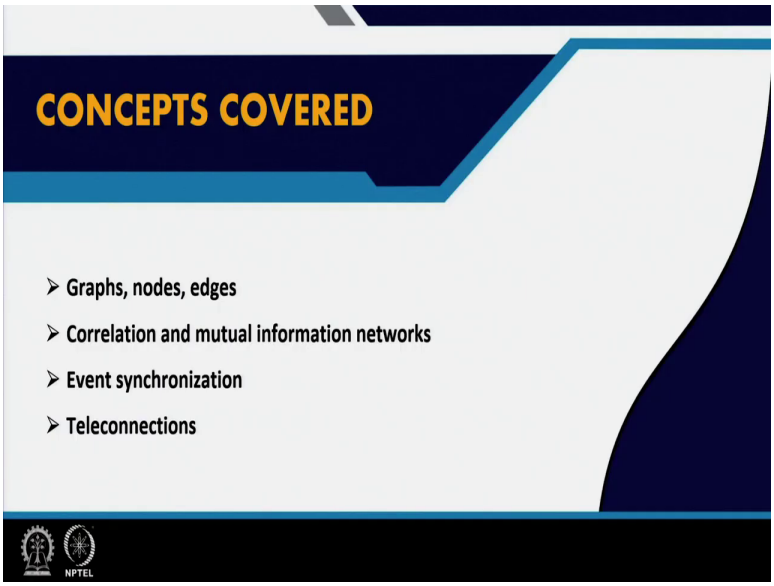
Machine Learning for Earth System Sciences

Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 01
Spatio - Temporal Statistics
Lecture - 09
Networks

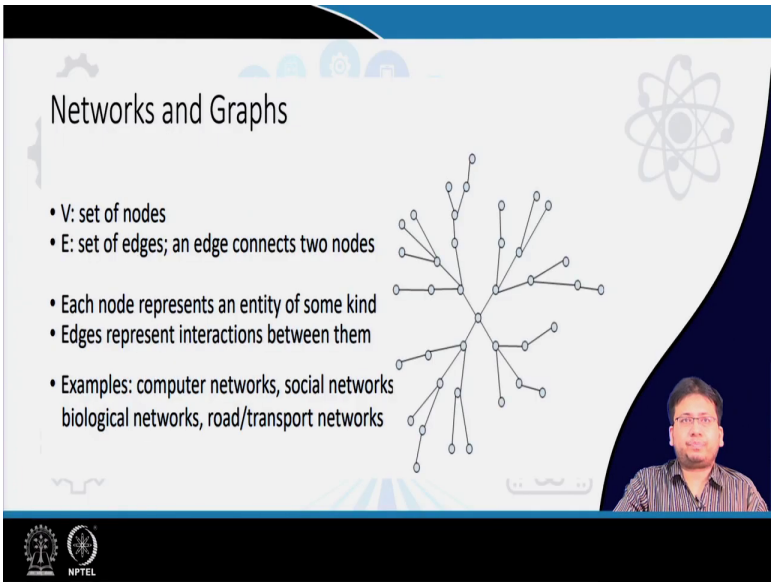
Hello everyone, welcome to lecture 9 of this course on Machine Learning for Earth System Science. We are still in Module 1 about Spatio Temporal Statistics and in today's lecture we are going to focus on the topic of Networks.

(Refer Slide Time: 00:41)



The concepts we will cover today are graphs, nodes and edges, correlation and mutual information networks, cvent synchronization and teleconnections. We will also see various small applications of these concepts in the domain of earth systems science especially in the domain of climate.

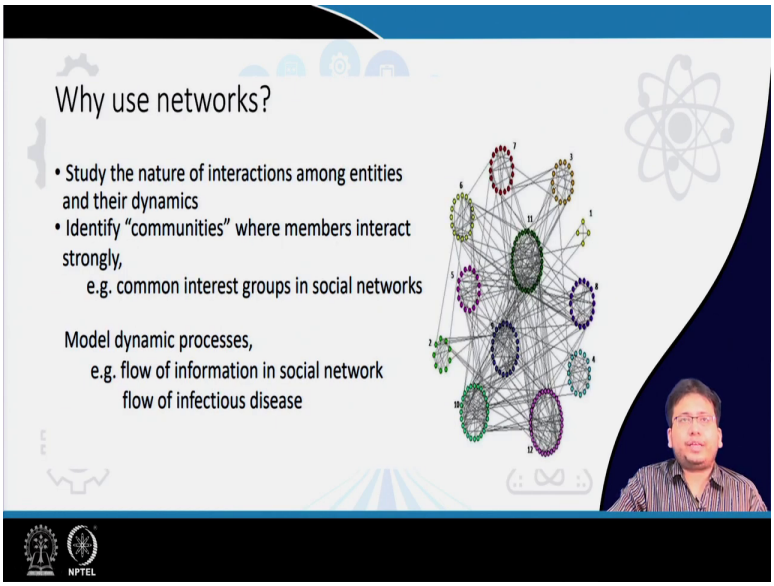
(Refer Slide Time: 00:59)



So, first of all the concept of network or graph most of you will be familiar with. So, it is a system where we have a set of nodes denoted by V and a set of edges denoted by E. Now each edge connects 2 of these nodes, each node can represent some kind of an entity and edges represent interactions between such entities. For example, we have like computer networks, social networks, biological or transport networks.

So, in a computer network each of these nodes will be different computers and 2 computers which are connected to each other or can exchange data among themselves will be connected by an edge. Similarly in a social network all these nodes will denote human beings and if 2 human beings are connected socially or if they know each other then we will put an edge between them.

(Refer Slide Time: 01:57)



So, the so why do we use networks? So, the study of the nature of interactions among the entities and their dynamics, this is one of the main motivations of studying networks. For example, we may want to identify a communities where the members interact strongly. So, if you look at this network, so you can see lots of nodes and edges among them, but here you can like identify certain communities.

