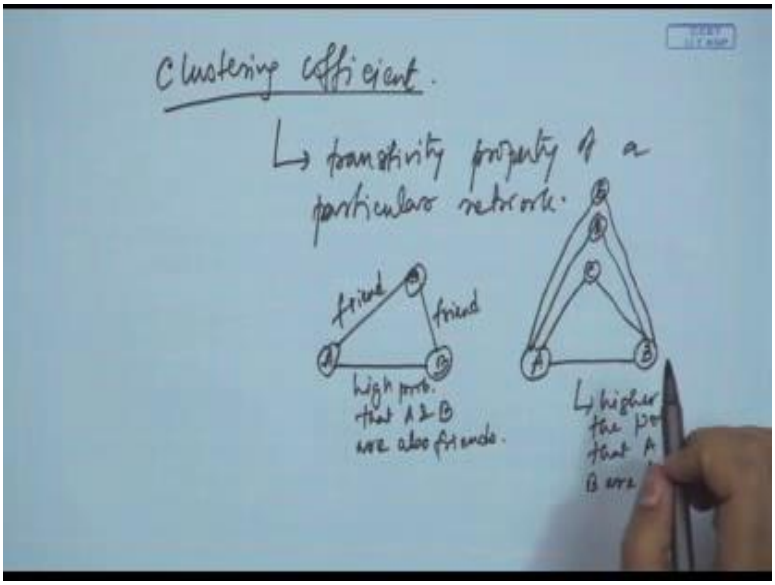
Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 03
Network Analysis – II

So, welcome back to this session on network analysis. So, we will continue with the second part of network analysis. And, today we will start with the second metric that we will get ourselves introduced to, and this is called clustering coefficient.

(Refer Slide Time: 00:34)

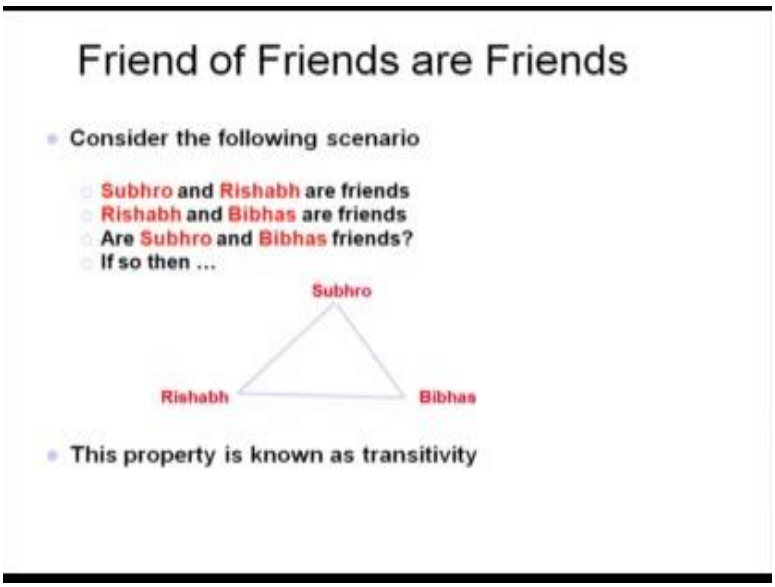


So, the clustering coefficient actually comes from the transitivity property of a network; transitivity property of a particular network. So, the idea is very simple that if you have, if there are two nodes in the network say A and B having a mutual friend say C, then there is a high probability that A and B are also friends in the social network. So, A and C was a friend, B and C was a friend, then there is a high probability. High probability that A and B are also friends. So, this is what is called the famous transitivity property of a network.

Now, this property actually comes from the observation that in many cases. So, this probability actually will be higher and higher, if you encounter situations like the one that I am drawing.

