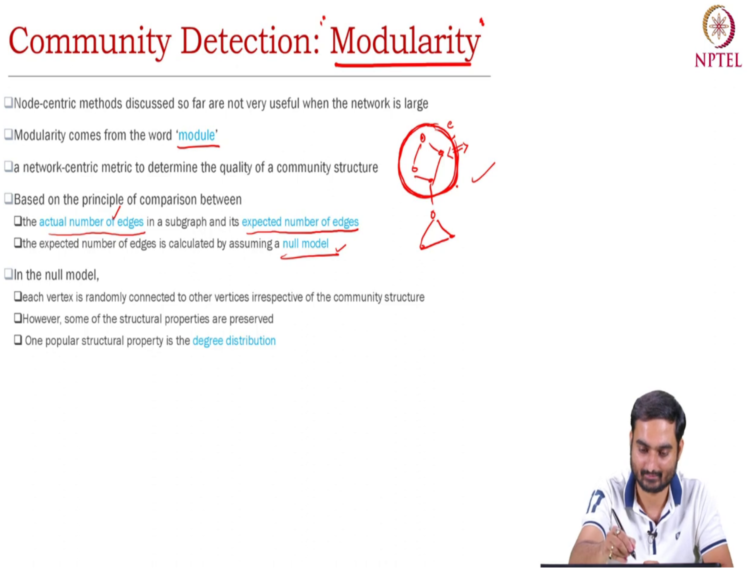
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 05
Lecture - 04

In the community detection, you know chapter we have already discussed simple measures, simple ways of detecting clusters. For example, we discussed you know all these nodes centric measures like K-clique, K-clan, K-club. We also discussed you know a metric called card, right and different versions of card, normalized card, ratio card. And we have seen how we use this matrix for community detection, right.

(Refer Slide Time: 00:57)



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, today we will discuss another such metric called modularity which is I think by far one of the best metrics that people use for community detection and this is widely studied across different networks, different areas of research and so on, ok. So, this this metric is kind of you know equals to m c square in the domain of network science, ok.

So, history, so modularity metric was proposed by Mark Newman. He is a English-American scientist, a faculty a professor at University of Michigan. And he proposed this method in 2006, which was published in Pinus, which is one of the top journals. And before that, before the you know proposition of modernity, people generally used to use cliques, different versions of cliques, people also used ah you know conductance cart ratio, these kind of measures. But this is these turned out to be statistically you know significant metric, you can detect modularity, you can use modularity for detecting quality community structure across different types of networks.

So, as the name suggests modularity is it was basically you know motivated by the term module. So, module is basically community or cluster, ok. And what does it what does it indicate? The term modularity, the quality, I mean the quantity you know the kind of quantity that is modularity expresses, this basically indicates the quality of a community structure.

I mean let us assume that you have an algorithm to detect community structure, you have already detected it using the algorithm and you want to measure the quality of this community structure, whether the community structure is at all you know good or not. So, how do we evaluate it? So, modularity was initially proposed to evaluate a community structure. And later people you know use this metric for detection as well.

So, we will first discuss this metric. We will try to understand that given a community structure how the modularity metric can give you the quality of a community structure, and then we will see how we can use this metric for community detection from the scratch, ok.

So, what is the philosophy behind this metric? So, basically the philosophy is that random network, random graph, does not have any community structure. We discussed what is random graph in chapter 3, where we discussed you know different types of network measures. We also discuss ER model, random graph model.

And we also discussed that you know how this random graph is basically formed, you choose a pair of nodes, you toss a coin, if the coin is head for example, right you connect these two nodes, otherwise you do not connect, ok. So, there is no preferential attachment as such. There is no hemophilia as such due to which these two nodes are connected, ok in a random network. So, the idea is that in a random network, there should not be any community structure, ok.

And given a network, once you identify communities in some ways, right, let us say let us say you are you have identified, let us say you have identified, right, you have identified this as a community, ok. What we will see? We will see what is the random configuration or random you know random counterpart of this particular community.

