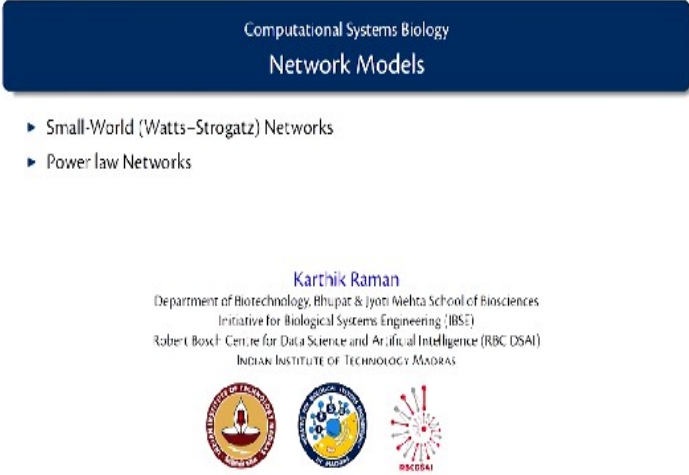


Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology - Madras

Lecture – 20
Network Models


(Refer Slide Time: 00:11)



Computational Systems Biology
Network Models

- ▶ Small-World (Watts–Strogatz) Networks
- ▶ Power-law Networks

Karthik Raman
Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences
Initiative for Biological Systems Engineering (IBSE)
Robert Bosch Centre for Data Science and Artificial Intelligence (RBC DSAI)
INDIAN INSTITUTE OF TECHNOLOGY MADRAS



In today's video, let us continue with network models and look at 2 interesting real-world network models, namely small-world networks or also known as Watts-Strogatz networks and power law networks which was first proposed by Barabasi and Albert in 1999.

(Refer Slide Time: 00:27)

Small-world networks
 Watts-Strogatz model, Watts DJ & Strogatz SM (1998), Nature 393:440-442

Regular Small-world Random

$p=0$ ————— Increasing randomness ————— $p=1$

For a large region between $p = 0$ and $p = 1$, we find

- ▶ high clustering, and
- ▶ low path lengths

In both ER random graph and the small-world model, connectivity distribution peaks at an average value and decays exponentially. Such networks are called "exponential networks" or "homogeneous networks", because each node has about the same number of links.

What do you find in the real-world? You find a lot of clusters, right. You always have this notion of a small-world, right. You say that you know it is a small-world because you met somebody from some other walk of life and so on. So what Watts and Strogatz suggested was, they had a way to build networks such that they had some interesting properties compared to random networks.

So, they started with the regular network as you see here. So, this is essentially what they call a regular lattice. It's very regular. You know what a regular graph is, right? So, every node has the same degree and so on. So here it is similar, you have a very set pattern of connections, right. So, think of this as a, you know, a round table where everybody knows 2 people to the right and 2 people to the left, right and then what they started doing was, they started rewiring these networks with some probability.

So, on the left hand side, this regular network is 0 probability, right and then you slowly start rewiring the network. When you have a small probability of rewiring, you will change a few edges in the graph. When you have a very high probability of rewiring, you basically change almost every edge in the graph. So, what you do is, you incrementally make the network more and more disordered in some sense.

But what you find is, between the extreme of $p=0$ and the other extreme of $p=1$, you find some

very interesting properties, what is called the small-world. So, what is interesting about small-world networks? So, they find that small-world networks exhibit very high clustering and much lower path lengths. And in fact, lower path lengths almost on the order of what you find in the random network.

But the clustering is much much higher than what you would see in a random network, right. So, quoting the paper, "in both ER random graph and the small-world model, the connectivity distribution peaks at an average value and decays exponentially. Such networks are called exponential networks or homogeneous networks because each node has roughly the same number of connections."

So how will you say, how will you make a comment like my network has much higher clustering, not random or my network has approximately the same path length as random networks. So, what you need to do is, given a particular small-world network, you build an equal sized random network and you compare the properties and obviously you don't build one random networks but you build like 100s or 1000s of random networks and then you compare the average properties.

(Refer Slide Time: 03:35)

The slide contains the following handwritten notes and diagrams:

- Random Network (ER):** A node i connects with probability $p = \frac{E}{N(N-1)}$.
- Small World Networks (WS):** A node i connects with probability $p = \frac{E}{N(N-1)}$.
- Binomial Distribution:** A graph showing $P(k)$ vs k for a Binomial distribution, with a peak at $k = np$.
- Small World Network (WS):** A graph showing $P(k)$ vs k for a Small World Network, with a peak at $k = C$.
- Small World Network (WS):** A graph showing $P(k)$ vs k for a Small World Network, with a peak at $k = C$.

