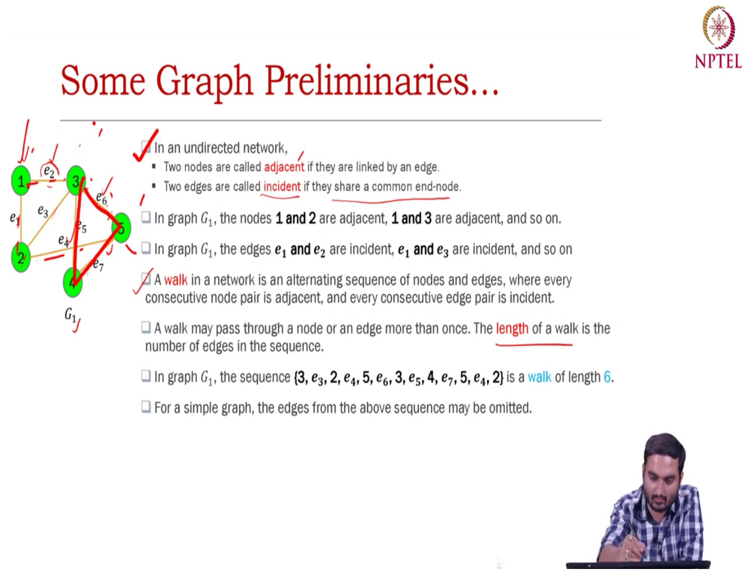
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 02
Lecture - 08
Lecture - 03

We have been discussing about you know different network measures, right. We have so far we have looked at degree distribution, right; why degree distribution is an important metric; we have also discussed power law, scale free property and so on. Now, we move to some of the other preliminaries of a graph, ok. And, you know those who have taken graph theory course right in the past, for them this is basically a recap.

But, for those who you know do not have any idea about the graph right graph theory or whatever the foundation of a graph, then this terminologies you know might be interesting. And, we will use these terminologies throughout the course.

(Refer Slide Time: 01:13)



Some Graph Preliminaries...

☑ In an undirected network,
- Two nodes are called adjacent if they are linked by an edge.
- Two edges are called incident if they share a common end-node.

❑ In graph G_1, the nodes **1 and 2** are adjacent, **1 and 3** are adjacent, and so on.

❑ In graph G_1, the edges e_1 **and** e_2 are incident, e_1 **and** e_3 are incident, and so on

☑ A walk in a network is an alternating sequence of nodes and edges, where every consecutive node pair is adjacent, and every consecutive edge pair is incident.

❑ A walk may pass through a node or an edge more than once. The length of a walk is the number of edges in the sequence.

❑ In graph G_1, the sequence $\{3, e_3, 2, e_4, 5, e_6, 3, e_5, 4, e_7, 5, e_4, 2\}$ is a walk of length 6.

❑ For a simple graph, the edges from the above sequence may be omitted.

So, in an undirected network we will call something called adjacent, right. We will say that two nodes are called adjacent if they are linked by an edge. For example, this node in this graph G1 node 1 and 3 they are adjacent because they are connected by an edge e2, right. Two edges are called incident if they share a common endpoint right say edge e2 and edge e6 they share a common end point 3. Therefore, these two edges are incident, alright.

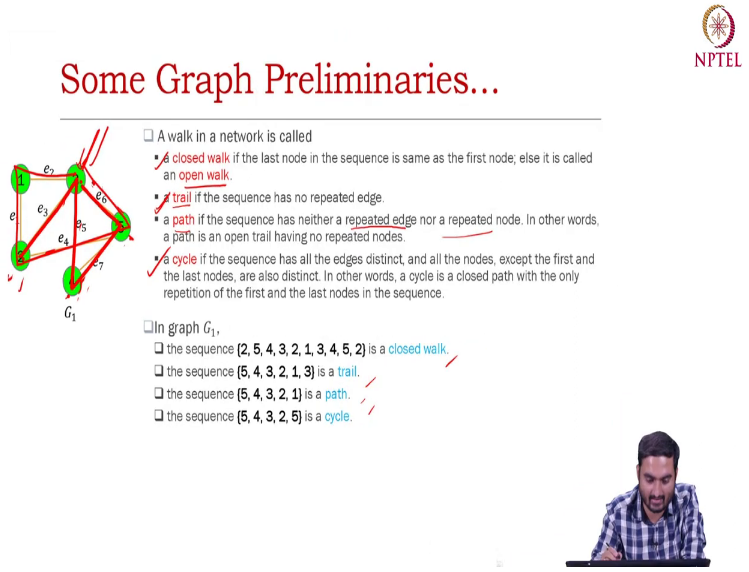
They are called incident so, adjacent and incident adjacent with respect to nodes incident with respect to edges. What is called a walk, right? So, walk in a network is an alternating sequence of nodes and edges where every consecutive node pair is adjacent and every consecutive edge pair is incident, right. So, this is a very you know very loose definition ok, will gradually make it tight or make it more constrained, right.

