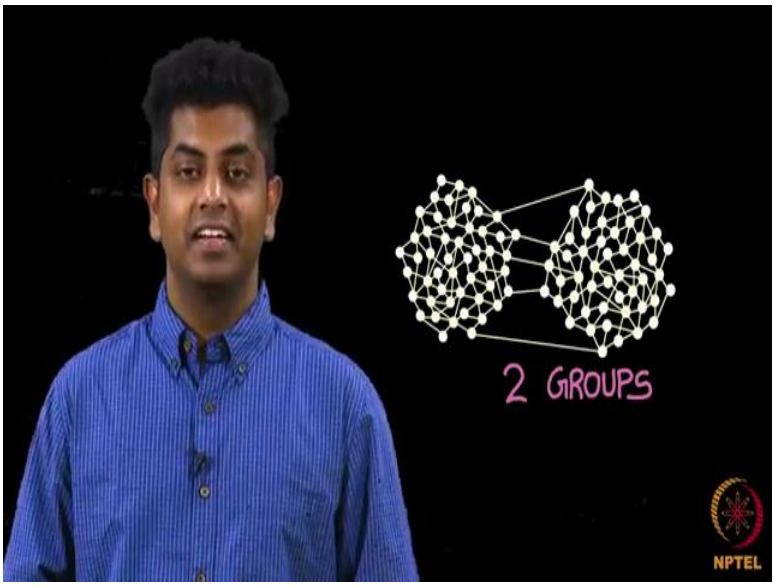
Social Networks
Prof. S. R. S. Iyengar
Department of Computer Science
Indian Institute of Technology, Ropar

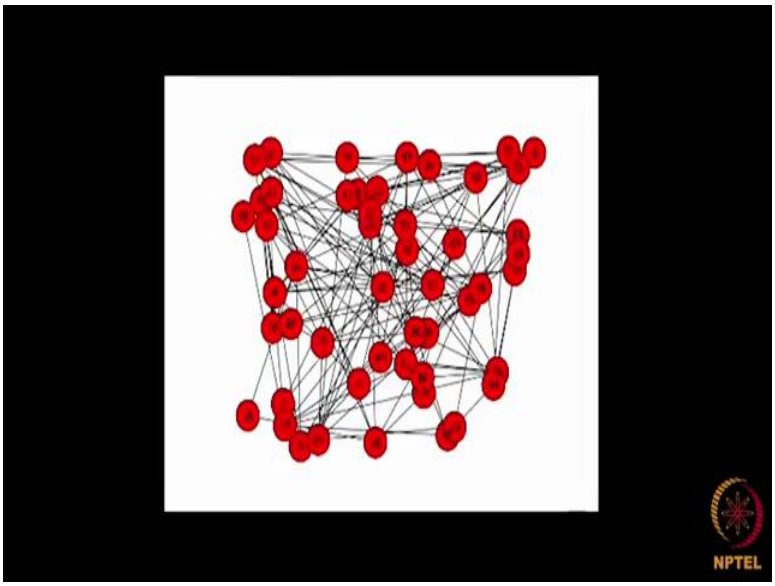
Lecture - 39
Strong and Weak Relationships (Continued) and Homophily
Betweenness Measures and Graph Partitioning

(Refer Slide Time: 00:08)



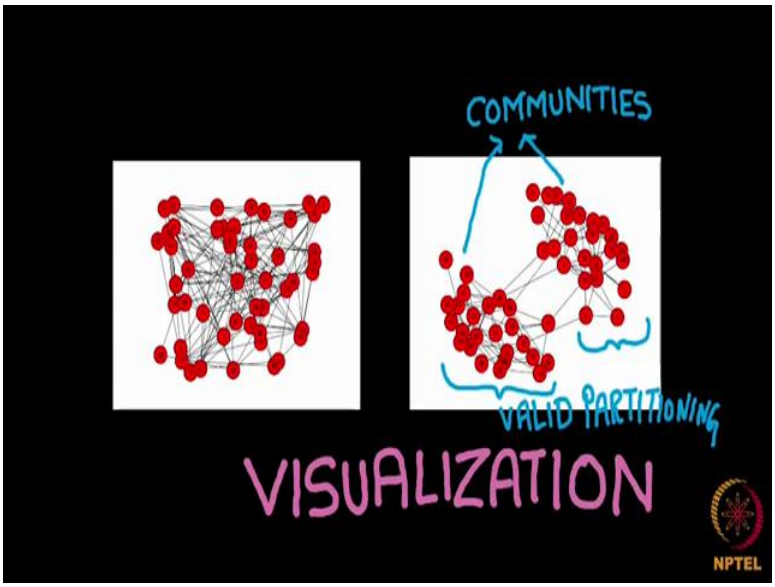
Let us start with the nice puzzle. Here is the classroom of 100 people and there are precisely two teams in this, two groups. Now what do I mean by two groups? Guess what I am saying, two groups means what? We say there is groupism in this place, in this organization means there are two teams, two groups and friendships are within and not across. By not across I mean, I do not mean there is no nothing across, I mean it is less across and more within then we say oh the classroom right now has two segregations, is segregated into two, right, and there is groupism here.