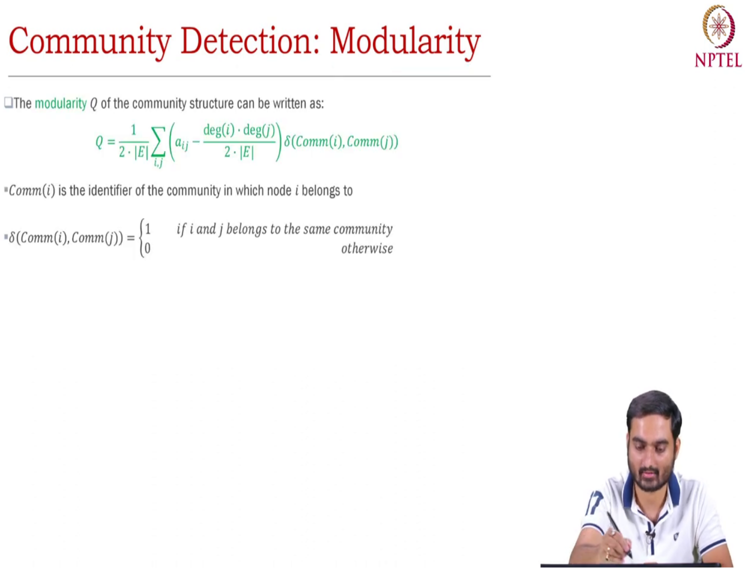
Meaning, that if you had drawn this network using random graph model, what would have been the quality of the of this group, ok, and how the existing network the existing community differs from the community that you could have detected from a random network. And the higher this difference, the better the quality of the community structure, ok. I will quantify each and everything one by one.

So, basically the idea is that you basically measure the actual number of edges within a group, within a community which has already been detected, and what would be the expected number of edges within the group if the network you know had been constructed using a random graph model, ok.

So, we define something called a null model which is basically a random graph model, right. So, we define a random, we define a null model and using the null model we draw a random counterpart of the original graph. And then, we quantify the expected number of edges within this community, within this group. And then, we see how this how these two quantities are different, the actual number of edges and the expected number of edges, ok, alright.

So, let us now try to first understand how do we create this random graph, right.

(Refer Slide Time: 06:25)



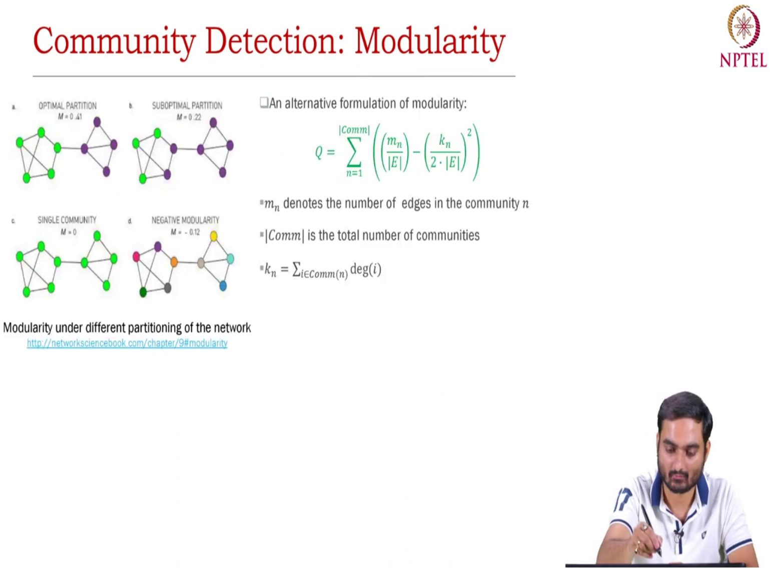
Community Detection: Modularity

❑The modularity Q of the community structure can be written as:

$$Q = \frac{1}{2 \cdot |E|} \sum_{i,j} \left(a_{ij} - \frac{\deg(i) \cdot \deg(j)}{2 \cdot |E|} \right) \delta(Comm(i), Comm(j))$$