So, what is the walk? You basically start from a particular node 3, right and you follow edges. So, from 3 you can move to 5 for example, move to 5, then move to 6, move to 3, again move to 5 again, move to 2, move to 3 and 1 and 2 and 3 and so on. This is a walk ok. I mean it is not the case that you start from 3 and then you suddenly you know. So, say you start from 1 and then you suddenly move to 5 this is not possible because 1 and 5 are not directly connected, right.

So, you I mean to in order to move to 5 you basically move from 1 to 3 then 3 to 5 and so on and so forth, right. So, this is a walk. Now, in the walk as you have seen that a node can also repeat and edge can also repeat, right. So, for example, if you start from 3 from 3 to 5 from 4 from 3 you can again move from 4 to 3 that is possible, that is allowed. From 3 to 5 you can again come to 5. So, in that case node 5 will be repeated and this edge e 6 will also be repeated.

So, walk is a very you know very top level definition, there is no constraint, right and what is the length of a walk? The length of a walk is the number of edges that you have traversed through this walk, right. For example, if you write if you say move from say from 3 to 5 to 4 to 3 how many edges you have covered? 1, 2 and 3, right number of edges is the length of the walk, ok.

(Refer Slide Time: 04:13)



Some Graph Preliminaries...

❑ A walk in a network is called

- a closed walk if the last node in the sequence is same as the first node; else it is called an open walk.
- a trail if the sequence has no repeated edge.
- a path if the sequence has neither a repeated edge nor a repeated node. In other words, a path is an open trail having no repeated nodes.
- a cycle if the sequence has all the edges distinct, and all the nodes, except the first and the last nodes, are also distinct. In other words, a cycle is a closed path with the only repetition of the first and the last nodes in the sequence.

❑ In graph G_1,

❑ the sequence {2, 5, 4, 3, 2, 1, 3, 4, 5, 2} is a closed walk.
❑ the sequence {5, 4, 3, 2, 1, 3} is a trail.
❑ the sequence {5, 4, 3, 2, 1} is a path.
❑ the sequence {5, 4, 3, 2, 5} is a cycle.

So, now, let us make it little you know constrained, right. So, there is something called a closed walk. What is a closed walk? A closed walk is a walk whose start node and the end node are same. So, you start from 3, then 5, then 4, then 3, this is a closed walk. It can also be 3, 5, 4 you know 3, 2, 5, 3 this is also a closed walk you know, right. So, this is a closed walk.

And, what is open walk? As intuitively, right so, you can understand that in an open walk; in an open walk the source node and the destination node the final node they are different ok. So, you start from 3, then 5, then 4, right, then 3, then 2 and that is all. So, you start from 3 and you stop at 2, this is a closed walk, this is an open walk, right.

