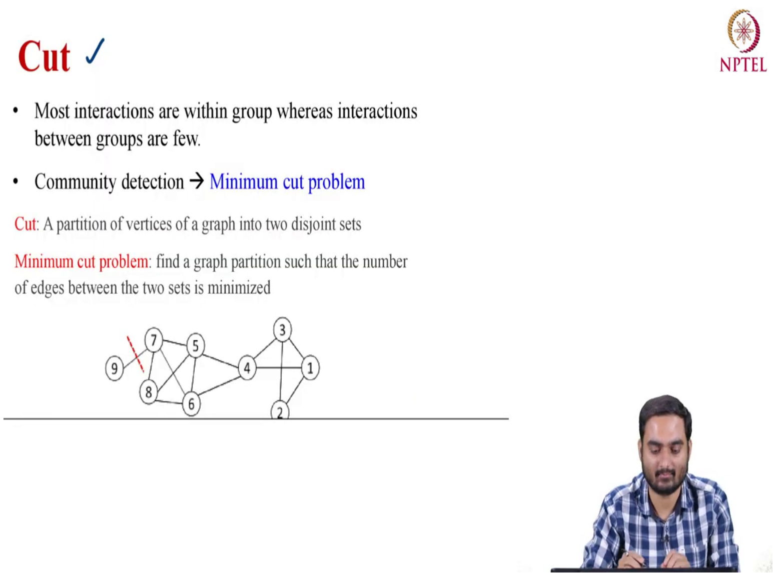
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 05
Lecture - 03

So, in the last lecture we have started discussing about simple ways to detection communities we have discussed you know clique, k-clique, k-clank, k-club right, k- core and so on. So, today we will discuss two important metrics very important metrics that we use for community detection for graph clustering ok.

(Refer Slide Time: 00:43)



Cut ✓

- Most interactions are within group whereas interactions between groups are few.
- Community detection → Minimum cut problem

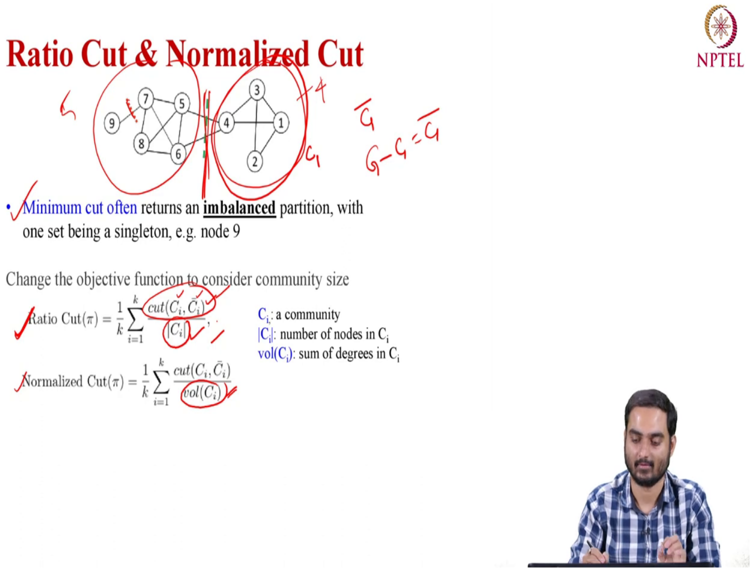
Cut: A partition of vertices of a graph into two disjoint sets

Minimum cut problem: find a graph partition such that the number of edges between the two sets is minimized

The first one is cut, I am sure that you have heard about this terminology before is called cut and mean cut problem particularly right. So, what is a cut? So, you know, so basically a cut is a partition of nodes of a graph into disjoint sets. So, the idea is that, say you have a graph like this and you want to identify edges which upon deletion; after deletion creates you know it after deletion create two different partitions.