So, this comes back to some sort of classic hypothesis testing. Let's look at small-world networks. So, this is usually called WS, Watts-Strogatz, 1998. How would I substantiate these?

So, we say that the characteristic path length of these networks is more or less in the range, in the ballpark of random networks but the clustering coefficient happens to be much much higher than for random networks.

So, to prove this, let's say I have a WS graph with N nodes and E edges. I create the corresponding ER graph with N nodes and E edges or rather I would make some 1000 realizations of this graph, ER graph, right and make a plot. I would plot let's say the average, the characteristic path length, right. So, this is basically some probability or like this, essentially a histogram and this is the average path length.

How will this graph look like? It will be normally distributed. We will get a nice bell-shaped curve. Now the question is where does, so this is your ER path length. The question is, where does your WS path length lie? So, this will have let's say this gives you a number L_{WS} , where does that lie? You will find that, that lies somewhere in this zone. So, nothing surprising, nothing interesting out there. It is not statistically very different from your random path lengths.

“Professor - student conversation” Plot of what? L_{WS} is not a plot. L_{WS} is a point, right. It's one network we have. L_{WS} is one network, corresponds to one network and I am plotting on the L_{ER} distribution. L_{ER} , I have 1000s of, 1000 networks here and this is the histogram corresponding to that, right.

Now I take these 1000 realizations and try to get the clustering coefficient as well. The average network clustering coefficient. **“Professor - student conversation”** Same probability, same probability, right, because you will get some variation, right.

So, you need to get some, so this is why we would, this is essentially a bootstrapping to basically try and understand what is the, you know, distribution of the test statistic that you want, right. From one network you will get one value but if you bootstrap, you will get a bunch of values. So, you know what is the distribution of your statistic of interest and then you see where your test statistic falls but that won't, they won't be comparable to your N, E network, right.

If you take different probabilities, it means that even if you keep the nodes same, the edges will be changing, right. See E is roughly pNC_2 , right. If p is the probability of a random edge in the ER graph, E is essentially pNC_2 , right. So now if I take these 1000 realizations and I plot the clustering coefficient. So, I will call this C_{ER} . I will again get a curve like this, right. This is just a histogram.

So, this is some probability or frequency count. Where does C_{WS} fall now? You will find that C_{WS} falls at the tail of this distribution. Why? It is not easy to answer. You have to see it for yourself but you basically find that if you take this kind of a regular network, start rewiring it in this fashion, you find a good amount of clustering, that will give you a clustering coefficient that is substantially higher than random graphs.

(Refer Slide Time: 08:29)

Small-world networks

- Small-world networks have a characteristic path length of the same order as random networks ($L \gtrsim \log N$), but have $C \gg C_{\text{rand}}$.

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.98	0.79	0.00027
Power grid	18.7	12.4	0.090	0.005
<i>C. elegans</i>	2.65	2.25	0.28	0.05

Characteristic path length L and clustering coefficient C for three real networks, compared to random graphs with the same number of vertices (n) and average number of edges per vertex (k). (Actors: $n = 225,226, k = 61$. Power grid: $n = 4,241, k = 2.67$. *C. elegans*: $n = 282, k = 14$.) The graphs are defined as follows: Two actors are joined by an edge if they have acted in a film together. We restrict attention to the giant connected component¹⁰ of this graph, which includes ~90% of all actors listed in the Internet Movie Database (available at <http://us.imdb.com>), as of April 1997. For the power grid, vertices represent generators, transformers, and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction. We treat all edges as undirected and unweighted, and all vertices as identical, recognizing that these are crude approximations. All three networks

So, let us look at an example quickly. So, these Watts and Strogatz in their paper, classic paper, they studied 3 different kinds of networks. The first one was a film actor network, right, where the nodes are film actors and the edges correspond to movies where they acted together, right and the next one is a power grid where nodes are power stations and edges basically corresponding to power lines and so on.