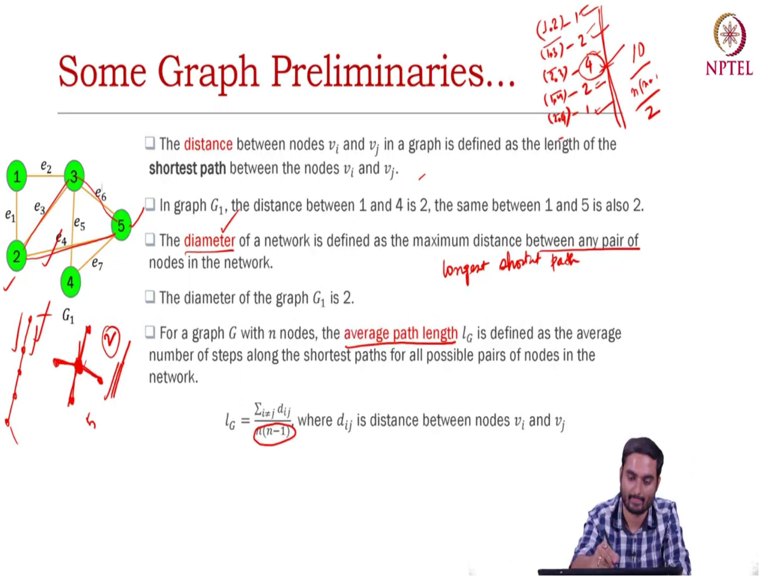
So, what is a trail? A trail now we are again making it more constrained. So, walk was a very general concept. Then we said that you know if source node and destination node are same then it is closed otherwise it is open. Now, I am saying that if a walk has no repeated edge, then this is called a trail ok. There is there should not be any repeated edge, right.

For example, you say you know start from 3 right, then from 3 to 5, then 4 to 3, then 2 this is a trail because here a node gets repeated, but no edges are repeated. This is a trail. Now, what is a path? A path is again a walk where no edge and no node will be repeated. So, the one that I mentioned 3 to 5 to 4 3 to 2 this is not a path why? Because the node 3 has been repeated.

So, trail is constant in terms of edges, path is constrained in terms of both nodes and edges, right. So, what is a cycle? A cycle is essentially you know a path which is closed, right. So, a path means no nodes, no edges will be repeated and if it is a closed path meaning the source node and the destination node should be same. Say for example, you start from 3 to 5 to 2 to 1 to 3, this is a cycle ok.

So, you see examples you know written on the slide. So, you know these are not that complicated.

(Refer Slide Time: 07:15)



Some Graph Preliminaries...

- □ The distance between nodes v_i and v_j in a graph is defined as the length of the **shortest path** between the nodes v_i and v_j.
- □ In graph G_1, the distance between 1 and 4 is 2, the same between 1 and 5 is also 2.
- □ The diameter of a network is defined as the maximum distance between any pair of nodes in the network.
- □ The diameter of the graph G_1 is 2.
- □ For a graph G with n nodes, the average path length l_G is defined as the average number of steps along the shortest paths for all possible pairs of nodes in the network.

$$l_G = \frac{\sum_{i \neq j} d_{ij}}{n(n-1)}$$ where d_{ij} is distance between nodes v_i and v_j

longest shortest path

What is the distance between two nodes? The distance between a pair of nodes is the shortest path between two nodes. So, remember when I talk about the path nodes and edges will not be repeated, right and the length of the path would be the number of edges that you traverse through.

Now, between two nodes there can be multiple paths one can be longer one can be shorter, right. For example, between 2 to 5 you see here you can directly move from 2 to 5 through this edge or you can move from 2 to 3 then 3 to 5, right. So, there are two paths which one is shortest? This one is shortest from 2 to 5. So, the distance is the shortest path between a pair of nodes.

You are given a pair of nodes and you are asked to ask to measure the distance between that pair you will say that I will look at the shortest path and that should be; that would be my distance, right. How do you measure shortest path? There are many algorithms Dijkstras algorithm and so on will you know those are standard algorithms that people generally use, but that is not our concern here.

What is the diameter? The diameter of a network is the longest shortest path between any pair of nodes in the graph, right; a longest shortest path. So, remember longest shortest path between any pair of nodes in the network. What does it mean? So, remember diameter is a property of graph whereas, distance is a property of a pair of nodes, right.

So, what is the diameter? You take all pairs of nodes, now for every pair you measure the shortest path and you then you choose that pair whose shortest path is maximum longest among all the other pairs. Let us say; let us say you have again hypothetical example 1, 2, 1, 3, 2, 3, 1, 4, 2, 4, right. Shortest path is 1 between 1, 3 – 2; 4 say for example, 2 and 1, these are the shortest paths.

So, what is the diameter of the network? Diameter is 4, the longest shortest path, ok. What is the interpretation of a diameter? Diameter basically says that how far two nodes can decide in a particular network, ok right and that is the maximum that is the worst case scenario, the maximum distance that is possible. So, whenever you take any example or whenever you think of an application you always need to remember that there is at least one such path one such longest path which exists, right.

