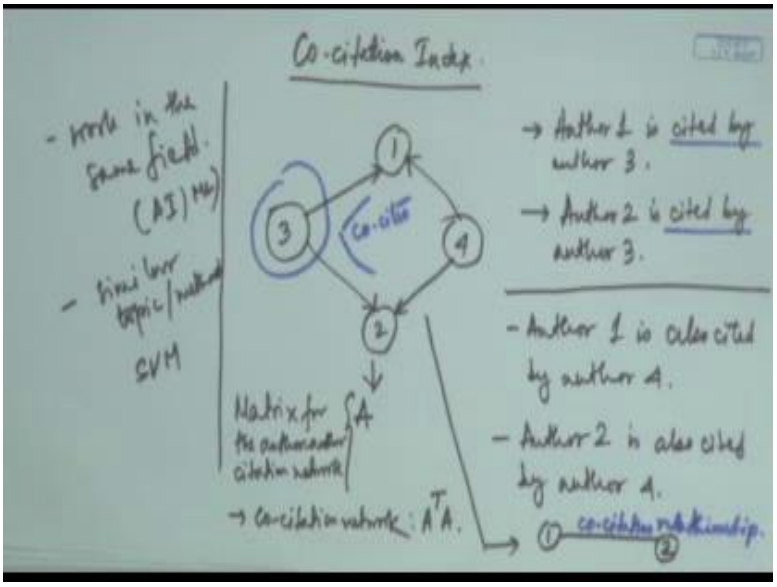
Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 07
Network Analysis – VI

Welcome back to this session on Network Analysis. We will continue with the ideas of network analysis that we have been talking for last few days. Today, I will introduce two important metrics that are very useful in the context of directed networks, and especially in the context of citation networks. The first one among this is called Co-citation Index.

(Refer Slide Time: 00:48)



The idea is very simple that when you look at a author author citation network, the type of network that we have discussing already in the context of degree distribution studies. If you look at an author author citation network what you observe is that, there are at times where two authors who are not connective by a citation relation among themselves have some relationship in an indirect manner and that is what we want to study here and that is what we quantify through this metric.

For instance consider this small hypothetical network here. From this network you immediately see that author 1 is cited by author 3. And similarly author 2 is cited by author 3, so mark this word cited by. Basically, there is this author three who actually co cites both author 1 and author 2. You even get a stronger evidence of such co-citation

behavior when you look at author 4. We find that author 1 is also cited by author 4 and author 2 is also cited by author 4.

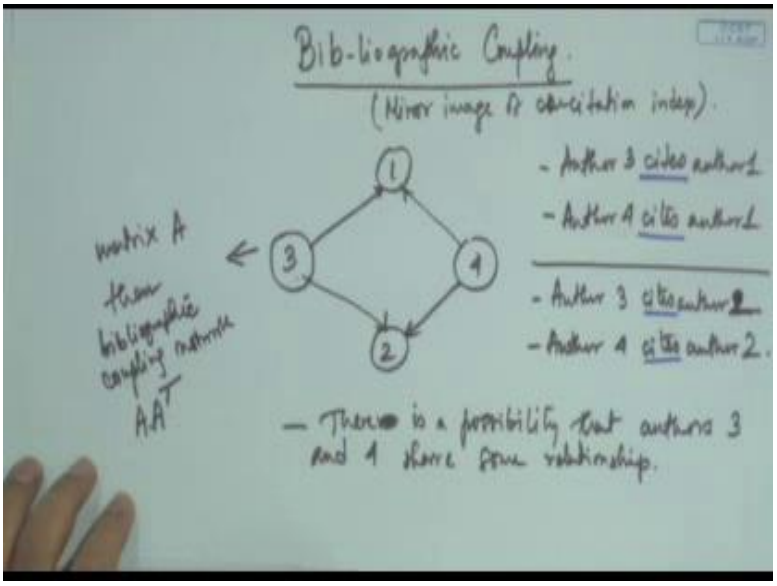
This gives us additional evidence that although author 1 and author 2 are not connected by any citation relation they perhaps have a relationship between themselves. And that relationship could be many things, could be in the form of that both author 1 and author 2 works in the same field say for instance you can imagine that probably both author 1 and author 2 works in AI or machine learning. Also it might be the case that author 1 and author 2 work on similar topic or method.