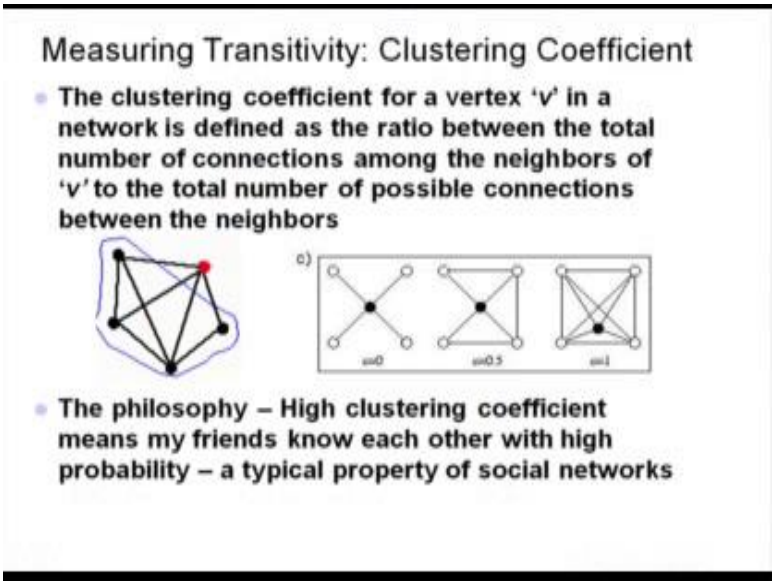
Suppose there are many such common friends, the larger the number of common friends between A and B, the higher is the probability. So, the large, higher is the probability that A and B are friends. So, the larger is the number of mutual friends between A and B, the higher is the probability that A and B are themselves friends. So, this idea actually is called the transitivity of a particular social network.

(Refer Slide Time: 03:01)



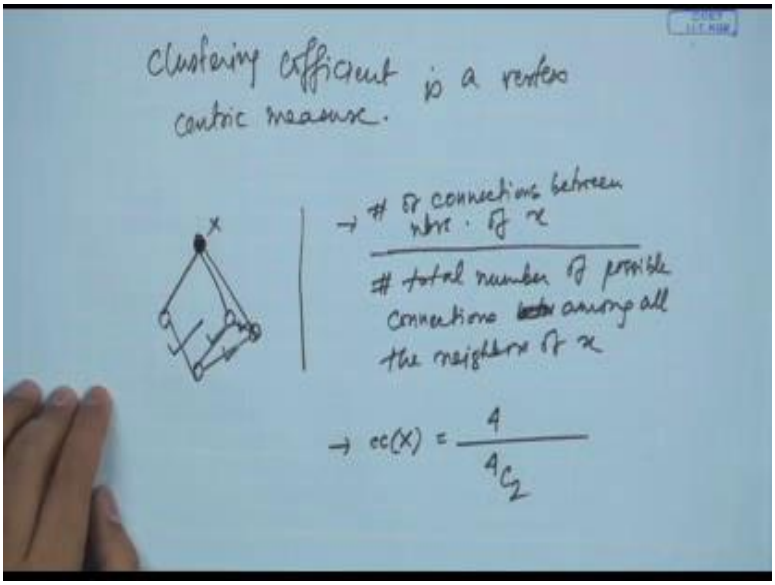
And, this transitivity can be quantified using something called the clustering coefficient. that we will define now.

(Refer Slide Time: 03:12)



So, the clustering coefficient can be measured for each vertex. So, it is a vertex centric measure.

(Refer Slide Time: 03:17)



So, clustering coefficient; clustering coefficient is a vertex centric measure. So, you can measure this value for each individual vertex separately. So what you do is, suppose you want to measure the clustering coefficient of a node say x here, you basically look into the neighbors of x. So, say x has 1, 2, 3 and 4 neighbors. You look into the total number of connections between the neighbors of x. So, in this particular example number of connections between neighbors of x; this is expressed as a fraction of the total number of possible connections between or among all the neighbors of x. So, this is basically the definition of clustering coefficient.

So, you look into the neighbors of a particular node x. So, if you are interested to find out the clustering coefficient of x, you look into the neighbors of x. Here for this particular example, you have 1, 2, 3 and 4 neighbors. So, you can write the clustering coefficient of the node x in this particular example is, among the neighbors there are how many edges? There is 1 edge, 2 edge, 3 edge and 4 edge. So, you have four divided by the total number of possible edges.

So, what is the total number of possible edges? It should be; so, since there are four nodes, there should be four C 2 possible edges. So, this is basically the clustering coefficient. This ratio actually defines the clustering coefficient of a particular node. So,