For example, if I cut here you get this as one partition and the remaining as another partition ok. If I say cut if I cut here right I get one partition here, another partition here right. So, what is the mean cut problem? The mean cut problem is to find out such a cut which is minimum right. So, basically this is the definition that, find a graph partition such that the number of edges between the two sets is minimized ok. So, let us think of this problem carefully.

(Refer Slide Time: 02:05)



Ratio Cut & Normalized Cut

- Minimum cut often returns an **imbalanced** partition, with one set being a singleton, e.g. node 9

Change the objective function to consider community size

$$\text{Ratio Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{|C_i|},$$

$$\text{Normalized Cut}(\pi) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(C_i,\bar{C}_i)}{vol(C_i)}$$

C_i: a community
$|C_i|$: number of nodes in C_i
vol(C_i): sum of degrees in C_i

So, if you think of this network right and if you use mean cut problem so; obviously, the output would be which one should cut here because this is only 1 edge that will be removed if we cut here right, but if we cut here 2 edges will be removed right, but think about it carefully.

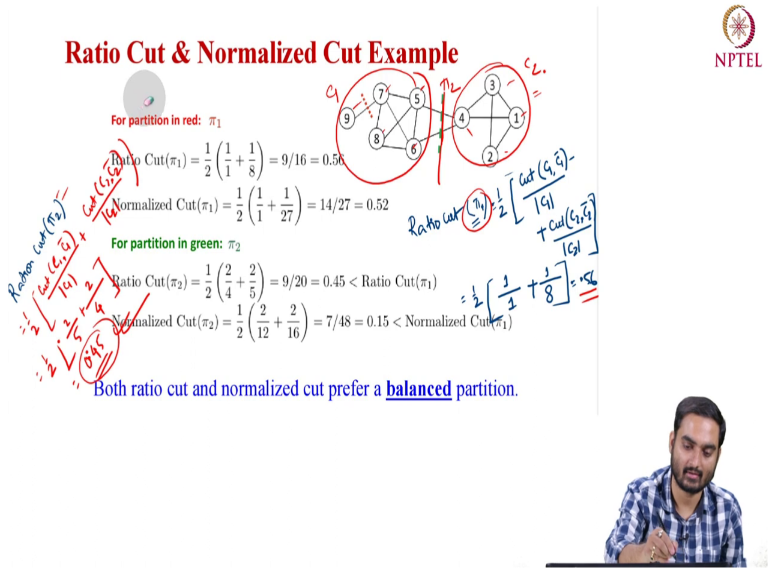
If I cut here right we will have one community with a single node and another community with a lot of nodes it is an imbalanced community imbalanced partition and we generally do not want to detect imbalanced communities ok. We generally want to detect communities which are balanced right.

So, the obvious choice would be to you know not to cut here not to cut here, but to cut here. So, if I cut here we will see that it will create to it will it will basically create two partitions in one partition there would be 4 nodes, in another partition there would be 5 nodes and the partitions will be balanced ok. So, mean cut problem oftentimes returns imbalance partition therefore, we will what we will do, we will modify the you know mean cut metric in a different ways. So, we define something called ratio cut, what is the ratio cut? Ratio cut is defined in this way.

So, it is a; it is a fraction where the numerator is the cut meaning that we have two partitions and cut off partition C i C i bar. So, what is C i C i bar? So, say this is C 1 right say this is C 1 and the rest of the network is C 1 bar ok. So, G minus C 1 is C 1 bar ok. So, what is ratio cut? So, the numerator is the cut itself the edges between two communities one is C i and another is C i bar and the denominator that is the numerator and the denominator is the size of the; size of the community ok. I will take an example and I will discuss ok.

So, this is ratio cut, what is; there is another modified definition of a cut called normalized cut. What is normalized cut? This is same as ratio cut the only difference is that in the denominator now we have instead of the size of the community now we have the volume of a community. What do you mean by volume of a community? Volume of a community is the sum of degrees of all the nodes present in the community, the sum of degree of all the nodes present in the community that is the volume of a community. So, I am normalizing it by the volume of a community ok.