So, like these communities these are like dense structure where almost every pair of node are is connected to each other, that is each node may have like may have some connections with nodes outside the community; but they are almost or they are connected to almost all members of the within that same community.

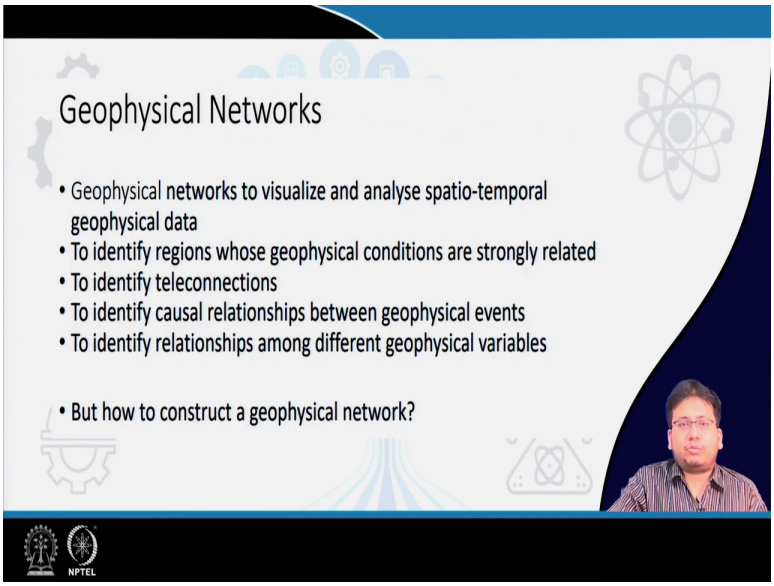
So, like for example, if we consider a social network these clusters of nodes which you can see here these densely connected clusters, like you can consider them to be some common interest groups or like other community based groups or family based groups whatever.

Where like the people know each other very closely with irrespective of whether they know, like other persons who are outside that network or not. So, this now if there is some kind of dynamical process like say flow of information in a social network. So, we know that in social media like Facebook if someone posts something that post can be shared by other people and the

friends of those people can see that post and then they can further share and then their friends can see the post and so on.

That is the way a post can become viral and so the so that is an example of a dynamic process in a social network. Now also on the basis of such flow of information it is possible that new edges might be established or old edges might get broken. So, the structure of the social network may keep changing over time. Flow of infectious disease is another thing to be studied using these kinds of networks.

(Refer Slide Time: 04:19)



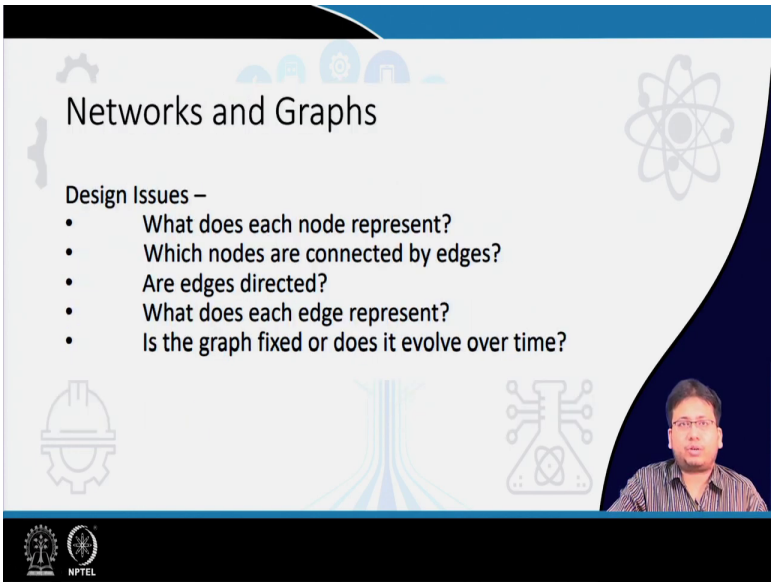
Now, coming to the domain of earth system science; so, geophysical networks are used to visualize and analyse spatio temporal geophysical data. So, the some of the applications might be to identify regions, where the geophysical conditions are strongly related. That is can we like identify homogeneous regions or spatial climate zones in different parts of the world or can like or like can we say that these is a zone of a particular soil type or a particular forest type etcetera or can we identify let us say seismic zones.

So, that is one kind of application; another example application is to identify teleconnections. So, teleconnections are like basically connections between geoscientific phenomena in different parts of the world. Say suppose something happens in Pacific Ocean and impact of that is felt in India

or in Europe. So, the like even though the locations are far away from each other, but one location impacts the other influences the other.

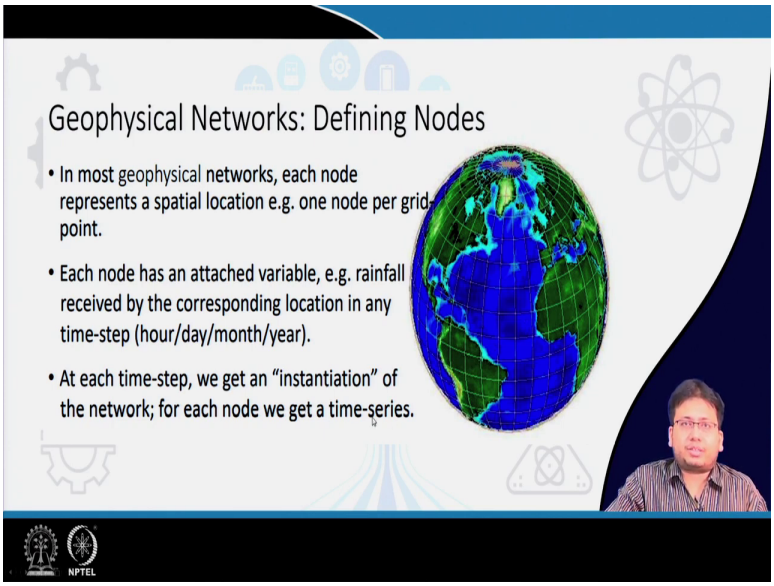
So, such long distance influences can be understood with the help of these networks or graphs. There are other possible usages like causal relationships between geophysical variables and so on. But the first question of course, is how do we construct such a geophysical network I mean what are the nodes and what are the edges?