For example, say SVM in the area of machine learning. So, such hidden relationships between pairs of authors or a group of authors might exist in an author author citation network. And unless you have design appropriate construction it is not possible to tear through these relationships. And such constructions are not very difficult to formulate. For instance, let us consider that this author author citation be named as A. So let the matrix for the author author citation network this be denoted by the matrix A.

Then the co-citation network is nothing but simply the product of the two matrices A^TA. If you construct the A^TA of this graph then you get a new graph which would look like there will be only two nodes 1 and 2, and there will be an edge between them indicating a co-citation relationship.

Now imagine that there are many such intermediate nodes like, node 3 and node 4 in this example, then the evidence that the nodes 1 and 2 shares some relationships becomes stronger and stronger. Then more the number of intermediate nodes like node 3 and node 4 the stronger is your evidence that there is some form of relationship that exists between node 1 and node 2.

(Refer Slide Time: 06:32)



Now, likewise there can be a symmetric or mirror image concept which is called the Bibliographic Coupling. So this is as I say is just the mirror image of co-citation index. To understand the concept let us again draw the same hypothetical citation network example. It is only in the difference of how you (Refer Time: 07:23) the citation relationship. In this case what you see that, author 3 cites author 1. The earlier relationship that we are talking about was cited by and here we are more concentrating on the relationship cits. Similarly, author 4 cites author 1.

Again you have second level evidence a stronger evident make making your observations stronger, where you see that author 3 cites author 1 sorry author 2. Similarly, author 4 cites author 2. The earlier case was considering the cited by relationship, the current case is by considering cites relationships. So, author 3 cites author 1 and author 4 also cites author 1, similarly author 3 cites author 2 and author 4 also cites author 2. This indicates that there is a possibility that authors 3 and 4 shares some relationship.

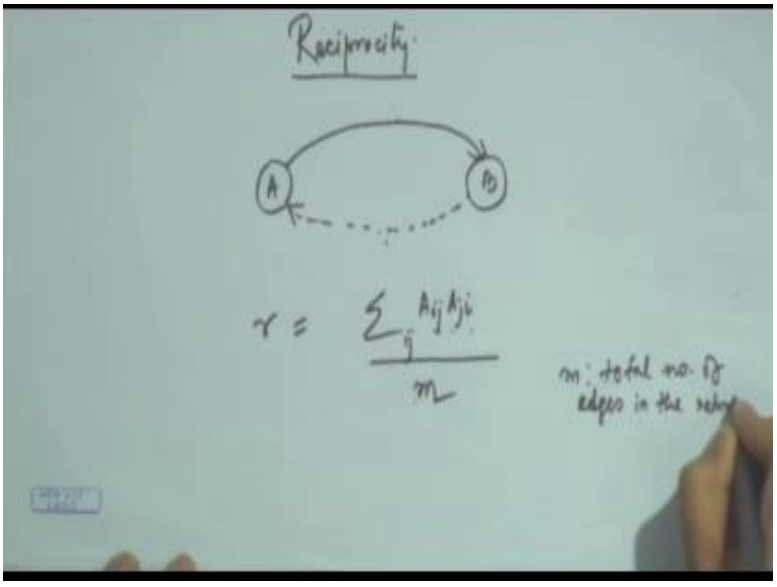
So, again extending the same idea as we talked about in the cite co-citation index context. Basically here you are no longer considering author 1 and 2, but authors 3 and 4 who have the similar citation behavior who have similar patterns in their reference list. The similar referencing behavior, you can also call it as a referencing behavior. The way author 3 refers to papers, author 4 also refers to papers in a similar way. Again the idea is

that probably these two authors work on a similar field or on a similar topic that is why very often they have to cite other similar people. So, there could be a relationship in that way existing between authors 3 and 4. This kind of a relationship is called the Bibliographic Coupling.

Again if we consider that the corresponding matrix for this network is a then for this citation author, citation network is A then the bibliographic coupling network can be obtained just by the product of the two matrices but now written like $A.A^T$. Whereas, the co-citation index matrix could be obtained by $A^T.A$, here the bibliographic coupling matrix can be obtained by $A.A^T$.

Actually these two networks are used in conjunction with the original author citation network matrix to do various sorts of recommendation task, citation recommendation, authorship recommendation, and various other recommendation tasks where people heavily use higher techniques but then at the back end some of the matrix that are used some of the quantification that go on are basically drawn from these three networks A A A^T and $A^T A$.