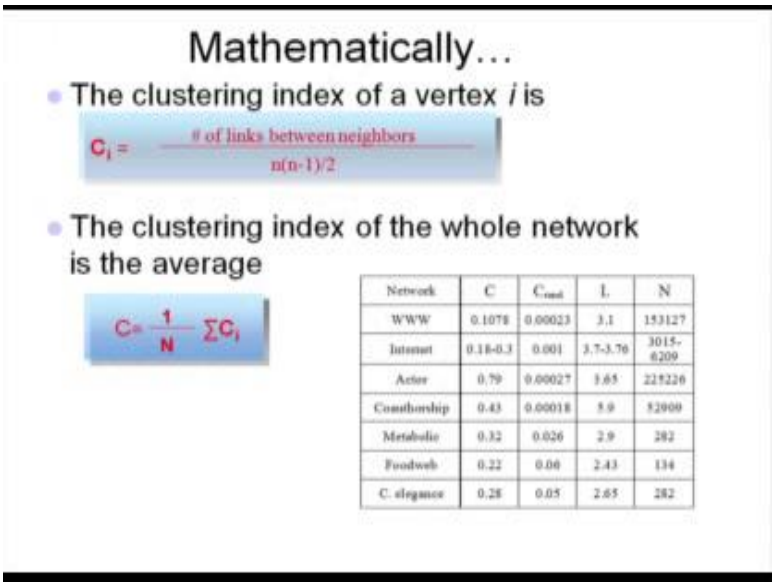
in this way you can measure the clustering coefficient for each individual node in the network. And, the clustering coefficient of the whole network is just an average of all the individual clustering coefficients of the different nodes.

So, so it is a node centric property. First of all, for each individual node you measure the clustering coefficient. Basically, for each individual node you try to estimate what is the extent of cliquishness, how complete or how cliquish the neighborhood of that particular node is. So, you try to express that as a fraction of the maximum cliquishness possible, and then this fraction can be estimated for each individual node. And then, for the whole network you just have to average out all the individual clustering coefficients for the different nodes. So, that is how you define the clustering coefficient of a particular network.

Now, for this part of the lecture I have shown you three examples. If you see in the slides, I have shown you three examples. There are three different examples, where the clustering coefficient for the first example for that black node, there in the center, should be equal to zero because there is no edge that exists between its neighbors.

So, similarly the clustering coefficient for the second example is actually 0.5 because there are three edges between the neighbors. There are three edges between the neighbors. And, there could be a possible of four C 2 edges. So, that is how you measure. Similarly, for the third example, the clustering coefficient is one because all the nodes that are neighbors of the black node are completely connected. That is how you actually measure the clustering coefficient for each individual node.

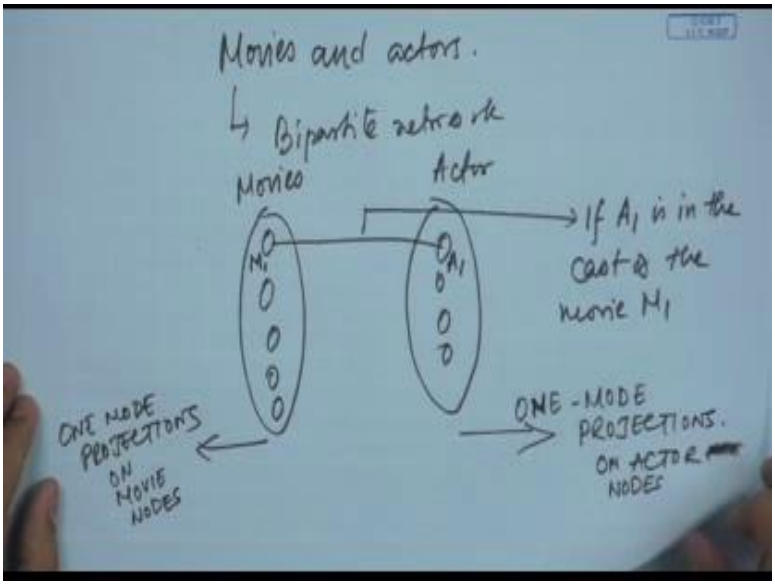
(Refer Slide Time: 08:01)



So, this is; in the next slide, I actually defined the formula more precisely. The same thing that I have written down in the text, now the interesting part is that you can measure. Once you have this quantity, you can measure the extent of transitivity of the different real world networks. So, and actually that was the thing that people were trying to do in early 2001, 2002. And, what they found is that so if you look into these networks the worldwide web, the internet, the co-authorship network, the metabolic network, the C. elegance network, you see most of them have clustering coefficients ranging from between, somewhere between 0 and 1.

But, what is interesting to note is that networks like co-authorship network has a very high clustering coefficient. That is, these are mostly social networks which have very high transitivity. That is, the idea is that if there are a lot of mutual coauthors between a pair of scientists, then it is highly likely that those two scientists have also coauthored a paper. So, that probability is quite high and is close to, roughly close to 0.43 as per the table shown in the slides.