The third network they looked at was the *C. elegans* neuronal network. So, a network of neurons in *C. elegans* and what did they find? L_{actual} . This is the, this is the same as the L_{WS} that I was

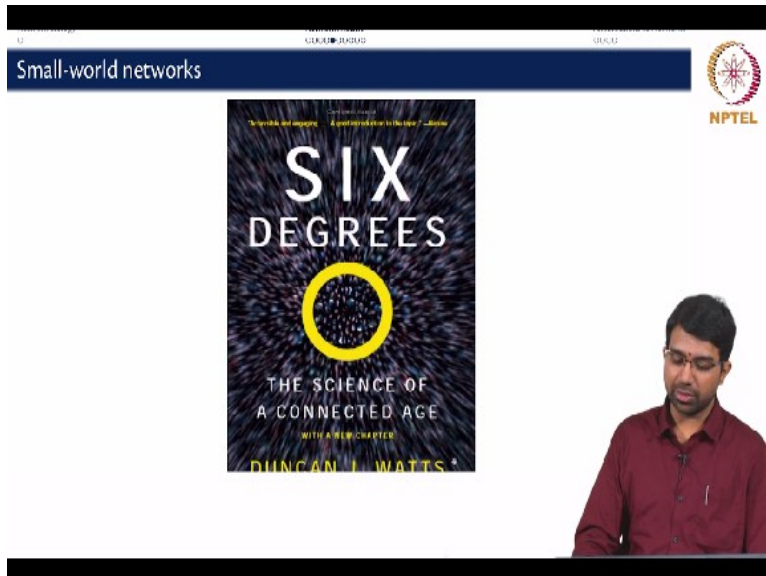
talking to you, right. This is L_{WS} and this is L_{ER} and you say they are all the same order of magnitude. There is obviously some variation but it is slightly higher approximately the same order of magnitude, that is as they say here but look at the clustering coefficient.

They find the random clustering coefficient is almost infinitesimally small but this is massive, okay. So, this 0.79 versus 0.00027. So, if you want to have a visual representation of that, it would be the equivalent of having a graph like this and having a real value somewhere here, right. So, this is a graph with mean=0.00027 and your real value of some 0.79 is somewhere here, right. So, it's completely outside your distribution which is, which means it's in the extreme, right.

Which means obviously that it is extremely statistically significant. So, if you want to compute a p value, it will basically be, mathematically it will be your less than epsilon, right. It will be essentially 0 because you find that there are no, no reading from your, from your observations that comes anywhere close to your real.

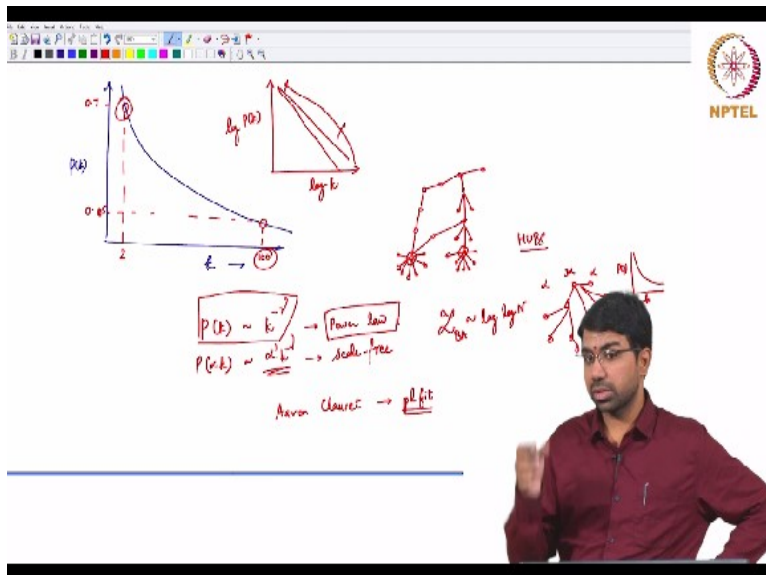
So, when you see these 3 networks are very different sizes, so the film actor network had 2,25,000 nodes. Power grid had about 5000 nodes and C elegans had about 300 nodes with different average degrees. So, all 3 networks show the small-world phenomenon where L is somewhat greater than but still equal to L_{random} but C is much much higher than random.

(Refer Slide Time: 11:19)



So, this was a very useful observation and Watts wrote a very interesting book on this topic known as Six Degrees. It is worth a read if you are interested. But then scientist found that most real networks look different. Many real networks, a lot of real networks, so you did see that the film actor network or the C. elegans neuronal network does follow the small-world kind of distribution but you find that many real networks have a completely different distribution, right. How do they look like?