- $Comm(i)$ is the identifier of the community in which node i belongs to
- $\delta(Comm(i), Comm(j)) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belongs to the same community} \\ 0 & \text{otherwise} \end{cases}$

(Refer Slide Time: 06:26)



Community Detection: Modularity

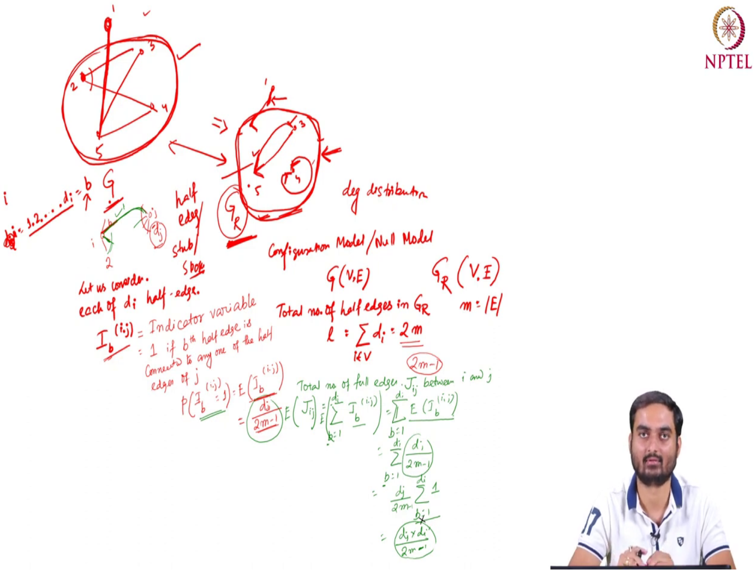
Modularity under different partitioning of the network
http://networksciencebook.com/chapter/9#modularity

❑ An alternative formulation of modularity:

$$Q = \sum_{n=1}^{|Comm|} \left(\left(\frac{m_n}{|E|} \right) - \left(\frac{k_n}{2 \cdot |E|} \right)^2 \right)$$

- m_n denotes the number of edges in the community n
- $|Comm|$ is the total number of communities
- $k_n = \sum_{i \in Comm(n)} \deg(i)$

(Refer Slide Time: 06:34)



$d_i = 1, 2, \ldots, d_i = b$

deg distribution

Configuration Model / Null Model

$G(V,E)$ $G_R(V,E)$

$m = |E|$

Let us consider.
each of d_i half-edge.

$I_b^{(i,j)}$ = Indicator variable
= 1 if b^{th} half edge is connected to any one of the half edges of j

Total no. of half edges in G_R

$l = \sum_{i \in V} d_i = 2m$

$P(I_b^{(i,j)} = 1) = E(I_b^{(i,j)}) = \frac{d_j}{2m-1}$

Total no of full edges J_{ij} between i and j

$$E(J_{ij}) = E\left(\sum_{b=1}^{d_i} I_b^{(i,j)}\right) = \sum_{b=1}^{d_i} E(I_b^{(i,j)}) = \sum_{b=1}^{d_i} \frac{d_j}{2m-1} = \frac{d_j}{2m-1} \sum_{b=1}^{d_i} 1 = \frac{d_j \times d_i}{2m-1}$$

So, remember when we create the random graph, we want that a property of a graph should be preserved. Let us assume that you know let us assume that this is a graph, ok and what you are trying to do you are say this is 1, 2, 3, 4, 5, you are trying to come up with. So, this is the original graph G and you are trying to come up with a random counterpart or random graph, which actually preserves some property of this graph of the original graph.

And what is the property that we are planning to preserve? We are planning to preserve the degree distribution, the degree distribution of G, ok. So, basically the idea is that in G R, G R should have the same degree distribution that graph G has, ok. So, then how do you; so, then how do you draw the I mean what is the easiest edges way to make sure that in the random graph the degree distribution will be preserved? If I make sure that the degree of each node will remain same, right.