(Refer Slide Time: 09:40)

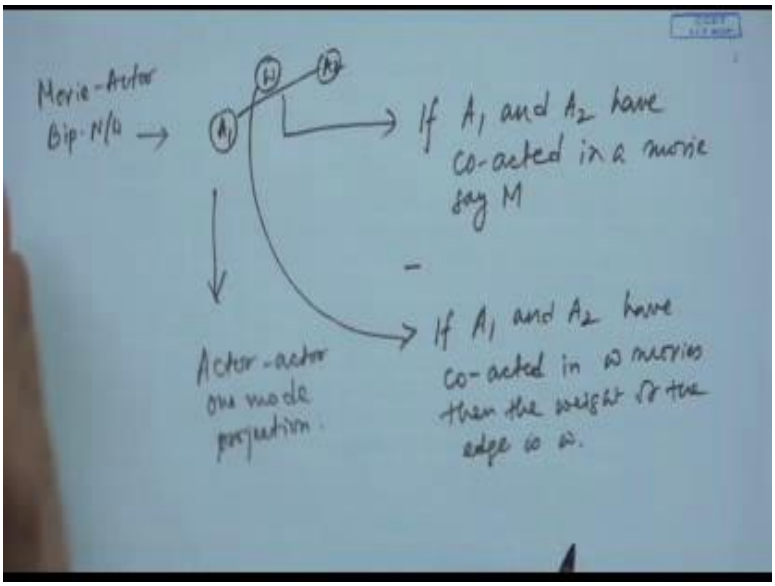


So, another interesting example is this actor network. This is again a, since this is an interesting example I would like to take up these and discuss a bit more. So, this example actually draws from the complex system of movies and actors.

So, you conceive first of all of a bipartite network, where one partition is basically the movies or the movie nodes and the other partition is basically the actors or the actor nodes. Now, you have some movies and you have some actors. Now, you draw an edge between a movie and an actor. If suppose, this say the name of this movie is M 1 and the name of this actor is A 1. You draw an edge between M 1 and A 1, if A 1 is in the cast of the movie M 1. So, this is how you construct a movie-actor bipartite network.

Now, from this bipartite network you can construct something called one mode projections. This one mode projections can be drawn; again on this side, one mode projections. So, these one mode projections can be drawn on actor nodes as well as on, sorry, on actor nodes as well as on movie nodes. So, what you would do in drawing the one more projection? So, let us try to define that in the next part.

(Refer Slide Time: 09:50)



So, suppose you have. So, from the movie-actor bipartite network, you construct one more projection say on the actor nodes as follows. Suppose there are two actors A 1 and A 2. You draw an edge between A 1 and A 2, if A 1 and A 2 have co-acted in a movie say M. Now, this graph as you can imagine can be a weighted graph. So, this can have a weight w. And, this w is nothing but if A 1 and A 2 have co-acted in w movies, then the weight of the edge is w. Now, in this way you can act, you can construct a actor-actor one mode projection.

So now, if you think of the movie-actor bipartite network, when you are constructing this projection on the actor nodes, what is happening? If you think carefully, what is happening is that for every individual movie there is a clique imposed on this network. So, all the actors in that movie will have an edge between them because they have acted in that movie. So, that forms a clique of actors for that particular movie.

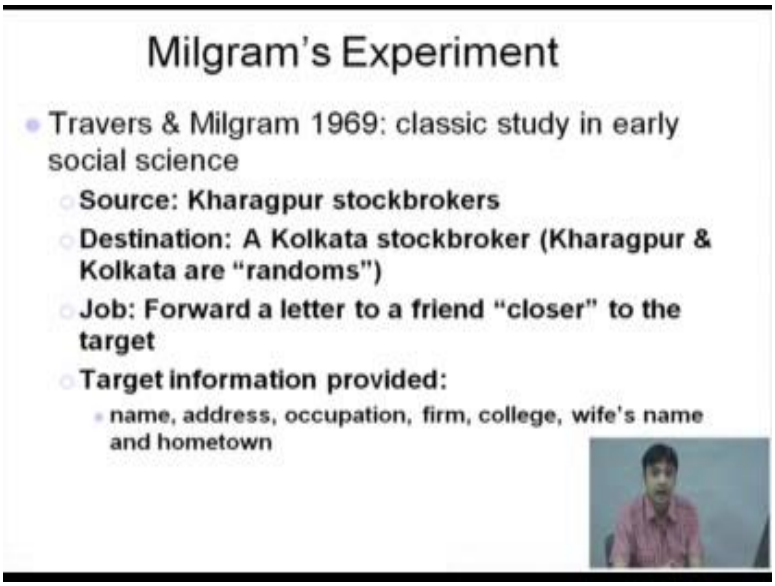
And for every individual such movie, you are imposing a clique on this network. So, that is why these kinds of graphs are usually pretty cliquish. And, that is why as you see the clustering coefficient of this network, as I have shown you in the slides, the clustering coefficient of this particular network, the actor-actor network is point seven nine, which is very high. So, the probability that there exist an edge between a pair of actors, if they have co-acted with many other actors is very high. So, that is what I wanted to actually point out.

