace{[\phi_3(C, D)\phi_4(D, A)]}_{\phi_6(D, \{A, C\})}$$

- We can say that  $B \perp D | \{A, C\}$  which is indeed true






 23/26



- In this example

$$P(A, B, C, D) = \frac{1}{Z} [\phi_1(A, B)\phi_2(B, C)\phi_3(C, D)\phi_4(D, A)]$$

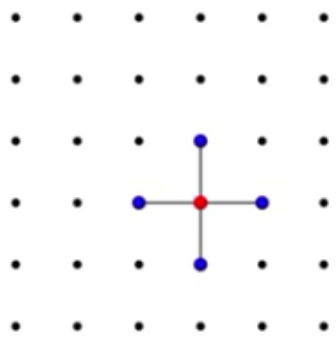






and it's also you could again do a different kind of rearrangement to get the other independence which was A independent of C given B, D, just need to arrange the factors a bit differently, okay. So if you could factorize the joint distribution as factors of the form  $\phi(XZ)$ ,  $\phi(YZ)$ , then X is independent of Y given Z, okay.

(Refer Slide Time: 10:24)

- For a given Markov network  $H$  we define Markov Blanket of a RV  $X$  to be the neighbors of  $X$  in  $H$





Now the next thing that we are going to define just as we had defined for a Bayesian network, we had define parents of a node. In the case of Markov network, we are going to define something known as a Markov Blanket, which is nothing but the collection of all the neighbors of  $X$ , right, so for any given Markov network, for a given random variable  $X$  belonging to this network, we can define the Markov Blanket of this  $X$  as all the neighbors of  $X$  and  $H$ , right, and this is illustrated in the diagram, okay.

Now what to consider as a neighbor is again something up to you, so can I consider these two to be neighbors? If I want I can, right, it's again a modeling choice which I make, say if I were talking about these things as pixels and image, I would probably decide to choose all of these as neighbors, right, but if it's some other application maybe where these diagonal neighbors don't make sense, so I'll just connect the horizontal and vertical neighbors, so that's completely up to me, but once I define these neighbors, and this is known as the Markov Blanket of  $X$ .

So now just as for the Bayesian networks we had these rule that a node is independent of all non-descendants given the parents, and now I have given you some kind of equalizer between parents and a Markov Blanket, can you tell me a rule for Markov networks? Is the analogy clear? How many of you get what I said just now, please raise your hands? Yes, so you had a question. No, you have to draw an  $H$ , okay.

So what I said is that I just started this discussion by saying that just as you had parents in the case of Bayesian networks, in the case of Markov networks I'm defining this Markov blanket,

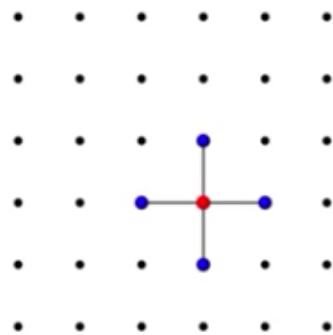


right, which is essentially everything that covers a given node, right, I've a case of Bayesian networks you had this rule that given the parents, the node is independent of all its non-descendants, right.

Remember in the case of Markov networks non-descendants does not make sense because there is no concept of descendants at all, okay.

So again given in Bayesian networks you had this rule that given the parents the node is independent of all its non-descendants, the parents analogy in the case of Markov networks is the Markov blanket.

So now can you give me a rule for the Markov networks? Given the Markov blanket node is independent of all other words, right, okay, is that make sensitive sense, right so given, (Refer Slide Time: 12:59)



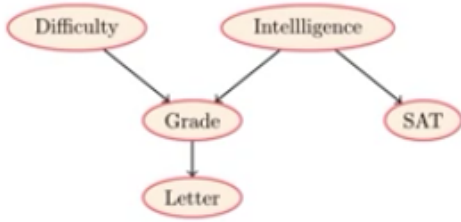
- For a given Markov network  $H$  we define Markov Blanket of a RV  $X$  to be the neighbors of  $X$  in  $H$
- Analogous to the case of Bayesian Networks we can define the local independences associated with  $H$  to be

$$X \perp (U - \{X\} - MB_H) \mid MB_H(X)$$



so  $X$  is independent of everything from the universal set except of course  $X$  itself and the Markov Blanket, given the Markov Blanket, does that make sense? Okay and it should see an analogy of this with the rule that we had for the Bayesian networks, okay, so this is what we had for the Bayesian network, (Refer Slide Time: 13:17)

Bayesian network



Markov network

Local dependencies

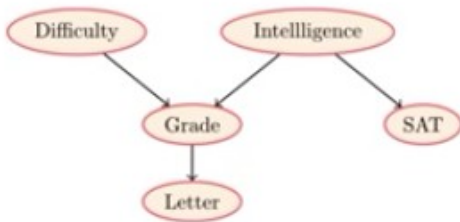
$$X_i \perp\!\!\!\perp \text{NonDescendants}_{X_i} \mid \text{Parent}_{X_i}^G$$



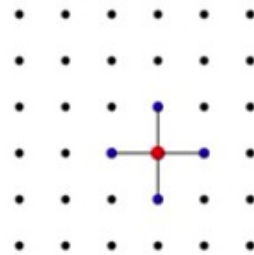
these were the local independencies again okay, we have fixed this.