Say this is 1, 2, 3, 4, 5, if I make sure that you know say in in say node 2 in or in graph G has degree 2; if I make sure that in a random graph node 2 will also have degree 2, node 3 will also have degree 2, net node 5 will have degree 3, node 4 will have degree 2 and node 1 will have degree 1, then the degree distribution will remain same, ok.

But you also need to remember that G R is a random graph. It is basically generated using random graph model. So, how do you do that? So, we define something called configuration model, right or null model, ok. And what is the I mean how do you connect these nodes? How do you connect nodes in G R? What you do, we actually break each edge in the original graph into two parts, ok. So, let us say this is your edge 1, 5 is an edge, right I will break it into two parts, this one and this one. I break it at the middle, ok.

And these parts are basically called, these are called half edge, half edge because you are basically you know cutting this into two parts or it is also called stub or spoke, right and so on and so forth; so, half edges. So, we break every edge in the original graph into two halves in the random graph G R, right.

So, remember these are half edges, and half edge are not connected, ok. So, by doing this we ensure that every node will have same degree, I mean the node say the node will have same degree in G R that it has in the original graph G. Then, how do you connect this half edges?

Then, we say we choose one half edge. So, let us say we choose this one, right and then we randomly choose one of these remaining half ages, right, say let us say we choose this this half edge, ok and then we connect, right. Then, let us say we choose this one, and we choose randomly one of the remaining half edges, let us say we choose this one and then you connect, and so on and so forth.

Remember, it may happen, it may happen that you choose this one and for this node you choose for this half edge you choose this half edge. And you will see a self-loop. That is also possible, ok. Self-loop is possible, in fact parallel edge is also possible, graph can be disconnected, everything can be possible, ok. But at least we make sure that in the random graph will have the same degree distribution as the one that you had in in graph G, ok.

Now, this is our G R, right. Now, what we do? We then you know try to understand, say let us say you know in the original graph G, we know that these 4 nodes form a community using some algorithm, right. So, let us also group these 4 nodes here in the G R, let us say these are, so this is the group.

Then, I will see that what is the number of edges what is the number of edges within the group in G and what is the number of edges within this group in G R, ok. And the more the difference between these two quantities number of edges, the better the quality of this community structure.

Remember, this G R, since this is a random graph you will have different such random graphs in different trials. In one trial you have this kind of structure; in another trial you may have a different kind of structure, right. So, therefore, we basically measure the expected number of edges because it is basically a stochastic process, right. We measure the expected number of edges that you know we could find if we had drawn this graph using a random graph model, ok.

So, then let us define the expected number of edges, ok. So, let us say, let us say we have this graph G, ok and you have V comma E, and we have a random graph G R which has same number of, same number of you know vertices V and same number of edges as well, because degrees are same, so number of edges will also be same, right.

So, what is the you know, what is the total number of total number of you know half edges in G R? So, total number of half edges if we denote it by l, this is essentially sum of degrees, sum of degrees of all the nodes here, right. So, you know d i, all i in V, right.

And so, we also know we have seen the theorem that sum of degree of degrees of all the nodes is twice the number of edges. So, this would be 2 into m. What is m? m is the number of edges m is mod E, ok. So, then, let us you know consider let us consider each of, right each of d i half edges. What is d i? d i is the degree of node i, ok. d i is a degree of node i. So, d i, so node i has d i number of half edges, right. Let us denote this half edges as say 1, 2, 3 dot dot d i. So, node node i has this many half edges, ok, and these are the basically indices of half edges.

So, and let us you know use the index b to denote this half edges, ok. So, now, let us quantify let us quantify you know a quantity which is called I b i comma j. What is this? This is basically, so this is basically, this is and this is called indicator, this is called indicator variable, right. These are this is called indicator variable.