It may happen that the other shortest paths are very small say length 1, length 2, length 3, but there is a shortest path whose distance is 10 for example, ok. So, this diameter also is useful to understand the property of a network. Let us say the diameter is small. There are two networks whose diameter is small and there is another network whose diameter is very large, right.

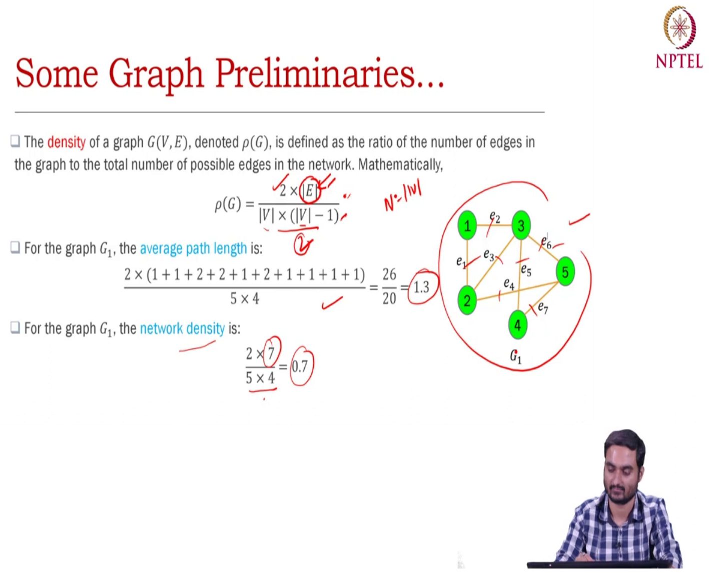
If the diameter is small you can think of it as a network like a star because in a star network you see that all these peripheral nodes have you know shortest path 2 because from here to here you can need to traverse through this path, this particular node, right. So, all most of the pairs of nodes have shortest path of 2, but from this node the shortest path is 1. So, the diameter is 2, right versus if you have a network like this, this is called a line graph or a chain, the longest path is from here to here.

So, let us say I told you that there are 5 nodes in a network; there are two networks both of them have 5 nodes. One network has diameter 2, other network has diameter 4. So, you can immediately understand that the network which has diameter 2 may have a star like structure and the network which has diameter 4, has a chain like structure ok like this.

Now, let us look at average path length. What is average path length? Average path length is basically simple. You be you take all pairs of nodes, you take the shortest path and the corresponding length, right? You sum them up and you normalize it, right. For example, here as you see we take all pairs, all pairs of nodes we sum them up 4 plus 2 plus 1 plus 2 plus 1, 8, 9, 10 and we divide it by n into n minus 1 by 2.

What is n into n minus 1 by 2 n into n minus 1 by 2 is the number of possible edges that can be; that can be there in a graph, right; n c 2 n choose 2, right. So, you normalize it by either n into n minus 1 or n into n minus 1 by 2 does not matter, just a average path length, right. So, the average path length for this network would be lower than you know a chain ok. Diameter average path length both of them are network centric properties ok, right.

(Refer Slide Time: 13:27)



Some Graph Preliminaries…

❑ The density of a graph $G(V,E)$, denoted $\rho(G)$, is defined as the ratio of the number of edges in the graph to the total number of possible edges in the network. Mathematically,

$$\rho(G) = \frac{2 \times |E|}{|V| \times (|V| - 1)}$$

N:=|V|

2

❑ For the graph G_1, the average path length is:

$$\frac{2 \times (1+1+2+2+1+2+1+1+1+1)}{5 \times 4} = \frac{26}{20} = 1.3$$

❑ For the graph G_1, the network density is:

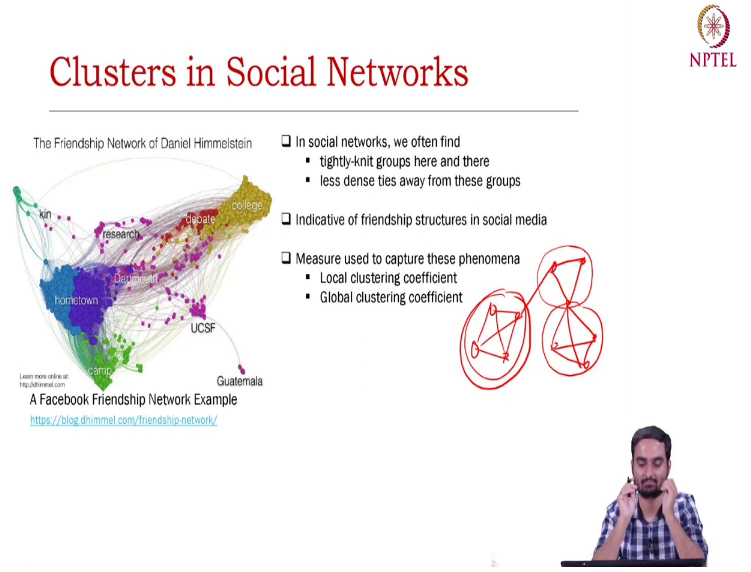
$$\frac{2 \times 7}{5 \times 4} = 0.7$$

Now, let us move to the next property which is called the density of a network. So, a density of a network, again this is very simple. The density is essentially the number of edges present in the graph and the maximum possible edges that can be there, right. So, number of edges present in the graph is mod of E, right.

So, mod E is the cardinality of the edge set, right and what is the possible number of edges there? It can it would be N into N minus 1 by 2. Now, what is N? N is mod V. So, mod V into mod V minus 1 by 2, right. So, I am moving to the numerator. So, this is the density. So, if the density is high meaning that the numerator and the denominator are kind of same ok, meaning that the graph is highly dense.

Meaning that you know all the nodes are connected to each other. It is like a click, a completely connected graph, right. So, I mean you see this example here right average path length for this particular graph it is shown as 1.3 and you know network density is you know. So, if you look at this one 1, 2, 3, 4, 5, 6, 7 – 7 edges and how many nodes are there? There are 5 nodes. So, 5 times 4 in by 2 right this is 0.7. This is kind of a very well connected graph, ok.

(Refer Slide Time: 15:04)



Clusters in Social Networks

The Friendship Network of Daniel Himmelstein

A Facebook Friendship Network Example

https://blog.dhimmel.com/friendship-network/

❑ In social networks, we often find
  - tightly-knit groups here and there
  - less dense ties away from these groups

❑ Indicative of friendship structures in social media

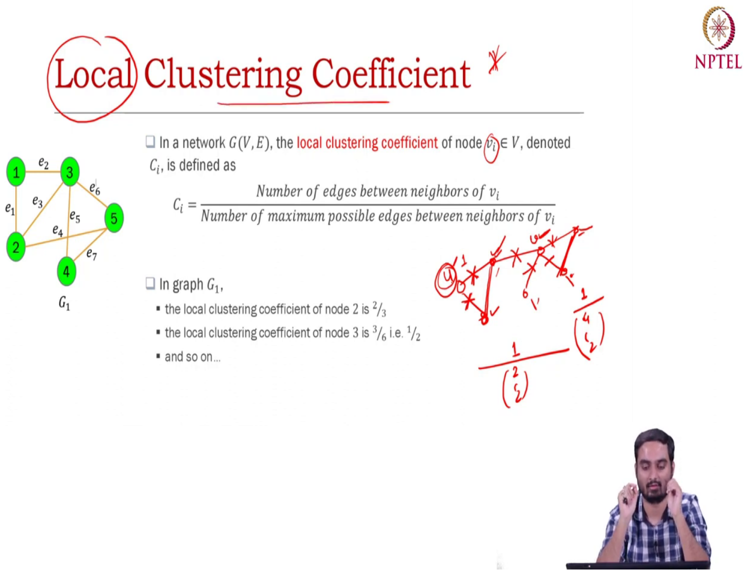
❑ Measure used to capture these phenomena
  - Local clustering coefficient
  - Global clustering coefficient

Now, let us look at another interesting property. This is called; this is called the cluster of a network. Cluster of a network is essentially groups of nodes which are densely connected to each other. Now, I am making it vague intentionally because we have a separate chapter on clusters graph clustering, graph community detection and there will try to make it concrete that what do you mean by cluster.

Now, for the time being let us assume that you know you have right a graph like this ok a graph like this and you see that this is this can be treated as one cluster, this can also be treated another cluster. You can also think of it as another cluster. So, cluster you can relate it to edge density, right. You can think of sub graphs within a sub graph nodes are densely connected, right across sub graphs nodes are sparsely connected.