(Refer Slide Time: 05:51)



So, these are of course, some of the design issues of networks.

(Refer Slide Time: 05:58)



So, in case in the case of earth system science the one standard way of defining these networks is where each node represents a particular spatial location. For example, 1 node per grid, so we can divide the earth surface into multiple grid points with the help of this latitude and longitudinal structures. So, as you can see like these 2 longitudes and 2 latitudes they divide the they enclose this region, so this is an example of a grid.

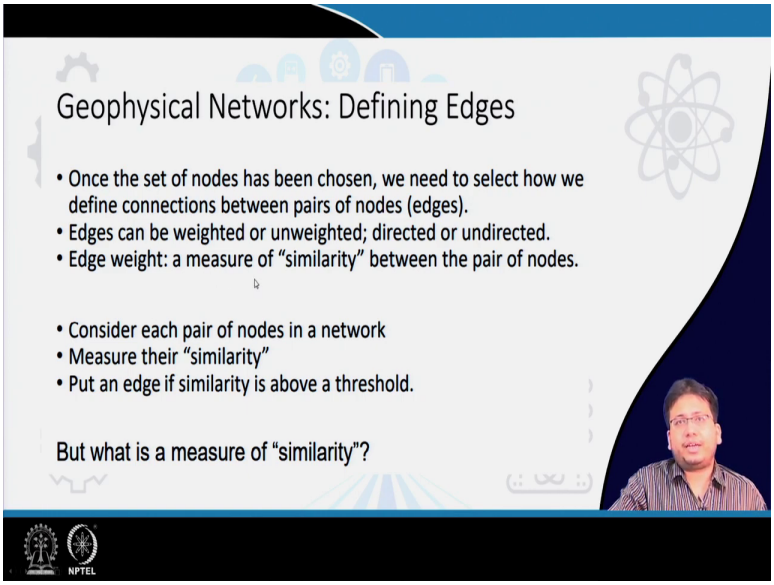
So, this grid can be treated as like as a location as a single location and it can be represented by a node. Similarly again this grid cell this again may corresponds to another node and so on. So, this way we get a collection of nodes all over the surface of the earth. Now if 2 nodes or 2 regions are close to each other or adjacent to each other we can connect them using an edge.

Also each node has an attached variable, for example rainfall received by the corresponding location in any time step, say this location or this location and so on. And at each time step we get an instantiation of the network where and that that means, at a any point of time all of these nodes they like they have a particular value, which is the value of the geophysical variable of interest in different parts of the world.

So, we say that it is an instantiation of a network where like each node has a has one like has some value of it is own. So, we can also call it as a snapshot of the network. Now these values of

course, keep changing over time at any other time point we can see some or all of those values have changed. Now, if we focus on a single node we can get a time series.

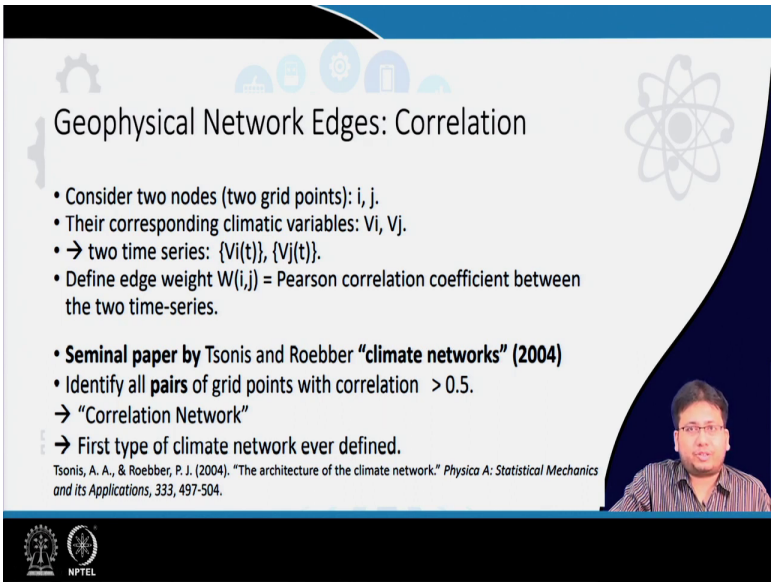
(Refer Slide Time: 08:05)



So, now this is how we define nodes, the next step is to define edges. So, once the set of nodes has been chosen the next step is we need to select how to define connections between pairs of nodes. Now edges can be either weighted or unweighted; either directed or undirected; this is something we know from graph theory. Now if they like if they are weighted nodes then each edge is associated with a real number, which in some sense is like the strength of the edge like if the strength is 0, then of course it basically means the edge is not present.

But in if it is non-zero then we can say that the edge weight in a sense is a measure of similarity between the pair of nodes. So, if the edge is absent that basically means the nodes at that corresponding 2 nodes are like they do not have sufficient similarity. Now consider each pair of nodes in a network measure the similarity and we can put an edge if the similarity is above the threshold. This is the standard way in which we define the network structure.

(Refer Slide Time: 09:19)



But the question is what is a measure of similarity? So, one possibility is correlation, so consider 2 nodes which may be 2 grid points i and j and their corresponding climatic variables are V_i and V_j. So, we get 2 time series which are denoted like this $V_i(t)$ and $V_j(t)$. So, like as you can see they are like expresses time series.