This is the value of this would be 1, if, so b is one such half edge, right one such half edge of i, ok; if the bth half edge is connected to any one of the half edges of G. So, let us say this is i, you have d i number of half edges and one such half edge is b, ok and you have j, j also have some half, j also has some half edges. So, this this quantity would be 1, if b is connected to one of this one of the half edges of j, ok.

So, what is the probability? What is the probability that this indicator variable equals to 1? This is basically you know expectation of this one expectation of this one. So, but so, now, let us try to understand.

So, this is how do we how do we quantify this expectation? So, this is basically saying that this b; what is the probability that b will be connected to any of this j any of the half edges of j? So, remember the degree of j is d j, right. Since, b can be connected to one of this d edges, d half edges or it can be connected to the other half edges present in the graph. So, how many after half edges are there? There are 2 m number of half edges, right.

And if I just ignore this one because I am calculating for this half edge. So, what is the remaining half edges present? 2 m minus 1. So, out of 2 m minus 1, if any of if I out of 2 m minus 1, if the this half edge b is connected to one of this d j half edges, then we will get the probability, will do, this will be 1.

So, the probability would be d j by 2 m minus 1. This is the total possibilities and d j is the possibilities, d j is the possibilities for which this i indicator value variable would be 1. This is the probability, right.

Now, what is the total number of and remember if this is one, the meaning that this will be connected to this one, ok. Now, there are other half edges also, right. So, what is the what is the total number of, what is the total number of. So, remember these two half edges will form a full edge, right. So, what is the total number of full edges between i and j?

Now, this is one case, we have other such cases, right. So, total number of full edges; remember all these things are calculated in the expected sense, ok. Total number of full edges we will denote it by J ij, capital J ij, right between i and j would be J ij is; what is this? This is basically summation of I b i, j, right for all b equals to 1 to, 1 to d i.

So, for each such d, for each such half edge of node i I will check whether this is 1 or not. So, number of 1s, right number of 1s will basically tell you the number of full edges between i and j. Say between this and this, this is 1. Say between this and this this is not 1, right. So, number of such 1s will be the total number of edges, full edges between i and j.

How do we quantify, how do we measure this? So, this is total number, ok; so, in the expected sense, what the expected number of full edges between i and j? This would be expectation of this one, ok. So, this would be expectation of this one, ok.

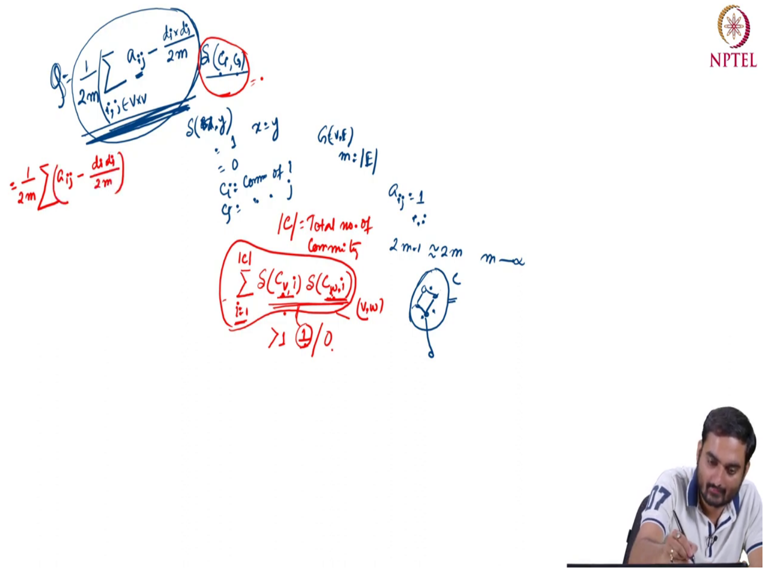
And so this would be ex; so we can take this summation outside. Let us take the summation outside b and, b from 1 to d i expectation of I b i comma j, ok. So, what is this value? So, b 1 to d i, what is this value? This we have already calculated, this is this one, this is d j by 2 m minus 1, ok.