(Refer Slide Time: 00:57)



Now, look at this graph what can you say about the graph? Does this graph have this is the friendship network of some 50 students, do you see any kind of what to say grouping here, are there two groups here, I think no there is a look like there are two groups.

(Refer Slide Time: 01:23)



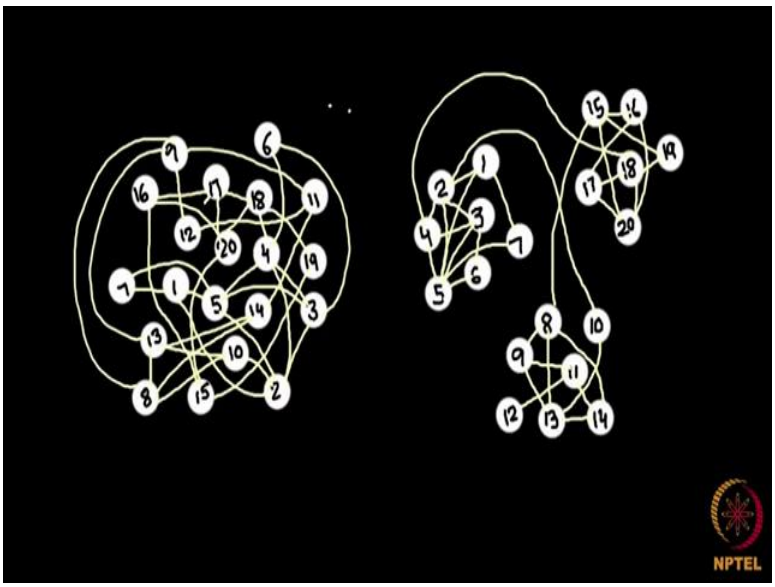
But probably you are wrong there are two groups here observe, there are two groups there. The way it was shown before to you probably did not recognize that there were two groups, but actually there were two groups it all depends on how you visualize the graph, I tried visualizing a different way and I see that there are actually two groups. The

visualization before was very important that said there is only one group, it is all in the visualization. In fact, visualization is a separate branch of research in computer science anyway I will we will not get there, but we will answer this question given a network like this how many groups are there.

These groups are called communities, in the languages or subject it is called communities. So, how do you define a community? A community is a bunch of nodes where the connections are a lot within them and a lot less outside. This is technically called, a bunch of nodes is called communities these nodes partitioning of a vertex at basically, you have 50 nodes that is the vertex said and you classify them as some 20 and 30 or something like that and you call this partitioning as a valid community partitioning if such a partitioning is such that there is a lot of intercommunity edge spar city and intra community edge density intra means what within inter means between. So, between it is less, within it is more then you call such a partitioning, or the vertex set as community partitioning.

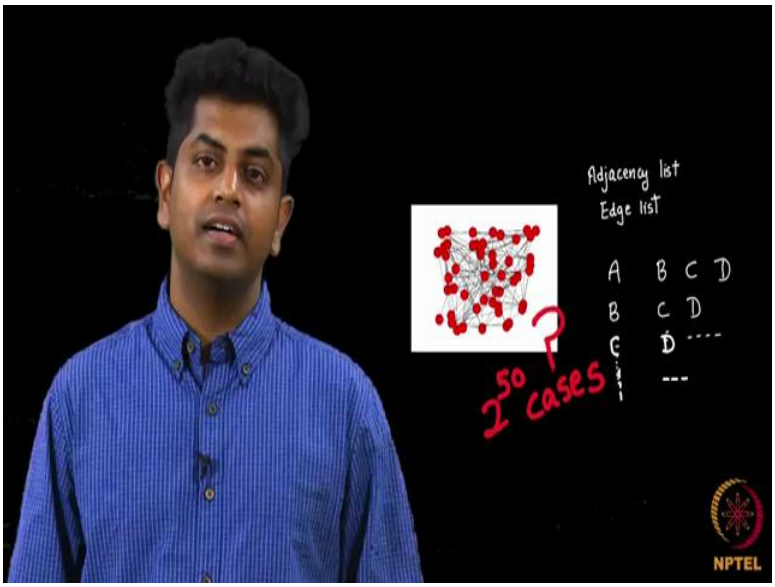
Now let us get back the network that I showed you 50 people you we all initially thought this only one group here, but we realized is two groups here. Given a graph g which is input to a computer how do you find how many groups are there, maybe there are two or may be more why more? I will give you an example.