Now, we define the edge weight between the corresponding grid points i and j, let us denote the edge weight by W. So, in this kind of network which we call as the correlation network the similar as I said the edge weight is basically a measure of similarity between the 2 between those 2 grid points and that in this case we will consider as the simply the Pearson correlation coefficients between the 2 time series that is at i you have this time series V_i.

And at j you have this time series V_j you can calculate the correlation coefficient between them. Which is of course, a quantity that will lie between (+1) and (-1). Now (+1) means they are perfectly correlated and (-1) means perfectly anti correlated that is whenever one increases the other decreases and so on.

So, now on the basis of this how to construct the edges so, one possibility is that we simply have a weighted network where like every edge we will carry this value $W(i,j)$ as the corresponding

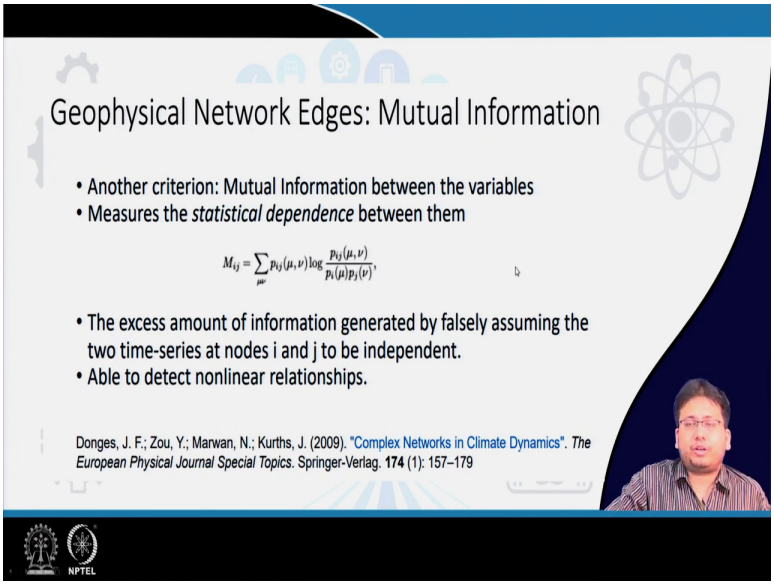
edge weight. So, that will be a very dense network that is the correlation coefficient will rarely be 0. So, in most cases it will have some value big or small. So, one alternative is to make it make the network sparse by considering only those pairs of points between which the correlation is above 0.5.

So, only the edges which edge strength above 0.5 will remain the other stages will just be dropped off. So, this is the of course this threshold of 0.5 this is something that was chosen by this seminal paper which appeared in 2004 called climate networks. But there is nothing special about 0.5 it could have been 0.8 or 0.2 or something else also.

But the basic idea is that instead of considering a dense network, where all edges have some weight associated with them we keep only those edges which have a particular weight which is high enough or low enough, it like we may also be interested in identifying anti correlation. So, in that case we may also want to identify those edges whose weight is say less than minus 0.5 or something like that that is also possible.

So, this is probably the first type of like geophysical network that is known to have been defined that was way back in 2004.

(Refer Slide Time: 12:17)



And now just extending the concept of correlation there is another possibility is to construct the edges on the basis of what is known as mutual information. So, mutual information it is basically it measures the statistical dependence between 2 variables. So, once again consider the 2 locations i and j. So, let us say that we are measuring the quantities μ and ν at i and j respectively.

So, the question is whether these μ like. So, now, we can consider these as random variables that is the observations at the 2 locations $i$and j we can just like we have been saying earlier we can consider those to be random variables and we can like see the relation between the joint distribution of μ and ν vis a vis the marginal distributions of μ and ν.

And we know that if they are independent then this the joint distribution can be expressed as the product of the marginal distributions. So, if that is the if they are indeed independent then this thing this fraction will become 1 and since we have taken the log of it will the mutual information will simply become as 0. But if that is not the case then we can like that is then this will be some non-zero value and accordingly we will have some value of the mutual information.

Now, in general it can be shown that 2 like when the variables are like they are strongly related to each other, then the mutual information will take some like some non-zero value and if they are independent of course, as we saw then their mutual information will be 0. So, in a sense it is like the excess amount of information generated by falsely assuming that the nodes at i and j are independent of each other.

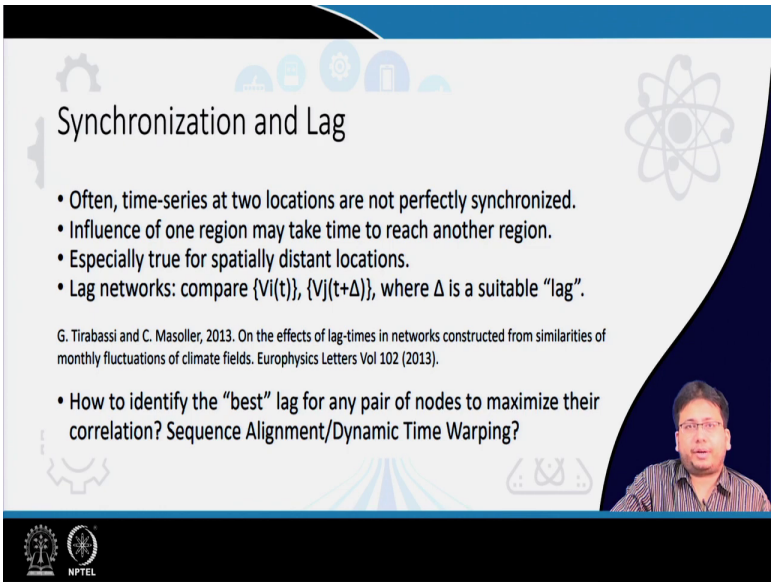
So, like this is so this mutual information this is another measure of similarity between 2 nodes which have a time series attached with them and this is also is able to detect the non-linear relationships. However, unlike the previous case where we are considering the in the correlation coefficient, when we are dealing with this mutual information like we are not explicitly considering the time series as such.