Now, note here is this quantity is independent of b. So, we can take this thing outside. So, d j by 2 m summation of b equals to 1 to d i and then 1, right. So, this summation would be d i. So, we have d j times d i by 2 m minus 1, ok. So, d i times d j by 2 m minus 1 is the expected number of full edges between i and j, ok.

So, I hope you understood how to compute the expected number of edges between two nodes when the graph is formed randomly, ok, using a configuration model.

Now, what is the formula of modularity?

(Refer Slide Time: 21:41)


$Q = \frac{1}{2m} \sum_{i,j \in V \times V} a_{ij} - \frac{d_i \times d_j}{2m}$ $\delta(c_i, c_j)$
$= \frac{1}{2m} \sum (a_{ij} - \frac{d_i d_j}{2m})$
$\delta(x, y) = 1$ $x = y$
$= 0$
c_i: Comm of i
c_j: " " j
$G(V,E)$
$m = |E|$
$a_{ij} = 1$
$|C|$ = Total no. of Community
$\sum_{i=1}^{|C|} \delta(c_v, i)\, \delta(c_w, i)$ (v,w)
> 1 1/0
$2m-1 \approx 2m$ $m \to \infty$
c
NPTEL

So, now modularity says modularity is denoted by Q, ok is denoted by Q is basically 1 by 2 m. What is m? m is mod E. 1 by 2 m, summation of i comma j which is basically V comma V, meaning you basically take all pairs of nodes, all pairs of nodes in the graph, right. We have this graph G which is V comma E, ok.

a ij, a ij, what is a ij? a ij is the entry in the adjacency matrix, i comma jth entry of the adjacency matrix. So, if a ij equals to 1 meaning that i and j are connected in the original graph, right. So, this is the actual number of edges between i and j. In a graph which does not have multi edges, this would always be 1 or 0, ok. This is the actual number of edges between i and j.

And what is the expected number of edges between i and j? We have already calculated, d i times d j by 2 m. 2 m minus 1, 2 m minus 1 is basically same as 2 m, when m tends to infinity for a large graph, ok. So, this is the modularity. But remember so far we have not considered the community information. What I have started, what I have said earlier that given a community structure C, right given a community structure C, I will check whether you know within the community nodes are how nodes are connected.

I mean what is the actual number of edges within the community of the original graph, and what is the expected number of edges within the same community in the random graph, right. So, I then, so in this metric, we have not considered the community information. So, we need to consider that. So, we then define, we add another indicator variable delta, right, delta. So, this is times delta C i comma C j. What is the delta? Delta i say delta x y, delta is basically an indicator again, kind of an indicator variable if x equals to y, this would be 1, otherwise 0.

And what is C i? C i is the community of node i. What is C j? C j is the community of node j. And this is multiplied by this quantity. What does it mean? It means that I will only consider this quantity when C i and C j are same meaning that i and j belong to the same community. So, if i and j belong to the same community, I will compute this the this metric, ok.

So, for a particular community, I will take all pairs of nodes, right, I will take all pairs of nodes in that community and then I will measure this one. So, this is modularity, ok. So, you actually can make it even concrete. So, so far this is you know this is based on all pairs of nodes and so on and so forth. We can even make it little concrete by define by you know redefining it in a different manner. So, let us try to understand what I am trying to say.

So, let us define let us define another thing, let us define. So, let us assume that there is more C number of communities, total number of communities; total number of communities that have been detected is mode C, ok. So, if it is the case, then if I denote all such communities by i, i is the index, right and delta say delta C v i, right delta C you know w, i. What does it mean? It means that for a particular pair of nodes v, w, right I will check all the communities.