(Refer Slide Time: 03:36)



Look at this network how many groups do you have? 2, 1, observe if you visualize it properly you see that there are 3 groups, how do we deducts its groups, such groups. I will give you a quick answer to it intuitive answer and a technical answer to it.

(Refer Slide Time: 04:02)



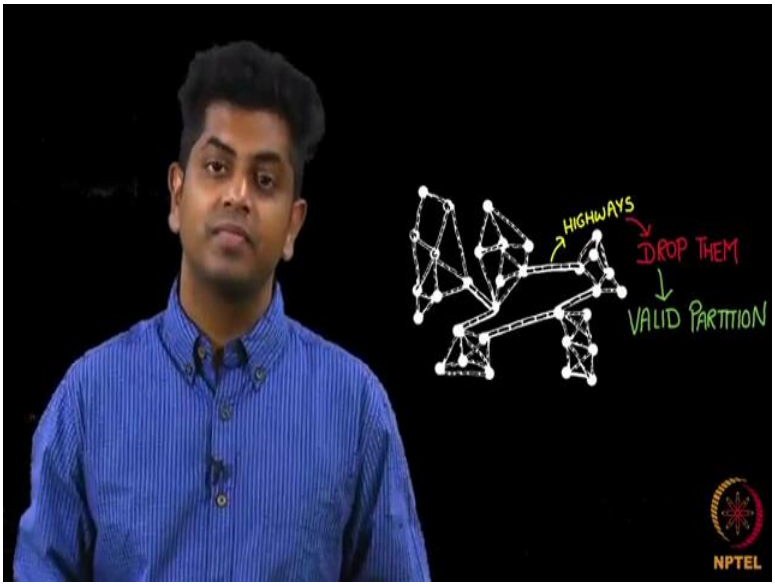
50 people there is some partitioning here and I need to know what is the partition. Is it even easy to answer this the graph is given to me in the form of let us say adjacency list or edge list, what do you mean by this I just know all the links in the graph it is given to me in a file let us say.

We have seen these things in our programming assignments right. This is how I know the graph let us say this way, now A is adjacent to B, C is adjacent to D so on and so forth. There is some complicated graph like this, and I see that there are 50 people, and these are the links and the links are known to me. Only with this information can I get to see get to know if there are teams here, this is hard, very hard to solve, it is a lot of computation whole lot of computation. In fact, precisely speaking you may have to, you observe that there are actually around 2^{50} possible partitioning of this 50 people, how is slightly complicated do not worry about it trust me it is 2^{50} number of partitions.

For each partition you must check is this partition leading to community, is there inter community spar city and intra community density, is technical languages all I am saying is there more adjust within and let us say just across then you say this is the community, I have found the community there are two communities. There are 2^{50} cases that you must

check that is the huge number, you probably cannot do it. Is there a easy way out? Yeah, there is an easy way out and I will give you the intuition the intuition is very clear. Assume I gave you all the road network of India, I will give you all the roads connecting two points you must and I will remove all the names, you do not know which node represents what just given the nodes. Nodes are all regions let us say town, cities, villages whatever and edges are the roads connecting these two places, direct roads connecting these two places.

(Refer Slide Time: 06:02)



So, where is a road that connects two points through region then you do not put a link between these two nodes, you put a link through this node. Let us say city A and city B do not have a direct road, but you should pass through C then you put a link between A to C and then C to B clear. How do you find out what are the main, what are states here let us say of the country? Yeah, states are basically probably communities here do you see, there are very few roads between the states and a lot of roads within this state. There are very few roads across cities and a lot of roads inside a city, grab on this sentence. There are a lot of roads inside the city, but a lot less roads across the city why? You see highways connecting two cities there will not be many roads between any two locations of the city, the only way you can get to this location of the city to some other location of some other city is through this highway correct.

This highways act as bridges in connecting two cities, if you can identify these highways and remove them then you get a partition that represents communities. I repeat in the road network if you identify the highways drop these highways remove them basically from the graph, what is remaining? Remember you are removing edges not nodes in the road network you have nodes and edges nodes represent cities towns and edges represent roads between them what you do is you look at the highways and drop these highways and the resulting network should be, the resulting nodes that you see should be partitioning of the graph. What you mean by resulting nodes? It results in a disconnected bunch of; you will see a lot of edges within when you remove the highways. They are probably the states or cities. This is basically the intuition behind what I am going to say next. Our puzzle that we are we were trying to solve is 50 student classroom how do you detect whether there are groups here or not.