We are like not considering like which value follows which value and so on. We are simply seeing which values of μ and ν can co-occur with each other. So, in a sense we have discarded the temporal information when we are going for the mutual information network. So, this kind of

mutual information network is may not be that suitable where if we are doing a time series based application.

I mean to say we like that is if we have such an application where it is whether similarity between 2 nodes should be defined in terms of the time series, in that case this mutual information network may not be that suitable. On the other hand it has a like additional advantage over correlation network namely that it is able to detect non-linear relations ok.

(Refer Slide Time: 15:53)



So, the this kind of network has also been studied in the domain of earth sciences, specifically climate. Now, it often happens that the time series at 2 locations are not perfectly synchronized. Now why that happens is because the influence of one region may take time to reach another region especially if the locations are far apart from each other.

So, suppose we are finding the relationship between or we are trying to quantify the relationship between say the sea surface temperature in the Pacific Ocean which is sometimes known as the Enoso the Elnino Southern oscillation index and that of monsoon rainfall in India.

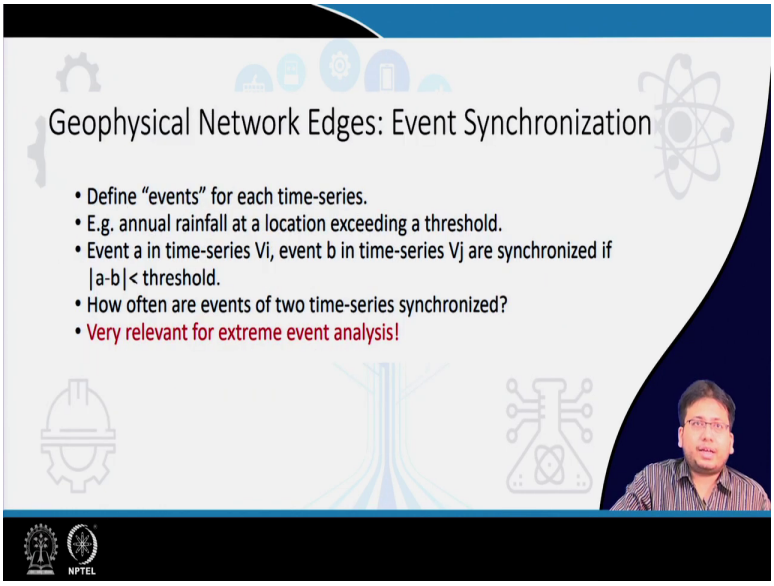
And it is generally known to climate scientist that if the pacific the central part of the Pacific Ocean if it gets heated up in given year, then that year India often experiences a drought that is less rainfall during monsoon. But there is a lag between the heating of the Pacific Ocean and the

deficient monsoon in India. Usually like if the heating of the Pacific Ocean happens say around the time of March and April; the deficiency is felt in the Indian region after 2, 3 months, like we know that the monsoon period in India is from June to September. So, like we can see a 3 to 4 month lag between the heating of the pacific and the rainfall in India. That is because pacific and India are far apart from each other and the influence takes some time to travel.

So, if we are trying to see the correlation between; or if you are trying to study the interactions between these 2 quantities with the networks, then we should not be considering the Pacific Ocean temperature in 1 month vis a vis the rainfall over India in the same month that will not make much sense. What we should be considering is Pacific Ocean temperature in one month vis a vis is the rainfall over India after 3 or 4 months. So, here what we should be considering is a lag, which we may denote with Δ.

So, in a normal correlation network or like or mutual information network we may be comparing $V_i(t)$ with $V_i(t)$, but in the presence of lag we should be comparing $V_i(t)$ with $V_i(t + \Delta)$ ok. So, now what is should be the ideal value of Δ; what value of lag should we consider? So, how to identify the best lag for any pair of nodes to maximize their correlation? So, there are like we can use various algorithms like sequence alignment, dynamic time warping and so on.

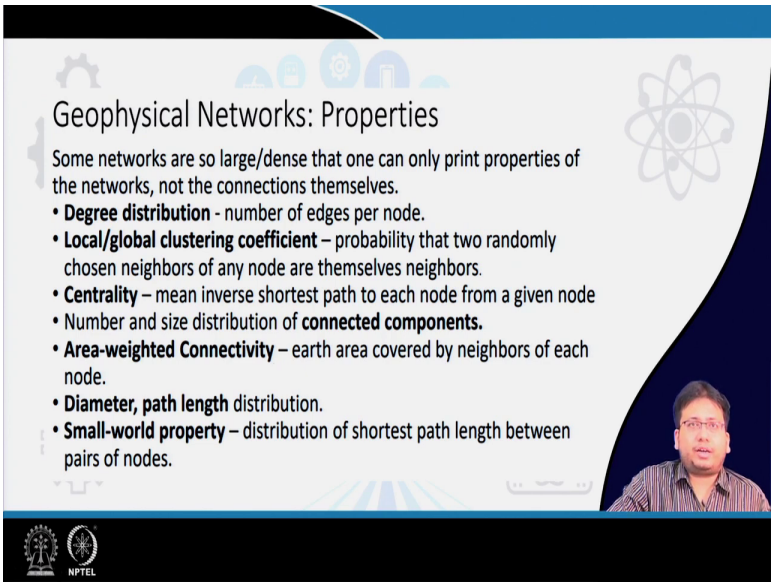
(Refer Slide Time: 18:47)



Another thing which is used in is that of event synchronization. So, we may define events for each time series. Like for example, like if we are having a rainfall time series we may consider those points or those time points where the value of the observation exceeds a particular threshold. So, event a in time series V_i and event b of time series V_j, we say that these 2 events are synchronized, if the like if the temporal difference between them is within a threshold.