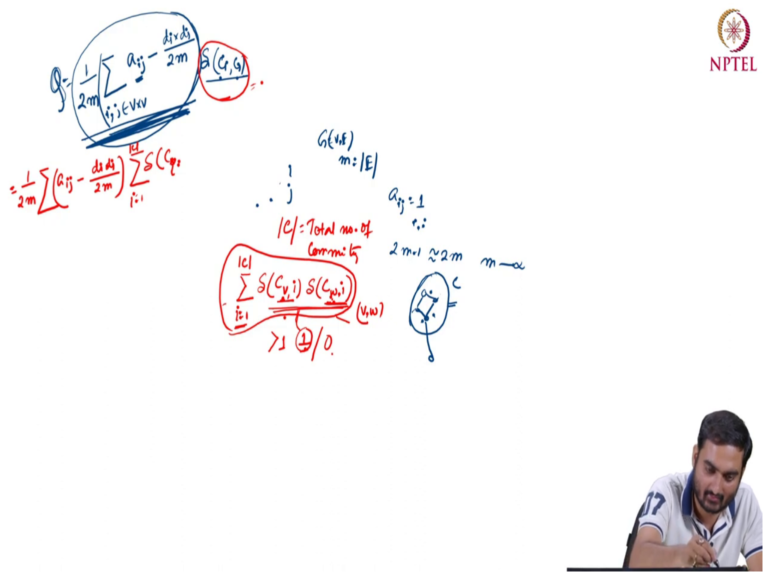
Remember, this is for disjoint communities, ok. So, a node can be a part of only one community. So, v and w can be a part of, v can be a part of one community, w can also be a part of one community, right.

So, if I check for all the communities for all the mod C number of communities, I check whether node v belongs to the community and node w belongs to the community. This will only return, this will return 1. This will return either 1 or 0. This will return 1, when v and w belong to the same community, when both v and w belong to i.

This should not be greater than 1; although it is summation, but it should not be greater than 1, why? Because v can only belong to one community, w can only belong to one community, ok. So, if both v and w belong to the same community then this would be 1, otherwise 0, ok. So, this is same as this one. This also returns 1 or 0, when i and j belong to the same community. This will also return 1 or 0, if v and w belong to the same community, ok.

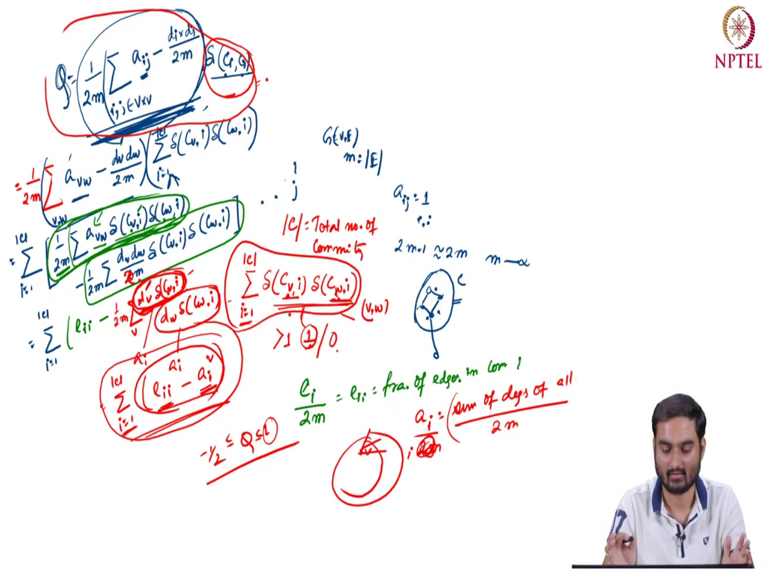
So, what I am doing here I am actually replacing, ok I am replacing this, ok, the remaining part will remain same d i, d j by 2 m, right.

(Refer Slide Time: 28:31)



And then we will have summation, ok i equals to 1 to mod C, delta C v, delta C i, ok. So, we have now index mismatch, so let us use different index, ok. So, let us use different index. Let us, ok let us just remove it, ok.