Local independencies in the Bayesian network and these are the local independencies in the Markov network, this already fixed you don't need to note this, okay, is that fine?  
(Refer Slide Time: 13:32)

Bayesian network



Markov network



Local dependencies

$$X_i \perp\!\!\!\perp \text{NonDescendants}_{X_i} \mid \text{Parent}_{X_i}^G$$

Local dependencies

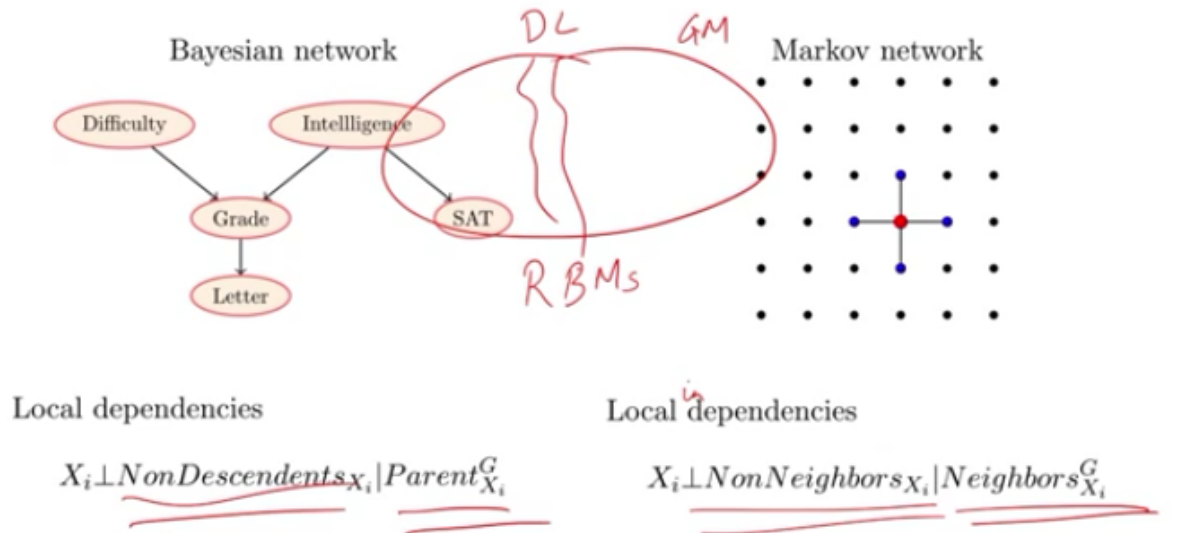
$$X_i \perp\!\!\!\perp \text{NonNeighbors}_{X_i} \mid \text{Neighbors}_{X_i}^G$$



So parents, neighbors, non-descendants, non-neighbors, was that fine? Okay, so that's the rule for Markov network, so this is I mean as I was explaining someone right, so we are in the course of deep learning, for deep learning we need some, we need to cover some topic known

as RBM's, for RBM's we needed this entire background of graphical models which is a separate course in itself.

What I've tried to do is whatever is the minimal stick part that I need to take through this jungle I have taken and from here we'll eventually try to reach RBM's, (Refer Slide Time: 14:10)



so that I'm trying to impress upon you is that graphical models even if you read one chapter from the book it is perhaps much, much more than what I have covered as background of directed models, or directed graphical models and directed graphical models and so on right, but my intention is not to do a course in graphical models, I've just done the minimal stick concepts that we need to eventually reach RBM's and from there reach auto and kudos and perhaps auto regressive models, right, so that's what we are aiming for.

And some minimal stick part I have taken already, and we'll continue probably exploring some more short parts in this jungle, and then eventually get to RBM's, hopefully by tomorrow or day after tomorrow, okay. So with that I'll just go to the next lecture.

### Online Editing and Post Production

Karthik  
 Ravichandran  
 Mohanarangan  
 Sribalaji  
 Komathi  
 Vignesh  
 Mahesh Kumar

## **Web-Studio Team**

Anitha  
Bharathi  
Catherine  
Clifford  
Deepthi  
Dhivya  
Divya  
Gayathri  
Gokulsekhar  
Halid  
Hemavathy  
Jagadeeshwaran  
Jayanthi  
Kamala  
Lakshmipriya  
Libin  
Madhu  
Maria Neeta  
Mohana  
Mohana Sundari  
Muralikrishnan  
Nivetha  
Parkavi  
Poonkuzhale  
Poornika  
Premkumar  
Ragavi  
Raja  
Renuka  
Saravanan  
Sathya  
Shirley  
Sorna  
Subhash  
Suriyaprakash  
Vinothini

## **Executive Producer**

Kannan Krishnamurthy

## **NPTEL Coordinator**

Prof. Andrew Thangaraj

Prof. Prathap Haridoss

**IIT Madras Production**

Funded by  
Department of Higher Education  
Ministry of Human Resource Development  
Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyright Reserved