So, the edge density within the induced sub graph, right of I hope you remember what is induced sub graph. We discussed in the first chapter right the edge density within the induced sub graph of a graph if this is high then you say that this sub graph is a cluster ok. We will make it concrete in one of the chapters later, right.

(Refer Slide Time: 16:32)



Local Clustering Coefficient

❑ In a network $G(V,E)$, the local clustering coefficient of node $v_i \in V$, denoted C_i, is defined as

$$C_i = \frac{\textit{Number of edges between neighbors of } v_i}{\textit{Number of maximum possible edges between neighbors of } v_i}$$

❑ In graph G_1,

- the local clustering coefficient of node 2 is $^2/_3$
- the local clustering coefficient of node 3 is $^3/_6$ i.e. $^1/_2$
- and so on...

$\frac{1}{\binom{4}{2}}$

$\frac{1}{\binom{2}{2}}$

Now, let us look at some of the properties that again node sending properties that this clustering you know the nodes within a cluster generally possess, right. Then this property is very very important. We will keep talking about this property in the remaining part of this lecture and the subsequent lectures as well.

So, there is something called clustering coefficient. The clustering coefficient is a property of a node. What is the clustering coefficient of a node? Clustering coefficient of a node v i is the number of edges within the neighbors the first of neighbors of a node divided by the possible number of edges among the first of neighbors of a node ok.

Let us look at example. So, let us take you know let us take let me draw something ok. Let us see, let us take this one – node u, node v. We will measure clustering coefficient for both u and v. So, for u there are two neighbors, this one and this one, ok. So, what I will see? I will in the numerator I will see the number of edges number of actual number of edges among the first of neighbors of u. So, these are two first of neighbors and how many edges are?

There is only one edge. So, this would be 1 in the numerator, denominator would be total number of possible edges between these two I mean among the neighbors. So, there are two nodes the possible edges would be 2 c 2 is also 1, ok. So, the clustering coefficient of u is 1. What about v? v has 1, 2 and 3 and 4, 4 one hop neighbors, right and let us look at and remember one thing very carefully when we calculate number of pages we are not considering those edges which are connected which are connecting the node the node under consideration with the neighbors. We are not considering this edge and this edge, ok.

Similarly, when we calculate clustering coefficient of v we do not consider this edge this edge this one and this one we will only consider which one? This edge we will only consider edges which are present among the neighbors ok. So, how many edges are there? Only one and how many possible edges are there are 1, 2, 3, 4 neighbors. So, 4 c 2, right.

Now, this is clustering coefficient. Now, let us the and this is also called local clustering coefficient, I will tell you why. This there is another counterpart called global clustering coefficient, but when you talk about clustering coefficient we generally refer to local clustering coefficient. Let us try to understand the interpretation of it, right. If I say that my clustering coefficient is higher than your clustering coefficient.

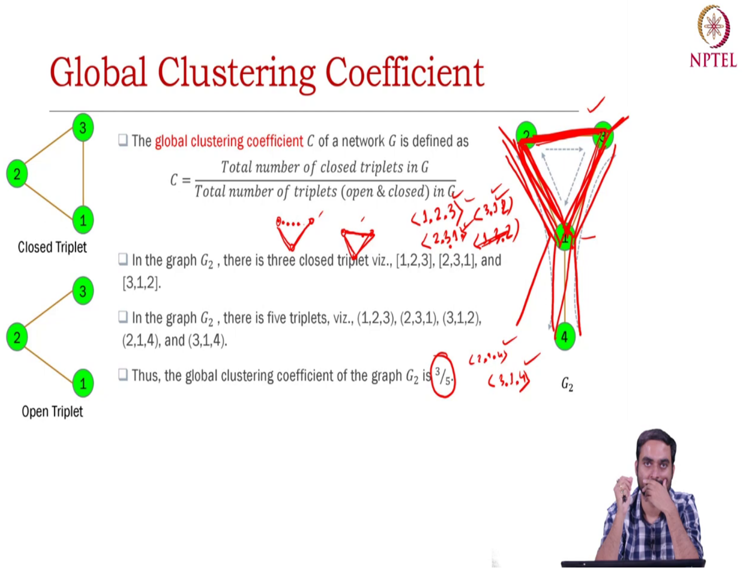
It means that my neighbors are highly connected. They may have high understanding among each other. Let us think of a friendship network; if my clustering coefficient is high that means, my neighbors are highly connected. Of course, I am also connected to my neighbors that is why they are my neighbors and my neighbors are also highly connected. So, and let us take another example.