(Refer Slide Time: 11:58)



So, continuing with other matrix we will now introduce another very interesting concept for directed networks and this is called Reciprocity. The idea is very simple, suppose there is a node A there is another node B in the network there is a directed edge from A to B. Now as the name suggest reciprocal means if there is a link from A to B there is a

reciprocal link from B to A. In directed networks there might be a link back from B to A; this is a very different link from the link from A to B in the context of directed networks.

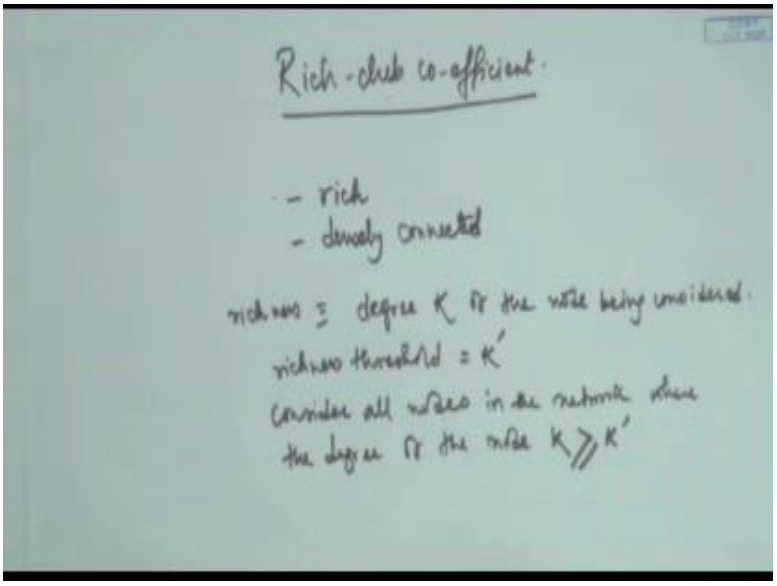
So if there is a link from A to B, then if there exists another link from B to A then we say that this is a reciprocation of the original link A to B. The reciprocation of the original link A to B is the link B to A. And the associated quantification is called Reciprocity. So the idea is very simple. So quantification is quantification of reciprocity which you define by the metric r.

So if you have m edges in the network, the total number of m directed edges in the network among this how many edges are actually reciprocated, that is the idea. If there are say n nodes in the network and there are some of the edges which are reciprocated and then there are some of the edges which are not reciprocated. So, you express the total reciprocation as a ratio of the total number of edges in the network. So that is how reciprocity is defined.

Basically, it is in the directed network A_{ij} A_{ji}, so there is one link from i j and there is another link from j i. If A_{ij} is 1 and also A_{ji} is 1 and this product will become 1. Otherwise, in all other cases this product will be 0. If there is 1 and 0, 0 or 1 in both the cases this product will be 0, only in the case where both of them are 1 that this product will be 1 and this is expressed as a fraction of the total number of edges in the network. So, m is the total number of edges in the network. So this is the how you define the reciprocity of a graph.

So, the next concept that we will talk about is called the Rich-club co-efficient.

(Refer Slide Time: 15:07)

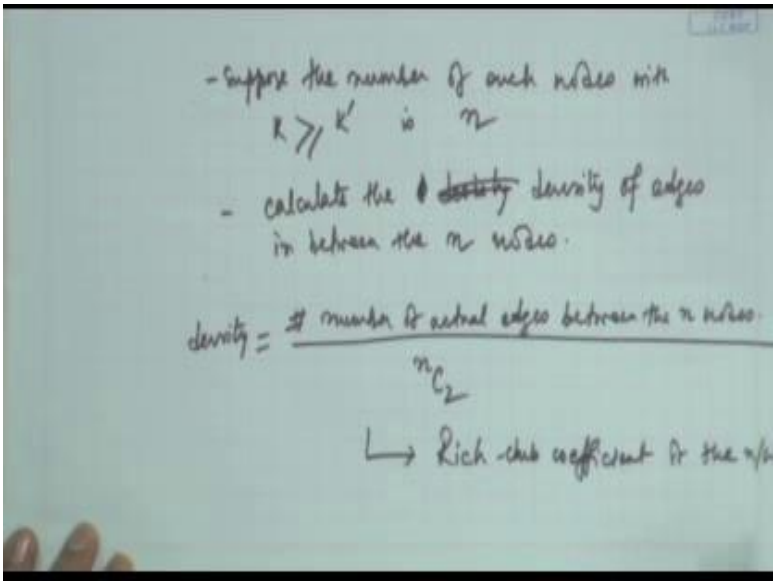


This is another very interesting phenomenon. For the time being we will only consider undirected graphs, unweighted graphs for estimating the rich-club co-efficient. The idea is very simple that given a social network or say given a collaboration network of scientist do you find a group of people or group of scientist or a group of nodes in the social network who are in some sense rich and actually form a dense connectivity, actually establish a dense connectivity among themselves. So they have to be rich and they have to be densely connected. These are the two constraints that this set of nodes need to satisfy.