(Refer Slide Time: 14:38)

So, then the next thing that we will talk about is this concept of small world and the 6 degrees of separation. So, so this idea is again very interesting. So, suppose look at the slides, suppose if I say that all late registrants in the complex networks course will be given ten marks bonus. So, how fast do you think will this information spread? In general, I would imagine that it would spread very fast among all the registrants of this course.

So if that is the case, then the question is that why is it? So that it spreads fast and why is it so that it actually spreads in the first place. And, to show that this actually happens, this spread actually takes place, the famous scientist Milgram designed a very interesting experiment, which he called the 6 degrees of separation experiment. So, what he actually did was something like this.

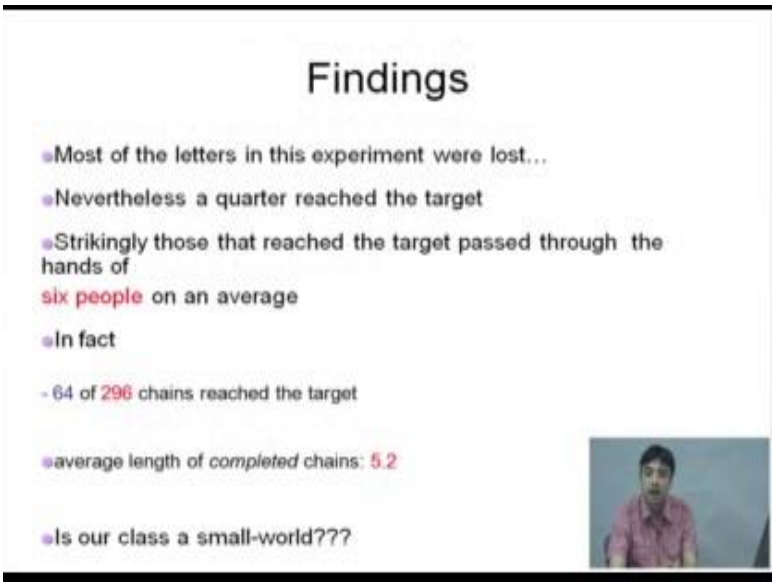
(Refer Slide Time: 14:46)



So, Travers and Milgram in 1969, they designed this classic study in Social Science. Actually, this is one of the very interesting in classic studies in Social Science. So, what they said is that suppose you have a source say Kharagpur, and there is a stockbroker in Kharagpur, who wants to send a message or a packet to some stockbroker in Kolkata. So, now this letter, what this person does, this stockbroker does is forwards this letter. So, he does not post this letter. He does not post it using the usual postal codes.

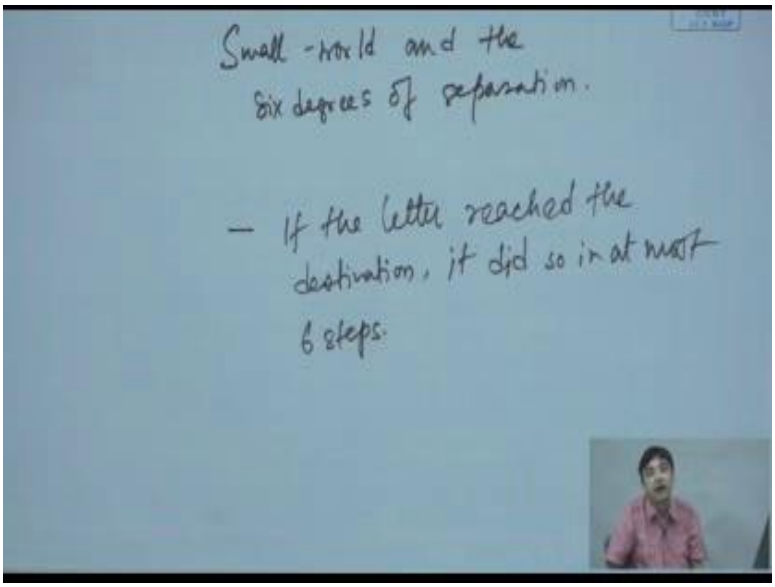
What he does is he passes this letter to one of his friends. So, this letter actually has the name, the address, etcetera of the destination stockbroker. So, what he does is the stockbroker takes up this letter and passes it to one of his friends, whom he believes would know the Kolkata stockbroker. The Kharagpur stockbroker picks up this letter and passes it to one of his random friends. So, and then he feels that this friend might be knowing the Kolkata stockbroker and would pass it on to him. So, now this friend will again pass it to some other friend of his, and that friend will again pass it to some other friend of his and in this way the chain might at some point complete, leading to the destination or otherwise it may fail.