(Refer Slide Time: 05:11)



Ratio Cut & Normalized Cut Example

For partition in red: π_1

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

Both ratio cut and normalized cut prefer a **balanced** partition.

NPTEL
C_1
C_2
π_2
1
2
3
4
5
6
7
8
9
$\text{Ratio cut}(\pi_1) = \frac{1}{2}\left[\frac{\text{cut}(C_1,\bar{C_1})}{|C_1|} + \frac{\text{cut}(C_2,\bar{C_2})}{|C_2|}\right]$
$= \frac{1}{2}\left[\frac{1}{1} + \frac{1}{8}\right] = .56$
$\text{Ratio cut}(\pi_2) = \frac{1}{2}\left[\frac{\text{cut}(C_1,\bar{C_1})}{|C_1|} + \frac{\text{cut}(C_2,\bar{C_2})}{|C_2|}\right]$
$= \frac{1}{2}\left[\frac{2}{5} + \frac{2}{4}\right]$
$= 0.45$

So, let us take an example ok, this example and let us you know let us measure the value of ratio cut and normalized cut ok. Let us assume that let us take this as a cut ok, if I take this as a cut what would be the value of the ratio cut and what would be the value of the normalized cut? Ok. If I cut here we will have two communities, this is one and this is another ok. So, what is the definition of ratio cut? So, the ratio cut and let us say this partition is called pi 1 and later on I will also; I will also calculate the same for this partition this would be pi 2.

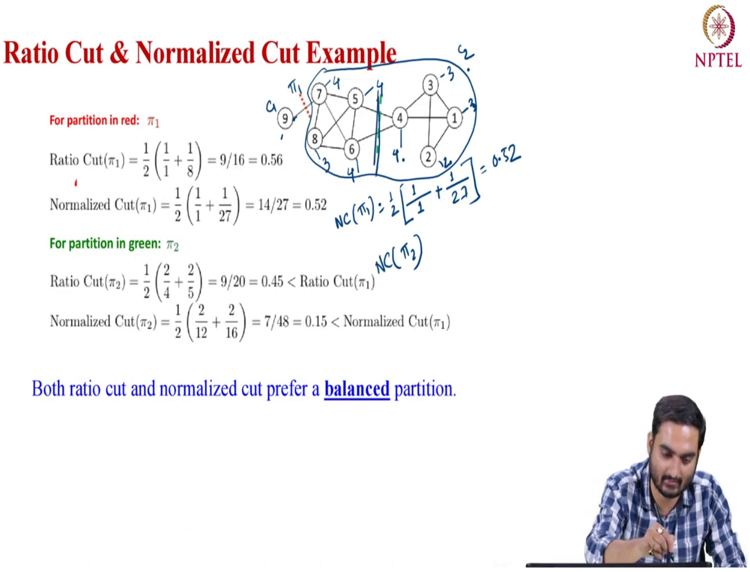
So, rate ratio cut of pi 1 is what? Again go back to the definition 1 by k, what is k, k is a number of partitions, k is the number of partitions that are going to be created ok. So, this would be if I cut here there will be two partitions, so 1 by 2 ok. Then what we have? Cut of; so, let say this is C 1 this is this is C 2 right. So, C 1 comma C 1 bar by divided by size of C 1 plus cut C 2 comma C 2 bar divided by mod of C 2, again cross check it this is the definition right, for every partition for every community you have this one ok.

So, this would be half, what is the cut of C 1? So, this is C 1 and this is C 1 bar ok the remaining graph is C 1 bar. What is the cut size? Cut size is 1. So, this would be 1 divided by what is the size of C 1, 1 plus cut C 2. So, this is C 2 and this is C 2 bar ok. So, cut would be 1 divided by size of the C 2 is 8, there are 8 nodes right if you calculate this would be 0.56 ok. So, this is ratio cut for pi 1. What is the ratio cut of; ratio cut of pi 2?