Basically, how to quantify richness? So richness in the context of rich club co-efficient can be quantified in different ways but the most basic one is to assume the degree of the node. If the degree of the node is above a threshold say k then you consider this node as a rich node. So richness is in this case analogous to the degree say k of the node being considered.

Suppose there is a richness threshold, say richness threshold is said to some k' then you consider all nodes in the network where the degree of the node k is greater than or equal to the richness threshold k'. So you consider all the set of nodes in this set where the degree of the node is greater than the richness threshold k' or k dash.

(Refer Slide Time: 17:49)



Suppose the number of such nodes with k greater than equal to k' is sum n nodes. Now you consider this set of n nodes and you see what is the density calculate the edge density or the density of the edges in between the n nodes. So, you calculate the density of the edges in between this n nodes which are actually the high degree nodes or nodes crossing the richness threshold bar of k'.

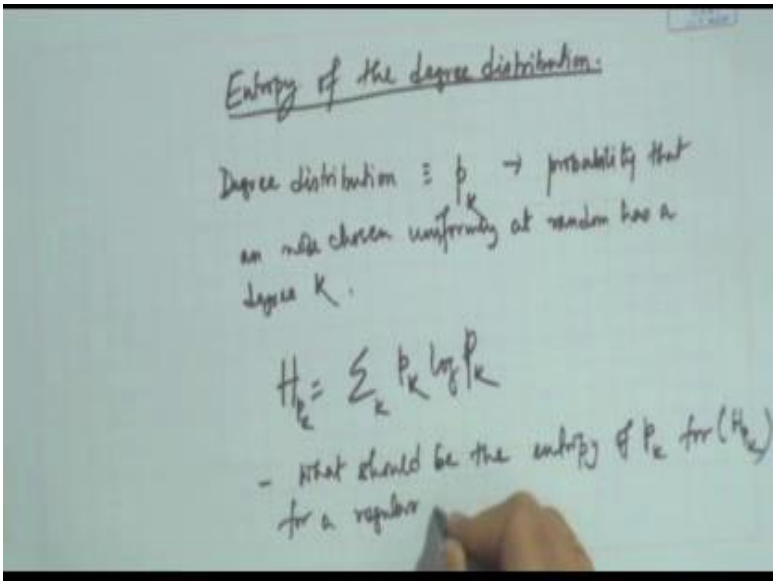
So that, how will you calculate the density? It is very simple. The density is nothing but the number of actual edges between the n nodes divided by the maximum number of edges which is possible between these n nodes. The maximum number of edges possible is $n C_2$. Whereas, you express the actual number of edges between these pairs of n nodes as the ratio of the maximum number of edges that is possible between these n nodes. And this is actually the rich-club co-efficient of the network being considered.

As soon as I talk about this you can very well imagine that the rich-club co-efficient is nothing but trying to express whether the nodes that are rich in terms of degree have a very high number of connections among themselves. For instance, take for the example the actor actor network from the movie actor example that I have cited earlier. So the actor actor network here, if you find a rich-club this would mean that the most influential actors packed together from a (Refer Time: 20:42) with each other from a group with each other and sign movies together so that the next movie release is really a box office hit.

Similarly, there are quite often we have found that in the scientific domain there are a bunch of scientists who are really very highly cited and very highly popular scientist they come together and do something very impactful something very seminal. So, whenever such a seminal thing happens there are a bunch of highly well known, well establishes scientists they come together form a rich-club and by the virtue of forming this rich club publish something very seminal.

Like for example, in the current last one or two years context one of the examples would be this hadron collider and the idea of got particles which you have probably come across in various new articles.