(Refer Slide Time: 17:38)



So, what happened was the experiment that they performed, what came out was something like this that if the letter was to reach the target, it would reach in roughly 6 steps.

(Refer Slide Time: 17:52)



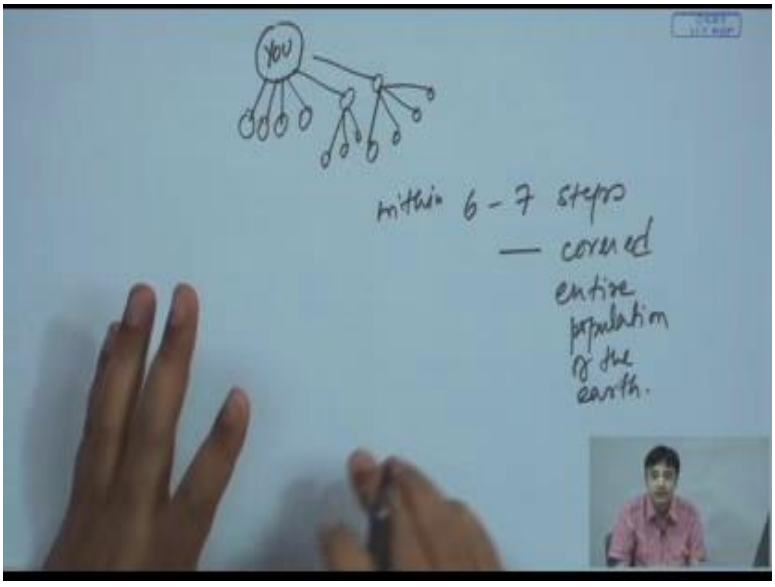
So, if the letter reached the destination, it did so in at most 6 steps. But, as you can understand that this is a, you know, this is a stochastic experiment and things happen by chance most of the times. So, what happened was like 64 out of 296 chains actually

reached the target. So, they initiate with 296 chains to start this experiment with, but then only 64 of them survived. Many of them dropped midway.

But, those that survived, among them, all of them reached the destination in an average of five point two steps. So, basically every letter, or every time the letter reached the destination, it reached within 6 hops, roughly within 6 hops. So, this was a very interesting observation that Travers and Milgram made. And, this is known as the 6 degrees of step separation. And, this is actually a very interesting trivia question in various quiz contests.

So, now the idea is like you might think like is this all a magic? By which this is happening? The point is intuitively if you try to explain this, there is a reasonable explanation that exists. It is nothing happening by magic. There is an intuitive; there can be an intuitive explanation for this. And, in the next slide we will try to discuss this intuitive explanation.

(Refer Slide Time: 19:40)



So, imagine that you are a person in the network. So, think of your Facebook friends. How many Facebook friends roughly you have? So, I would imagine somewhere between 500 to 1000. So, let us be much more moderate.

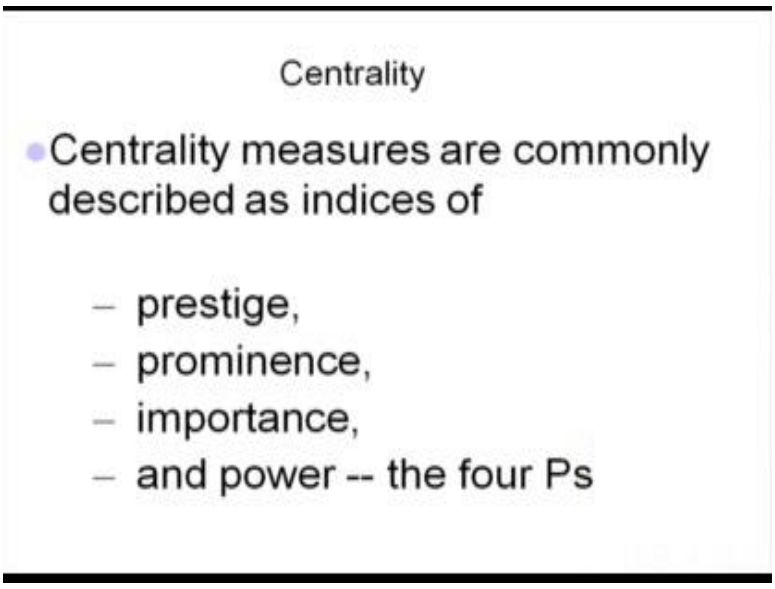
Let us take that you have say some 100 friends. And, I am assuming that these 100 friends are not connected among themselves. So you have, actually you have more than