So, if I calculate in the same manner I will have this one I will have. So, this is pi 2 C 1 C 2. So, this would be 1 by 2 because again another two partitions right, cut of C 1 comma C 1 bar by mod of C 1 plus cut of C 2 comma C 2 bar by mod of C 2. So, half cut C 1 C 1 by 2 because there are there are you know 2 edges. So, this would be 2 t by divided by what is the size of C 1 1, 2, 3, 4, 5 plus again cut is 2 and what is the size of C 2 1, 2, 3, 4. So, if we calculate this would be 0.45 ok.

So, ratio cut of pi 1 is 0.56, ratio cut of pi 2 is 0.45. So, which one to choose which partition to choose, remember our task is to minimize the ratio cut minimize the cut minimize the ratio cut here we are trying to minimize the objective function ok. So, we will definitely choose this one because this partition will give you minimum ratio cut. So, we will cut here and that that was also our choice. So, ratio cut will give you the ideal solution ok.

(Refer Slide Time: 09:28)



Ratio Cut & Normalized Cut Example

For partition in red: π_1

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

Both ratio cut and normalized cut prefer a **balanced** partition.

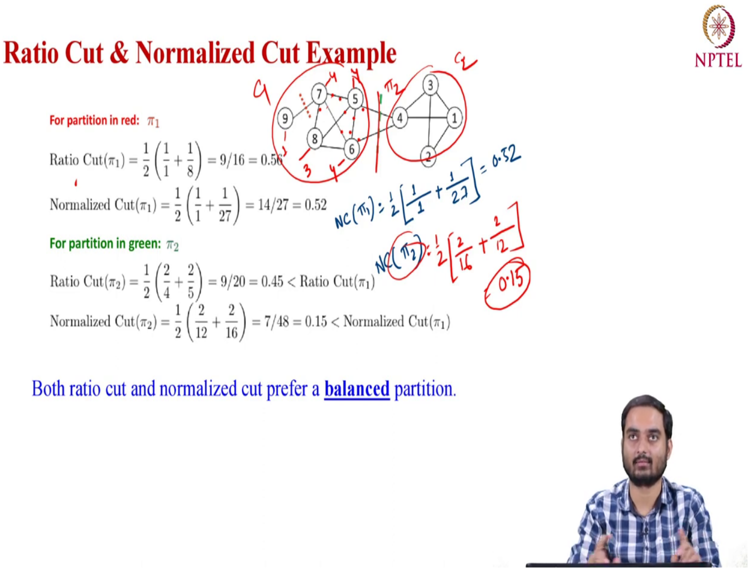
$$NC(\pi_1) = \frac{1}{2}\left[\frac{1}{1} + \frac{1}{27}\right] = 0.52$$

$NC(\pi_2)$

Now, let us look at the second way which is the normalized cut and let see whether the normalized cut will also give you the same result ok. So, pi 1. So, normalized cut of pi 1 is let us go back and see the definition 1 by k sum over all k cut of cut and then volume, volume is the sum of degree of all the nodes right. So, half right we have say let say this is C 1 and this is again C 2 right.

So, cut C 1 C 1 bar would be 1 divided by the volume of the C 1 is there is only 1 node and the degree is 1. So, this would be 1 plus cut of C 2 C 2 bar is 1 divided by the volume of C 2 would be the sum of all the degree. So, the degree is 3 3 2 4 4 4 and this one is also 3 and this is degree 1, 2, 3, 4 ok. So, 6, 8, 12, 16, 19, 27 this will be 0.52 ok and normalized cut for pi 2. So, this is this is pi 2 ok.