Another case where say let us take another node like say u as a node and your neighbors are not highly connected therefore, your clustering coefficient is less. It indicates that I might I am stable because my neighbors are connected and my meaning that my neighbors have high understanding among each other due to this friendship relations. And, of course, since I am also their friends. So, this group as a whole me and my neighbors they may form a cluster.

In contrast to your case your clustering coefficient is less, meaning your neighbors are not that related. They are not friends maybe, right. So, it is highly likely that gradually you also lose in your interest and you move out, ok. Using the clustering coefficient a whole bunch of algorithms have been proposed.

We will discuss another such algorithm one such algorithm in the community detection chapter there is something called permanence which our group you know proposed way back 2014. This permanence algorithm is based on the clustering coefficient notion ok.

(Refer Slide Time: 21:19)



Global Clustering Coefficient

❑ The global clustering coefficient C of a network G is defined as

$$C = \frac{\textit{Total number of closed triplets in } G}{\textit{Total number of triplets (open \& closed) in } G}$$

❑ In the graph G_2 , there is three closed triplet viz., [1,2,3], [2,3,1], and [3,1,2].

❑ In the graph G_2 , there is five triplets, viz., (1,2,3), (2,3,1), (3,1,2), (2,1,4), and (3,1,4).

❑ Thus, the global clustering coefficient of the graph G_2 is $\boxed{3/5}$.

Let us move on. The next concept is global clustering coefficient. We have discussed what is local clustering coefficient, now we discuss what is global clustering coefficient. The idea is same the global clustering coefficient is a property of a graph where the local clustering coefficient is a property of a network. So, you may say that you know it is very simple I take I measure the local clustering coefficients of nodes and I take the average, average local clustering coefficient of all the nodes in the graph and that would give you the global clustering coefficient of course. This is one way.

Another way is we look at something called triplet. What is a triplet? A triplet is a substructure with three nodes, right. This is a triplet this can also be a triplet a triplet is a substructure with three nodes. It is a motif, right and you know these two triplets are different. Why? Because this triplet you see this is not closed, but this is closed, right this is closed. Had it been an edge like this it would have been a closed triplet.

What we see we measure the number of closed triplet in the graph divided by we will then divide it by the total number of triplets can be closed, can be open and that would give you the clustering coefficient global clustering coefficient of a graph. Let us take an example let us take this graph right. So, here you see that one triplet is 1, 2, 3, 1, 2, 3 right. So, of course, this is a closed triplet.

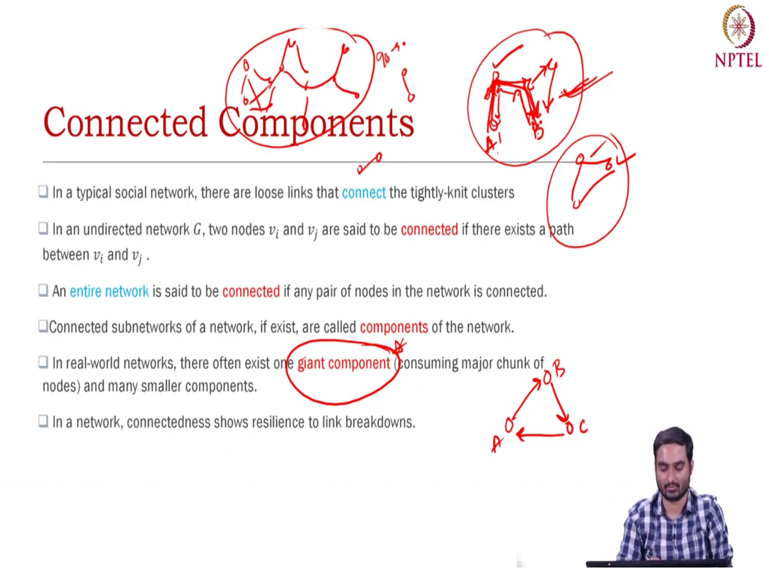
You can also think of from this closed triplet you can actually you know derive 3 triplets. Each of these triplets centered around three different nodes. So, one can be 1, 2, 3; other can be 2, 3, 1. So, here the 2, 3, 1 the central node is 3 the other can be you know 1, 3, 2 sorry. So, 1, 2, 3, 2 1, 2, 3, 2 2, 3, 1 and 3, 1, 2, 3, 1, 2 right 3, 1, 2 so, the center is 1.