So, then the question might arise like how often are the events of 2 times series synchronized. So, this is a very relevant for extreme event analysis.

(Refer Slide Time: 19:36)



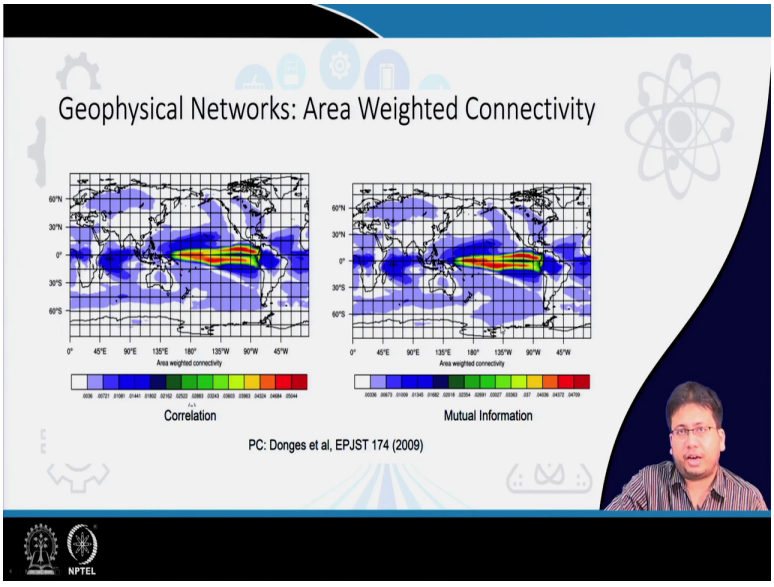
Now, when we have a geophysical network like this we may also be interested in various properties of the network. Say for example, the degree distribution that is the number of edges per node or the local or global clustering coefficients.

So, this is basically to see the like whether some kind of community structure is present in the network or not. So, like they the clustering as the name suggests indicates whether different nodes are like or neighboring nodes are densely connected with each other or not. Specifically it is the probability that 2 randomly chosen neighbor's of any node are themselves neighbor's ok.

So, then next there is the concept of centrality which means that like are there some central nodes in the network, are there some nodes which are like linked to most other nodes. Then there are other things like a number of connected components can we like does the whole network operate as one entity or like are they are like parts of the network which are disconnected from each other and they form what is known as a components.

Then there are other things like area weighted connectivity, diameter and so on and an interesting property of networks is small world property, which is like the distribution of the shortest path lengths between pairs of nodes. So, its small world network is such a network where the number of nodes may be very high, but if you take choose any 2 nodes at random the shortest path between them the length of the shortest path between them, Or if you want to say the number of minimum number of hops required to go from one of one node to another will not be more than small number, let us say 3 or 4 or something like that. So, that is called a small world property. So, these are various standard properties which we study in any kind of network whether it is a like a geophysical network or not.

(Refer Slide Time: 21:47)

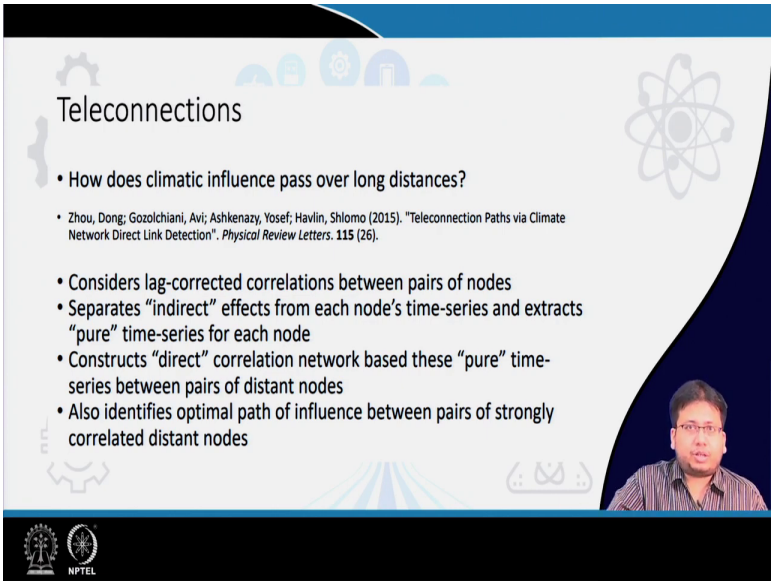


So now, these like in the domain of earth science many of the concepts discussed above they have been used.

So, like for example, here we are we have constructed 2 networks which one is the correlation network, the other is a mutual information network we have of course, discussed these quantities already. So, here the what these colors indicate is the like the area weighted connectivity. So, like here these long red or stripes that you are seeing these indicate that in a sense these are some kind of clusters or homogeneous regions which like that is which influence things or events in different parts of the world.

That is like these are regions which are connected to so many other places in the network in the whole climate network. Say like similarly these deep blue edges these like for them I mean they are also connected to many other regions of the world, but probably less strongly than this region. So, this central Pacific Ocean it so this what this analysis shows us is that this central Pacific Ocean this is a very significant or very sensitive region and whatever happens there may impact like climatic conditions at various other parts of the world.