that. But then, say there are at least 100, who are not at, who are no way connected among themselves. Now, these 100 friends by the same hypotheses will again have another 100 friends, like this. And, this will continue. Now, if you look at this tree of acquaintances or friendships, this completely desperate tree, where two nodes only know the parent, but nobody among themselves. So this, if you grow this tree, then what you see is that within 6 to 7 steps; so, if you have 100 friends in each level, within 6 to 7 steps you have, within 6 to 7 steps we have covered the entire population of the earth.

Of course, I understand that there could be transitivity triangles. But, what I am assuming is that there are 500 friends, and among these 500 or they are. So, in many of you will have 800, 1000 friends. But among this, there are at least 100 friends who are not connected among themselves. That is the assumption. And that is the assumption in every step.

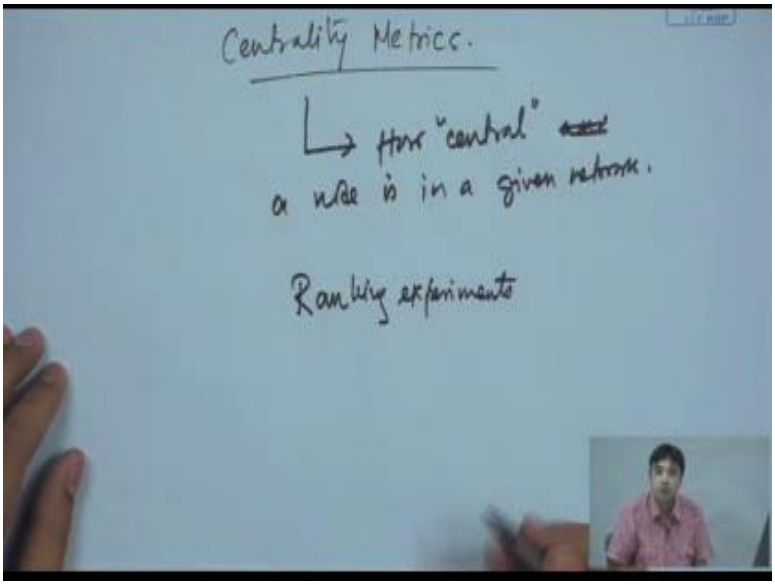
If you do this simple assumption, which I would say is a realistic assumption. And if you go by this, then at every step you spawn a bunch of new nodes. And, if you have spawned until say 6 steps, you have at least like a few billion nodes that you have covered. So, that is how. So, this is why Milgram and his colleagues could have each of the letters reach their destination within 6 steps. That is the basic idea of the 6 degrees of separation.

(Refer Slide Time: 22:12)



So then, after this we will start with another very important quantitative metric that people quite often use. So, these are called centrality metrics.

(Refer Slide Time: 22:31)