(Refer Slide Time: 11:00)



Ratio Cut & Normalized Cut Example

For partition in red: π_1

$$\text{Ratio Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{8}\right) = 9/16 = 0.56$$

$$\text{Normalized Cut}(\pi_1) = \frac{1}{2}\left(\frac{1}{1}+\frac{1}{27}\right) = 14/27 = 0.52$$

For partition in green: π_2

$$\text{Ratio Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{4}+\frac{2}{5}\right) = 9/20 = 0.45 < \text{Ratio Cut}(\pi_1)$$

$$\text{Normalized Cut}(\pi_2) = \frac{1}{2}\left(\frac{2}{12}+\frac{2}{16}\right) = 7/48 = 0.15 < \text{Normalized Cut}(\pi_1)$$

$$NC(\pi_1) = \frac{1}{2}\left[\frac{1}{1}+\frac{1}{27}\right] = 0.52$$

$$NC(\pi_2) = \frac{1}{2}\left[\frac{2}{16}+\frac{2}{12}\right]$$

$$\boxed{0.15}$$

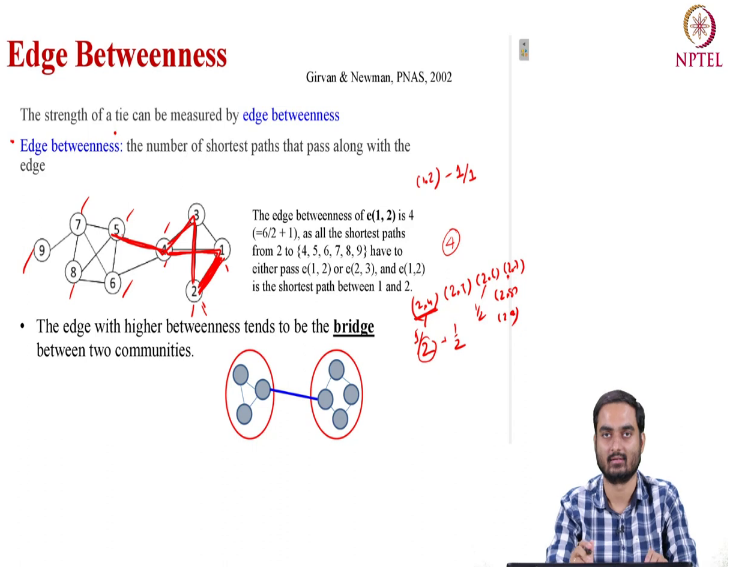
Both ratio cut and normalized cut prefer a **balanced** partition.

Let us again erase it and we will have another partition which is pi 2 normalize cut of pi 2 would be half right, what is the cut? So, say this is C 1 and this is C 2 ok, cut size is 2. So, 2 by the volume of C 1, the volume of C 1 would be if you calculate you see that I think this would be let us see this is 1, 2, 3, 4 right.

This is 1, 2, 3, 4 this is 3 the degree is 1, 2, 3, 4 right, 16 right plus 2 by if you sum it up you will see it is 12. So, the this is the value would be 1 point 0.15 right. So, which one is minimum, this one is minimum. So, again pi 2 will be chosen. So, ratio cut and normalized cut will produce the same output and the output will be a balanced partition ok.

Min cut problem is also NP complete right. So, we generally try to relax it. So, what is the; what is the you know what is the definition of a community according to the conductance or cut right. So, the definition of a community is that community is a group of nodes where we only look at the edges across communities right. So, it should be sparsely connected externally right. So, cut or ratio cut right. So, particularly cut. So, cut does not look at the structure within a community it only looks at the structure you know across communities and try to it basically tries to minimize it ok.

(Refer Slide Time: 12:45)



Edge Betweenness

Girvan & Newman, PNAS, 2002

The strength of a tie can be measured by edge betweenness

- Edge betweenness: the number of shortest paths that pass along with the edge

The edge betweenness of **e(1, 2)** is 4 (=6/2 + 1), as all the shortest paths from 2 to {4, 5, 6, 7, 8, 9} have to either pass e(1, 2) or e(2, 3), and e(1,2) is the shortest path between 1 and 2.

- The edge with higher betweenness tends to be the **bridge** between two communities.