(Refer Slide Time: 09:02)

You just go on the hunt for highways here. What do you I mean by highways here? Of course, there are highways here even in friendship networks, how do you see these are the edges that connect between these two clusters, they act as highways right. How do you detect them when the graph is not really viewer friendly, it does not tell you the communities when you see it unless you pluck it out like this, place it like this and then you see there – yes, there are two teams. These edges if you can detect and remove them then you see a disconnected two components - component 1, component 2 and then you can say ha yes there are two teams.

So, it all boils down to you detecting these links which act as highways. How do you detect them? So, we need to quantify this, how stronger a highway is this road correct, there is this notion of betweenness that does this. Now let us go slowly and let me define what one means by betweenness. Look at this edge in this graph take a node from this side and a node from that side, consider a path connecting these two nodes you see it always passes through this red edge, red link correct, it always passes through this red link. Take some other node these sides, some other node that side you have to invariably pass through this edge if you want to take a connecting path between these two nodes. But when you pick two nodes well within the left side then you can have a path which connects these two links directly like this correct.

Now betweenness of an edge is defined as the total number of paths that go on it, this is the very dumb definition. A very regress definition is slightly mathematically involved. All it says is the total number of paths that pass through this edge the total number of shortest paths that pass through these edges versus the total number of shortest paths between two vertices. What do I mean by this?

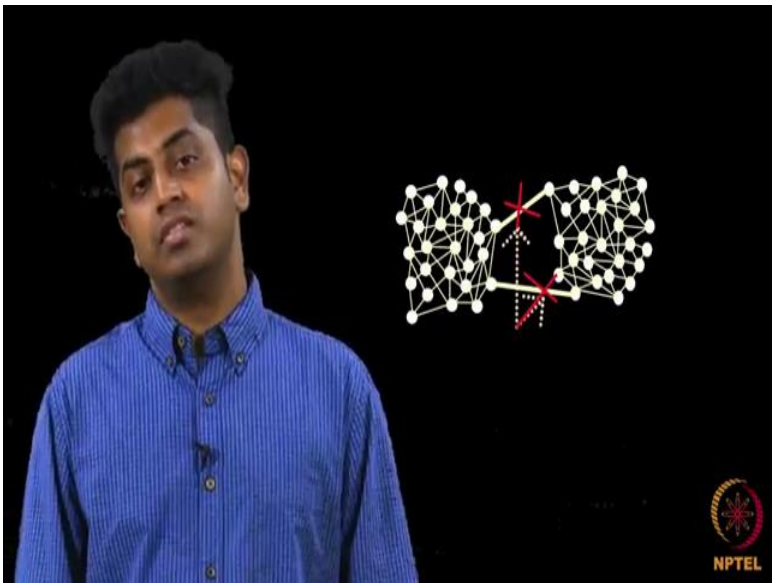
(Refer Slide Time: 11:35)

Betweenness of an edge is defined as $\sigma_{st}(e) / \sigma_{st}$, this fraction means what? You consider basically a path from s to t, you take a path from s to t and you see if edge e is coming on it; what are those shortest path from s to t where edge e comes on it; and what are all the shortest paths from s to t.

This fraction you see I take the sum of all such fractions across all possible s and t as a little mathematically involved all that it says is I am going to take the fraction of edges, a fraction of paths that uses this edge this is complicated version of this very straightforward statement and this is called the betweenness value of v. So, betweenness of value v is computed by using this rational of edge is has high betweenness if it comes in between as a link in the paths connecting two nodes. So, formally speaking betweenness a value of an edge is $\Sigma_{st}\ \sigma_{st}(v) / \sigma_{st}$ running through all the vertices by this simply speaking all that we say is straightforward intuition that what are the fraction of shortest paths that cross e.

This quantifies betweenness, now how can we use this. Do not you think intuitively betweenness means this, higher the betweenness more does it connect two components that otherwise are not connected correct.

(Refer Slide Time: 13:19)



Look at this example this edge there are two edges that connect these two components and the betweenness of these two edges will be high as compared to the betweenness of some edges within the clusters that is easy to see. So, what I do is remember the question that we started with, given a classroom of 50 people we check are there two groups here, how do we do that? You take this graph and compute the betweenness of all the edges here other is seemingly not so complicated, it is very well known that this runs in a time

that is doable on a computer it does not take two to the power of 50 computations, it can be done much faster.

I do this and I find out the betweenness values of all the edges and then look at those edges that I have betweenness I observed that these two edges have high betweenness and what I do is I remove these two vectors and I observed that the graph becomes disconnected. When the graph becomes disconnected it has two components and these two components are the two groups in the classroom. So, we can find out what are the communities by looking at high betweenness edges.