(Refer Slide Time: 23:14)



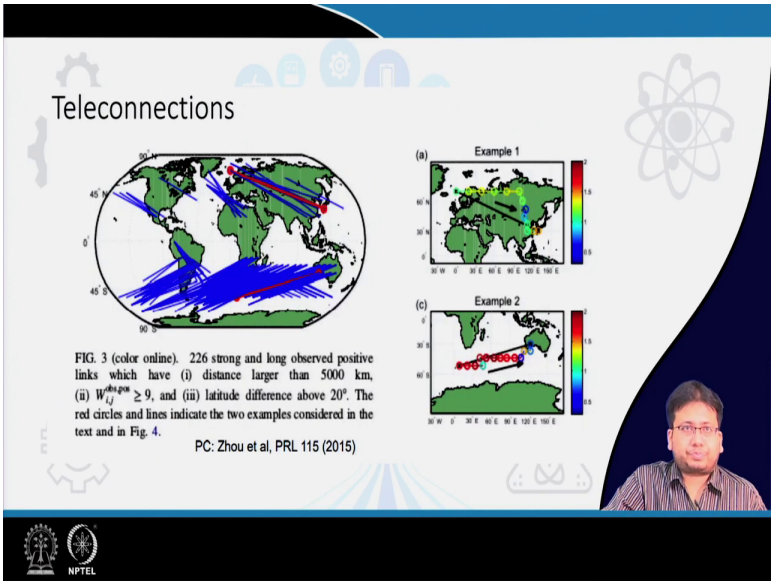
So, that brings us to the concept of teleconnections. So, the question is how does climatic or other I mean other scientific influence pass over long distances? So, here we may consider like a

lag corrected correlations between different pairs of nodes and separate indirect effects from the nodes time series and extract the pure time series.

So, that is we are like so indirect effects is like basically when the one the I mean a variable X at one location influences a variable Y at another location j through a set or a sequence of intermediate steps at I mean other variables at other locations. It is like saying that like variable X at location i influences variable Z at location k which in turn influences let us say variable W at location $m$and finally that influences variable Y at location j.

So, that is an indirect relation, but it is possible to so that is definitely. So, identifying these kinds of indirect pathways is one thing. The other thing is to identify the direct relations between X like X at location i and Y at location j. So, teleconnection is the study of teleconnection is to compare both of these things that is to see whether the influence is a direct or through a sequence of intermediate influences like this ok.

(Refer Slide Time: 25:04)



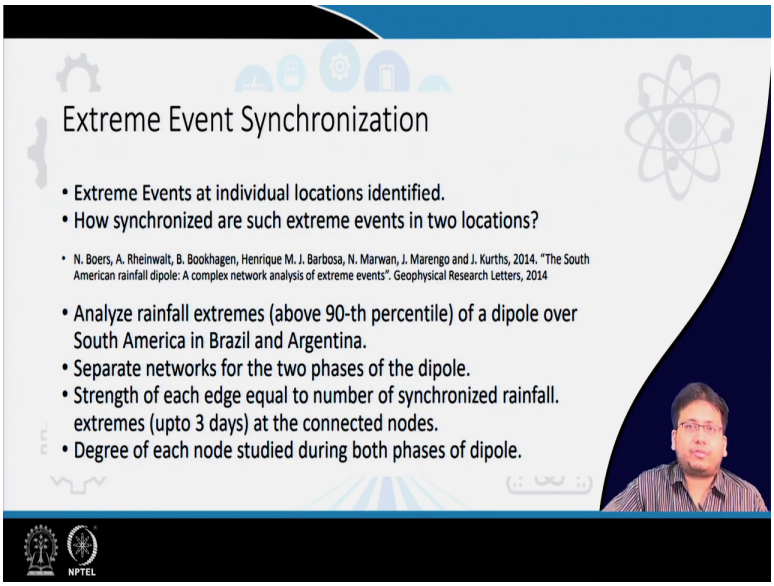
So, like identification of teleconnection patterns in different parts of the world and to like if they are found to identify the nature of the pathways causing them.

This is another like another set of or another body of research, where this concept of networks has been heavily identified. So, here like so this is a paper which appeared in 2015. So, these

edges which you are seeing here these are basically the teleconnection edges as you can see these are all long edges, which are connecting locations at far off places and so the blue and red they indicate whether it is a positive relation or a negative relation.

On the other hand, here like we are actually looking for the kind of indirect relation that I that we just talked about. So, here we see that from here to here there is one edge that is suggesting some kind of a like the teleconnection. But we may be interested to understand that is what is the optimal path along, which the influence study travel. So, that analysis seems to be a this seems to be the like that kind of an optimal path ok.

(Refer Slide Time: 26:24)



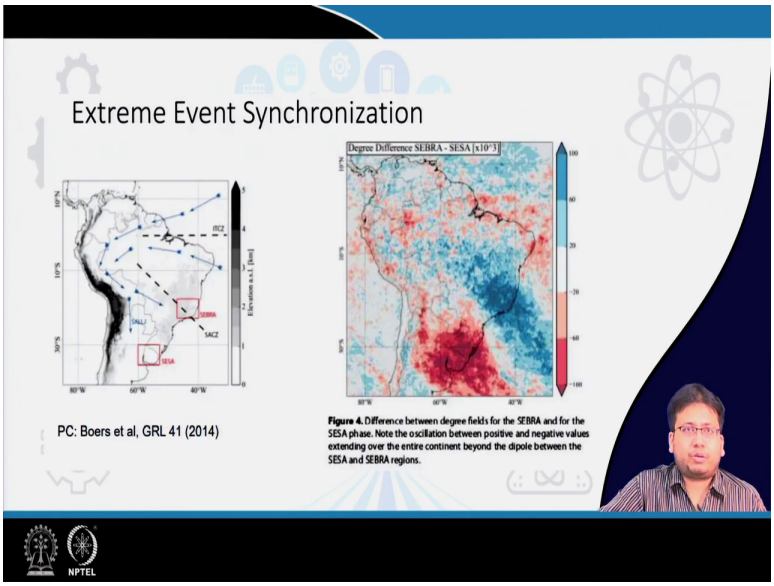
So, now there another interesting applications of climate of these networks in the domain especially of climate is that of extreme event synchronization. So, we have talked about event synchronization earlier, that is, we have said that we like we will connect 2 nodes by an edge, if the corresponding time series appear to be event synchronized. That is whenever one event some kind of event happens in one time series, the other time series will also have an event in the like in a at a nearby like at a short time interval.