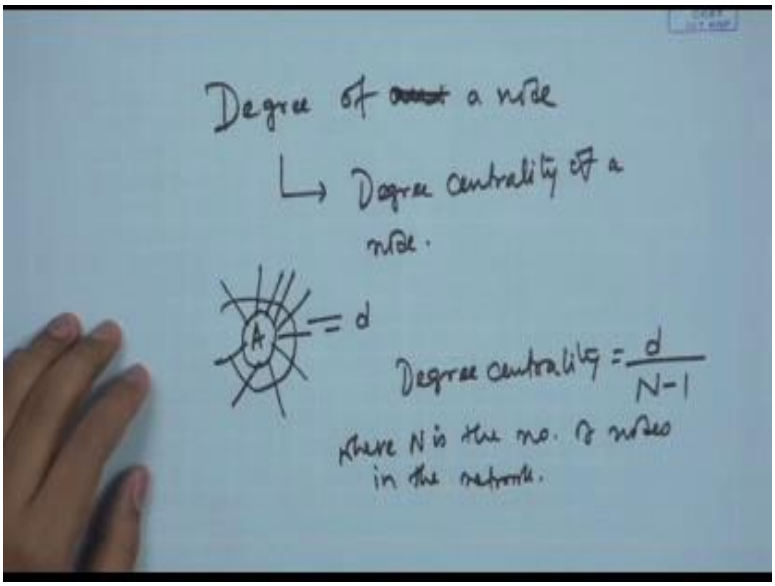


So, as the name suggests you can immediately understand that centrality would indicate how central a node is in a given network. So, you want to estimate how central a node is in a given network; and, as I write in the slides. So, you want to estimate basically centrality in terms of these four quantities. This 4 P's - prestige, prominence, importance and power; these are the 4 Ps of or the 4 pillars of measuring centrality. So, you want to identify prestigious nodes, prominent nodes, powerful nodes and important nodes. All of them more or less means similar; if you think carefully.

So, we have to have quantitative measures to identify such central nodes. And, this is actually important in various ranking experiments; because in various experiments, what you want is to rank the nodes according to their centrality values. So, the more central values go at the top, the more nodes with more central values go at the top and the nodes with less central values come at the bottom. And, this ranking is actually necessary for various other applications as we shall see in some of the later part of the course.

So, now say if we are like okay with the philosophical definition of centrality, we have to find out quantitative measures to identify centrality of the nodes in a network. Now, one of the simplest measures that come to once mind would be, in the context of a network, would be perhaps the degree of a node.

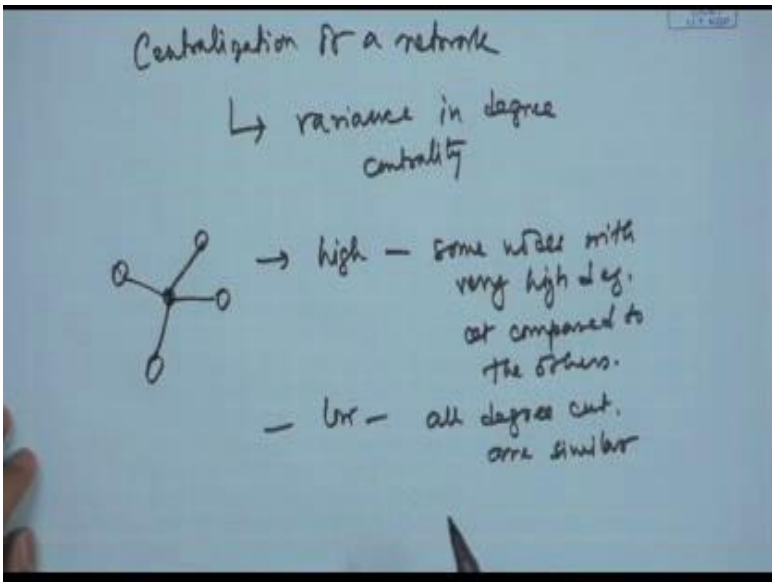
(Refer Slide Time: 24:38)



Degree of a node, so that is what I write in the next slide; and this is termed as the degree centrality of a node. So, and the definition is very simple. Suppose, there is a node say A and it has say a degree equal to d. So then the degree centrality, the degree centrality is equal to d by capital N minus 1, where N is the number of nodes in the network.

So, basically what you are doing? You are expressing the degree of a node in as a fraction of the maximum possible degree of a node. So, in all, a node can be connected to N minus 1 other nodes, if there are N nodes in the system. So, if the node has a degree d, then you are expressing this d as a fraction of N minus 1. That is what is the degree centrality of a node in a network.

(Refer Slide Time: 26:21)



Now given this definition of degree centrality, we can also define something called the centralization of a network. Centralization of a network; which means the variance in degree centrality; you try to measure. So, for each node you can estimate the degree centrality. Now, using these values you can measure the variance of these values. And, if this variance, so this variance is called the centralization of the network.

If this variance is high, this means that there are some nodes with very high degree centrality. If it is high, some nodes with very high degree centrality compared to the others. If it is low, then all degree centralities are similar; mostly similar to each other. And, an example of a case where the degree centrality is queued, that is, the centralization is high, would be.

If you can simply imagine a star network, where, the inner black node will have a high degree centrality; whereas the outer nodes will have low degree centrality. So, this has a high centralization, whereas a line network will have a low centralization because all of them will have almost equal degree centrality. So, that is why we end this part of the lecture.

Next day we will start with (Refer Time: 28: 12) centrality.