The second metric that we use is called edge betweenness and the edge betweenness metric is something that we you know we already discussed in one of the; in one of the lectures earlier right. We discussed something called betweenness centrality ok and we also mentioned that you can you can actually recast it for edges where we look at all pairs of shortest paths and in in how many such pairs of shortest paths this particular edge exists right, this is exactly same.

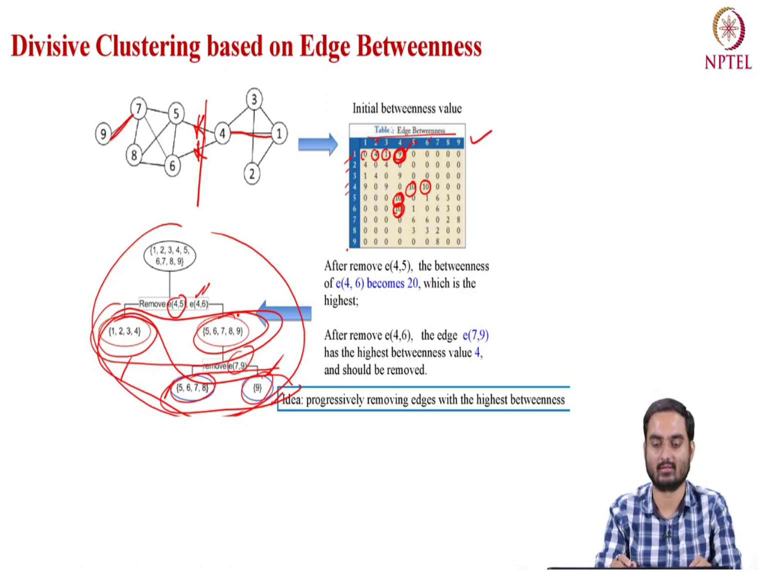
So, let us think of let us measure the between edge betweenness of this edge 1, 2 right. So, we take all pairs, but in most of the pairs this edge will not come therefore, so, those pairs will not be considered right, but we need to consider this pair 1, 2 because in 1, 2 there is only shortest path there is only one shortest path and in this shortest path the edge 1, 2 exists. So, this would be 1 by 1 ok and you also need to consider all the other nodes 4 5 6 7 8 9 with respect to 2.

So, we need to consider you know 2, 4 why, 2 5, 2 6, 2 7, 2 8, 2 9 why? Because when you measure the shortest path between say 2 to 4 ok, this is one shortest path size 2 this is another shortest path size 2 right. So, between 2 to 4 there are two shortest paths and among these two shortest path there is one shortest path where this edge 1, 2 exists.

So, for 1, 2 this would be 1, for 2 4 pair this would be 1 by 2 there are two shortest paths and in one shortest path the edge 1, 2 exist. Similarly, for 2, 5 2, 5 you can move through this or you can move through this right again 1 by 2 this would be again 1 by 2, 1 by 2, 1 by 2. So, there are 6 such pairs. So, this will be 6 by 2. So, 6 by 2 plus 1 is 4. So, 4 is the edge betweenness of this edge 1, 2.

So, how do we detect community here? So, first we measure edge betweenness of all edges ok. So, if the edge has higher edge betweenness it basically means that it comes across communities, it basically falls between communities therefore, whenever you basically wants to want to traverse through you know nodes and edges you need to encounter this edge.

(Refer Slide Time: 15:33)



Divisive Clustering based on Edge Betweenness

Initial betweenness value

Table : Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0

After remove e(4,5), the betweenness of e(4, 6) becomes 20, which is the highest;

After remove e(4,6), the edge e(7,9) has the highest betweenness value 4, and should be removed.

Idea: progressively removing edges with the highest betweenness

So, the first task is we measure the edge betweenness of all the edges and then we remove those edges whose edge betweenness is maximum ok. So, this is the edge betweenness metric, you see that these are nodes 1 to 1, 1 to 9 and again 1 to 9 and each such entry indication appear right 1, 2. So, this is 1, 3 right this is 1, 4. Now, if there is an edge between this pair, say for example, this 1, 4 right 1, 4 is this edge and this entry indicates the edge betweenness of this edge.