So, from a triangle a closed triplet is basically a triangle. From a triangle you can actually derive three triplets centered around three different nodes, right. So, and so these are three triplets, let us look at other triplets. So, this is another triplet 3, 1, 4 and this is another triplet 2, 1, 4 right. So, how many triplets are there? 1, 2, 3, 4, 5 total number triplets are total number of triplets is 5.

How many closed triplets are there? There are three closed triplets. Remember every triangle contributes to three closed triplets, ok. You may also argue that look 2 ,1, 4 is an open triplet; 4, 1, 2 can also be open triplet. No, why? Because actually you can differentiate between triplets based on the central node, the middle node. The middle node is same for both the cases ok.

So, the global clustering coefficient is 3 by 5, ok. Now, what does it mean? So, number of closed triplets indicates number of close friends. Now, think of a graph where you your global clustering coefficient is very high meaning that you have a lot of triangles right means graph this graph is very dense, ok. You can infer many things from the value of the global clustering coefficient.

(Refer Slide Time: 25:38)



Connected Components

❑ In a typical social network, there are loose links that connect the tightly-knit clusters

❑ In an undirected network G, two nodes v_i and v_j are said to be connected if there exists a path between v_i and v_j .

❑ An entire network is said to be connected if any pair of nodes in the network is connected.

❑Connected subnetworks of a network, if exist, are called components of the network.

❑ In real-world networks, there often exist one giant component (consuming major chunk of nodes) and many smaller components.

❑ In a network, connectedness shows resilience to link breakdowns.

Let us look at the last concept in this lecture which is called connected component ok. What is connected component? Connected component of a graph is a component where there are nodes of course, but nodes are connected within a common within a component, but they are separated from another component. If a graph is disconnected how do you know that a graph is disconnected? If there is at least one shortest path between every pair of nodes then the graph is connected.

Let us take a graph like this. This is one graph, but two components. If you take one node from here another node from here you will not see a shortest path of finite distance right. So, this is a disconnected graph. So, this is one component, this is another component. Now, depending upon the notion of connectedness, you can think of two different types of components – one is called strongly connected component.

In the strongly connected component nodes are you know I mean the this notion of strongly connected component and weakly connected component comes from the directionality of an edge. If you have an if you have a director graph right, then if you take a component and you and if you take a pair of nodes, and if you see a path between all pairs of nodes – now remember when I say that a graph is directed say this is the direction right, this is the direction.

So, from A to B, you can move through the direction of the edge, but you cannot move from B to A because the direction does not permit you ok. So, let us take an example a component in an in a directed graph where there exists a path between all pairs of nodes. So, from A to B there is a path, from B to A there is a path for all pairs of nodes. Then it is a strongly connected component meaning that you have an edge like this, you have a path like this, right.

Let us take an example of this one right A B C A cycle. This is a strongly connected component. From every node you can move to other node through the direction. Weakly connected component another notion where again for a directed graph if we forget about the directionality for a directed graph you think of it as an undirected graph, right.

An undirected graph, then you take every pair all pairs and you see a path. If a path exists then it is a weakly connected component. Think of this one. This is not a strongly connected component because from B you would not be able to move to A, but this is a weakly connected component because if you forget about the directionality you can move from B to this, right.

And, in the next chapter we will talk about something called giant component. What is a giant component? Giant component in a disconnected graph giant component is a component which constitutes 90 to 95 percent of nodes in a network. Think of a network like this right, network like this right and you have few other components.

This is a giant component because this constitutes 90 to 95 percent of the nodes present in the graph, ok. And, this giant and the concept of joint component is very important. We will see in case of random network theory random graph theory in the next chapter that how giant component emerges you know in the process of network evaluation.

So, we stop here. In the next part of the lecture which is basically a is a continuation of this chapter. We will discuss about something called centrality. It is a very important concept right centrality of a node and we will see how we you know use centralities interpret centralities for different applications, right. With this, I stop here and let us meet in the next chapter next lecture.

Thank you.