Now, extreme events is something which we have already studied in this in a previous lecture. So, we know that a rainfall extremes and things like that we know we now know how to define

these rainfall extremes. And now the question is can we identify sets of locations where extreme rainfall events are synchronized. That is we can say that if our rainfall extreme rainfall happens in one location then within 2 days or 3 days another extreme rainfall event will happen at another location.

I mean for nearby locations of course, this is this might be trivial as we have already seen the extreme events tend to be specially clustered. But suppose we are considering locations that are far off from each other, then it is not very straight forward to see that extreme in one location will also happen will impact the extreme in the other location. But such things also happen and they can be identified with this kind of networks ok.

(Refer Slide Time: 28:08)

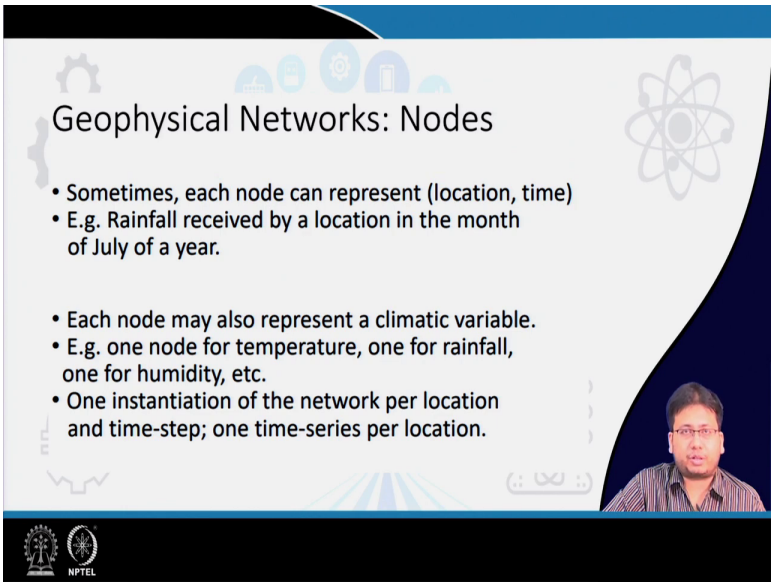


So, like this is a study of such a thing where like in South America, like they studied the oscillations of some kind of rainfall dipole and it was found that that is when this like when these regions are having say high rainfall then these regions are having less and vice versa. So, like the here actually the, these blue and red they indicate the in some sense the positive and negative values.

And like so this map basically explains the what is known as the degree differences, that is the by a degree we know that it means the number of edges. So, here we can see that there is a very like

it is like a dipole kind of structure which has been created. So, like there is one region where the like that is the degree difference is like very high and in here it is like something like a very low.

(Refer Slide Time: 29:22)



So, there are of course, like there are many other examples in which geophysical networks can be defined and the nodes which we mentioned they also can be they need not be correspond to locations as we have been saying so far. But they can be they can mean something else also. For example, a node may represent a climatic variable. So say for example, one node for temperature another node for rainfall etcetera, one for humidity and so on.

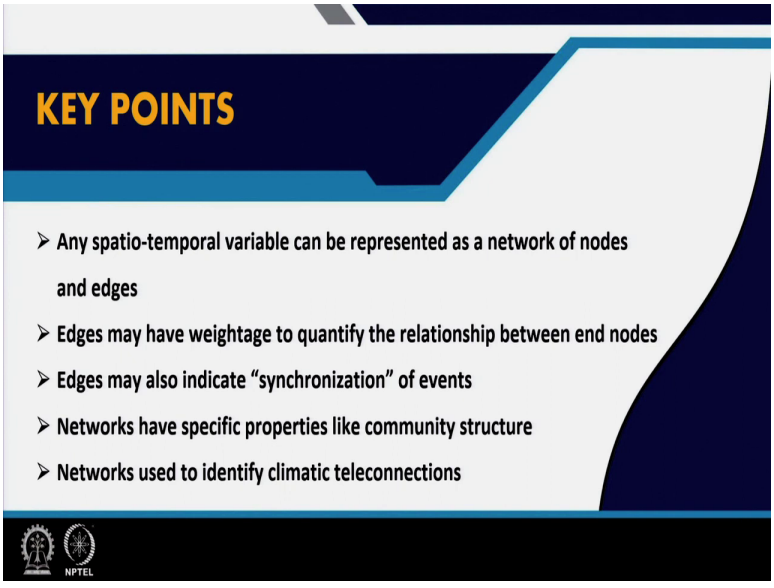
And like one instantiation of the network per location like we will get per location and per time step. So, that like at any given point of time the values stored in the different variables like they form something like the snapshot or instantiation of the network and of course the values keep varying over time. So, like note that the concept of location is not present here.

So, we are assuming that we are doing everything for the same location. Alternatively we can have multiple parallel networks which I each of which are corresponding to one location and where and in each of the network the nodes corresponding to the different climatic variables. And then there can be some hyper edges across the different networks that is also possible.

(Refer Slide Time: 30:41)

And so there are lots of references in the domain of art system sciences which deal with the which deal with this concept of networks.

(Refer Slide Time: 30:49)



Like we have already seen some of those discuss some of those references.

So, the key points to be taken away are any spatio temporal variable can be represented as a network of nodes and edges. The edges may have some weightage to quantify the relationships between the end nodes, the edges may also indicate the synchronization of events. The networks have specific properties like the community structure, which we discussed and they may be used to identify teleconnections or like synchronization of extreme events and things like that.

So, that brings us to the end of this lecture, so that this is the lecture 9. So, after this we will have just one lecture next on the left in this module after, which we will move to the next module. So, till then bye bye.