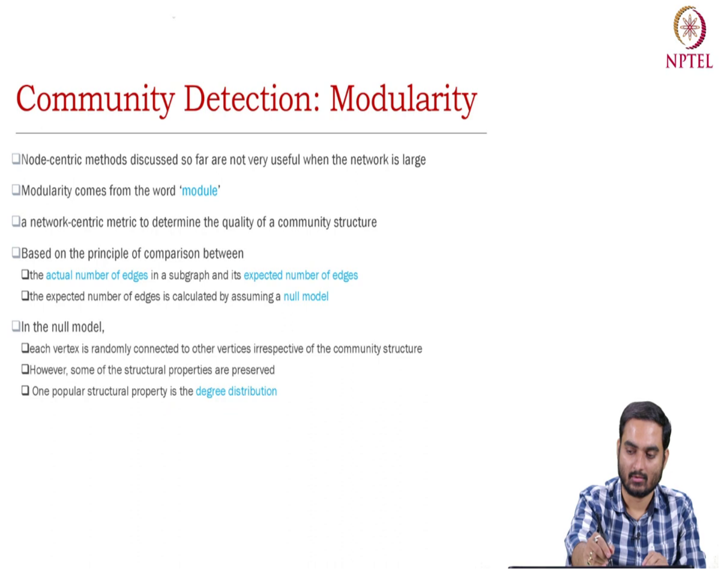
So, 9 is the edge betweenness of the edge one 4 right. So, I calculate the edge betweenness of all the edges and then I identify the edges which have maximum edge betweenness. So, the edges you see here this and this meaning 4, 5; 4, 5 and 4, 6. So, this is also 5 4 and 5 4 and 4, 5 and 4, 6, this is this is also 4, 5 and 4, 6. So, 4, 5 and 4, 6 are the edges which you remove.

So, if I remove these 2 edges what would happen, we will have the resultant graph, on the resulting graph we again measure the edge betweenness of all the edges identify the edges which have maximum edge betweenness remove them and you keep on doing this thing again and again right. So, think about it when we remove 4, 5 and 4, 6 you will have 1, 2, 3, 4 and this one 5, 6, 7, 8, 9 right.

Then again then which one you remove, if you calculate you will see then this edge will be removed after that, then you will have 5, 6, 7, 8 and 9. So, it depends on how many I mean either you keep on removing edges until unless you see all the nodes you know are separated or you stop at certain point if you see that ok I do not need more than three communities for example, more than 4 communities. So, if you stop here you will get 1, 2, 3 communities ok.

Now, if the number of communities is available beforehand, so you can stop here right you can stop accordingly. And remember, this also kind of gives you the notion of hierarchical community right, because you see here this is the first hierarchy right, this is the second level hierarchy; this is the second level hierarchy. So, this is basically hierarchical structure. So, therefore, I mentioned earlier that you do not need to explicitly detect hierarchical communities it will automatically get identified when we detect community structure ok.

(Refer Slide Time: 18:25)



Community Detection: Modularity

- ❑Node-centric methods discussed so far are not very useful when the network is large
- ❑Modularity comes from the word 'module'
- ❑a network-centric metric to determine the quality of a community structure
- ❑Based on the principle of comparison between
  - ❑the actual number of edges in a subgraph and its expected number of edges
  - ❑the expected number of edges is calculated by assuming a null model
- ❑In the null model,
  - ❑each vertex is randomly connected to other vertices irrespective of the community structure
  - ❑However, some of the structural properties are preserved
  - ❑ One popular structural property is the degree distribution

So, I stop here in the next part we will discuss a very important metric called Modularity. So, modularity is a metric which is which has been extensively studied in the network science community and we will see that how we can use modularity for community detection ok.

Thank you.