(Refer Slide Time: 21:57)



Now, the next concept that we will talk about is the Entropy of the Degree Distribution. So this is another interesting idea, very simply put. If you recollect the degree distribution is encoded in the variable p_k, which is nothing but the probability that a node chosen uniformly at random has a degree k. So, this was the very simple idea of degree distribution. Since, this is a probability distribution one can always estimate the entropy of this distribution. So, simply put the entropy H is nothing but $\Sigma_k p_k \log p_k$; and as you can well imagine that entropy actually encodes the randomness, the extent of randomness in a distribution.

f this quantity is high then the distribution is relatively random and the network structure is also relatively more uniform, whereas if it is queued, if it is lower than probably it is

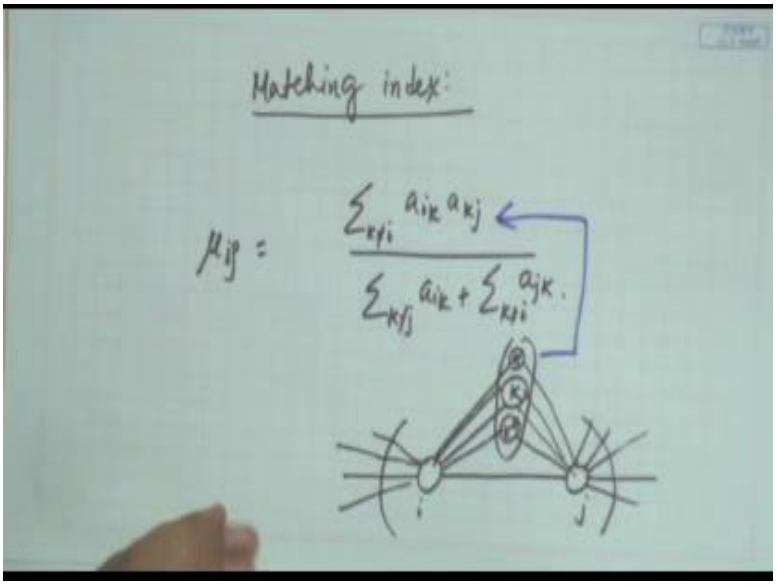
not the case. Some interesting exercises could be like what should be the entropy of p k for, let us call this H_{pk} , the entropy of p_k that is H_{pk} for a regular graph. A regular graph is a graph in which all nodes have the same degree. So, a (Refer Time: 24:42) is also a regular graph where each node has a degree n - 1, where n is the total number of nodes in the networks.

Similarly, if a network each node has a degree k then that network is called a k regular graph. Basically, in such a system, in such a network what do you have p_k is 1 for that particular value of k for which you are constructing the regular graph and p_k is 0 for all other values of k. In such a case what should be the entropy, it is very simple to find. So, you can easily compute the entropy. For that particular value of k there will be it is 1, p_k is 1 1 log 1 plus the rest is there is nothing.

Such a network does not encode any diversity that is what is been actually talked about. So in such a network you do not observe any diversity, all nodes have similar degree. So that is why the entropy goes to 0. Whereas, if there is a network where this degree (Refer Time: 25:55) then the entropy is more close to non-zero values. So this again gives you an idea of the topological structure of the underline social network.

So, one is the degree distribution itself and on top of it you can also measure the entropy of this distribution to understand to get an idea better feel. Suppose, if the entropy of such a network is close to 0, then you immediately get to know that this a more sort of a regular structure; whereas, if this entropy is not close to 0 then it is a more non-trivial structure that is there in the social network.

(Refer Slide Time: 26:54)



There is this last concept that we will get introduce to which is called the Matching Index. Basically, what this matching index tries to say is tries to quantify is that how similar are two nodes in terms of their connectivity patterns. The matching index μ_{ij} is actually expressed using the following formula. As soon as I write the formula it becomes very clear to you. So what we are trying to see, suppose there are two nodes i and j in the network and there is some other node k sitting out here, and we assume that there is already an edge between i and j.

Basically, if there is a lot of connections like k, see there is some k', see there is some k'', all of these actually connect i and j. So these k' k'' k are the number of nodes that is being measured by the numerator of this formula a_{ik} a_{kj}. So, basically you are trying to count how many such triangular shapes or triangular completion exits between i and j and that you express as a fraction of the sum of the degrees of i and j.

Basically, this gives you an idea of the balance. If there are two nodes i and j and there is an edge between them and if all other connections all other intermediate nodes between i and j there is a connection from actually i to k and k to j, if the structure is like that there is no other node than this set of k values then the matching index is maximum. Now if there are many other nodes.

Basically it might be heavy on i, there degree i might have a very high degree similarly j might have a very high degree, but the overlapping set of nodes is not so high then the

matching index is low. So, basically you see how well is the match between the neighborhood of i and the neighborhood of j, how balanced is this match. That is what you try to quantify using the matching index.

So we stop here.