(Refer Slide Time: 28:59)


NPTEL
G(V,E)
m: |E|
$a_{ij} = 1$
|C| = Total no. of Comm.ts
2 m-1 ≈ 2m
m → ∞
(v,w)
>1 1/0
e_{ii} = fra. of edges in com i
a_i = sum of degs of all / 2m
$-1/2 \le Q \le 1$

459

So, we have a v w for all such pair v, w, right minus dv dw by 2 m, summation of i equals to, i tends ranges from 1 to mod C delta C v i delta C w i, ok.

So, what I will do? I will keep this summation outside, this one. So, this would be summation over i equals to 1 to mod C, right 1 by 2 m, ok a p w, right. And what I will do also I will multiply this with because remember this is two multiplication, right a times b into c. So, this is a times b into, a minus b into c. So, if this is a into c minus b into c. So, this would be delta C v comma i, delta C w comma i, right.

So, this is one part. And then we will have another part which is again 1 by 2 m, summation dv dw by 2 m, right delta C v i delta C w i, ok. So, now, right; so, now, think about it. What is this one? So, this is again; what is this quantity, this one? So, it is basically saying that, you just focus on this part, ok. So, a v w, which is the adjacency matrix in a value, so when this would be 1, meaning both v and w belong to the same community, I will only consider a i's in that case, a v w's in that case.

So, this is basically total number of edges in community i, think about it. It is a total number of edges within the community i, and right. And then, so let us say let us say the total number of edges within the community i is denoted by e i, ok. Now, then we have this 2 m. So, this is fraction of edges in community i, right. And let us denote it by e ii, ok. So, this is basically this total quantity is e ii. What is e ii? e ii is the fraction of edges in community i. This is basically intra community edges, fraction of intra community edges in i, ok.

Now, let us try to understand what is this one; what is, ok. Now, let us look at this one, ok. You can actually further break it down into two parts, ok. You can break it into this way, right. So, 1 by 2 m, ok d v delta C v comma i, right so, this will not be 2 m, right, this will be 2 m. So, you have 2 m, ok that is fine, this is 2 m, alright.

We already have a 2 m outside, right, ok so, d v delta C v i and d w delta C w i. So, what is d v delta C v i? So, it basically says that if v belongs to community i, I will take the degree and this is sum over all the vertices. So, in a community let us take a node v, and if this, since v belongs to the community i, I basically take the degree. And I basically take the sum of all the degrees within the community, ok.

So, this is sum of all the degrees within the community. And this is also same, sum of degrees of all the nodes in the community, sum of degrees of all the nodes in the community. So, if I denote a i as the sum, so and we also have this 2 m, right. So, 2 m is a; so basically this is sum of degrees of all nodes in i divided by 2 m. This is just a fraction, normalization of fraction.

And we denote this a i as this one, ok. So, this would be a i, this would also be a i. So, the final formula is i equals to 1 to mod C e ii minus a i square, ok. Now, this a i is basically a fraction, a normalized version of the sum of degrees of all the nodes in the community. And what is this? This is the total number of, this is again fraction of intra community edges with respect to community i. So, this is again as concrete version of the modularity, ok.

Now, this is node centric definition. We take all pairs of nodes. And this is community centric definition because we take a particular community i and then we measure the intra community edges, and sum of degrees of all the nodes in that community, ok. So, when you are given a community structure, we choose a particular community and then for that community you measure this thing, and this one, you measure the same quantity for other communities, you take the sum, and that will give you the value of the modularity, ok.

So, there are theoretical papers which basically showed that modularity maximum machines NP-hard and then modularity ranges between minus half to 1, right. If you have a better community structure, it should the modularity tends to be 1, otherwise it tends to be minus half and so on, right. So, this is about modularity.

In the next part of this chapter, we will discuss how you can use modularity for community detection. So, here we stop by saying that given a community structure how do we evaluate the quality of the community. Now, let us, and then, we will discuss given a network how can we use modularity for community detection, ok.

Thanks.