(Refer Slide Time: 11:53)



So, these are all different kinds. So, some real networks fall into this class. No real network will fall into the random class practically, right. So that is like a null model that you want to always compare against and see how different you are and so on but real networks will fall into the

small-world class or what we call the scale-free or power-law class or have a mixture of the 2, right or it will have different behaviours in different zones.

So, you will never see a perfect degree distribution. We will look at it shortly. So, real networks have a very interesting degree distribution that looks like this. What does this mean? Let's look at 2 points in this degree distribution and let me say this is 0.7, let me say this is 0.1 and may be let me say this is about 2 and this is 100. So, what you have is most of your nodes are boring. They have a degree of like less than 1 or 2, right.

So, 70% of your nodes lie in this region, right but about 10% of your nodes or may be even less than that lie in this region where they have 100 links. So typically, you will get a network that looks like this. What you have here, most of your nodes have degree of 1 or 2 but you have these 2 nodes which have very high degree. So, these are your hubs. So, if you translate it this to a log log plot, you would essentially get a straight line. So, this is $\log p(k)$, this is $\log k$

or

$$p(k) \sim k^{-\gamma}$$

So, this is called power law. And if you take

$$p(\alpha k) \sim \alpha^{-\gamma} k^{-\gamma}$$

So, this is why it is called scale-free. If you scale k , right, the behaviour does not change. This is called, so I prefer to call them power-law networks rather than scale-free networks. These networks have a very interesting property. What is that property?

(Refer Slide Time: 15:28)


Power-law networks
 Barabási-Albert power-law model: Barabási AL & Albert R (1999) Science 286:509-512

NPTEL

- ▶ Characterised by a power law degree distribution

$$P(k) \sim k^{-\gamma}$$

- ▶ Such networks can be generated using the "preferential attachment", or "rich get richer" model
- ▶ Networks generated by this model have a power-law degree distribution characterised by $\gamma = 3$
- ▶ Scale-free networks with $2 < \gamma < 3$, a range commonly observed in many biological networks, are *ultra-small*, with a characteristic path length $l \sim \log \log N$, significantly smaller than that of random networks ($\log N$)



They have been created by growing a network. You can actually create these networks by using what's known as the rich get richer model or what is more scientifically called preferential attachment. So, you preferentially attach nodes to nodes that are already rich. Rich in terms of degree. So, you prefer to attach to a node that already has a higher number of edges. So, if you were to grow this network, let's say you start off in this fashion, you start with a network that looks like this, then you have a third node incoming, it will connect with equal probability to either of those. Then you have a fourth node, the probability of connecting to these nodes is 0.25, 0.5, 0.25. So, let's say it connects to this. Now the probabilities will become even more different. So, this is going to be 1, 3 and 1, right. So, your probabilities of connecting to these will become alpha, 3 alpha and alpha, right or essentially 0.2, 0.6 and 0.2. So, maybe the node connects here.

The next incoming nodes connect here, here, here, here and as you start growing this network, you will find that you end up with a power-law degree distribution of $p(k)$ versus k . So, basically you can generate these networks using a power-law, right. So, the preferential attachment model generates power-law networks.

“Professor - student conversation starts” Well it is, there is no why to that. If you follow this method of generating networks, you end up with a power-law network, that is the way to look at it, right. It turns out, so probably this is how real networks grow, right. So, if you look at a classic example in this regard is the internet router network or even the worldwide web, right.

So, you see that they just have that a lots more links are more likely to get more links than pages that have very few links. They remain with very few links, right. So, when a new site comes in, it will attach, it will connect with higher probability to already popular websites and so on, right. So, this kind of behaviour seems to be present in many real systems.

But in practice, you may not find this exact perfect power-law but you will find some, some deviation, right. This also looks linear almost but you will essentially see something like this or you know, different flavours of this, right but you can still say that in this zone, I have like a proper power-law but then there is a rapid fall of our whatever or it remains like, a little flatter, different kinds of behaviours can be seen in real networks.

So, you want to see, unfortunately there came about a big obsession for power-law networks. This classic paper on power-laws was published in 1999. After that everybody started publishing a paper saying I have a power-law network, right but you have to actually do a statistical test. Can you really make a proper fit of this sort, right? With what accuracy can you make a fit of the form p_k is some α into k to the $-\gamma$, right and what are your values of γ , right?

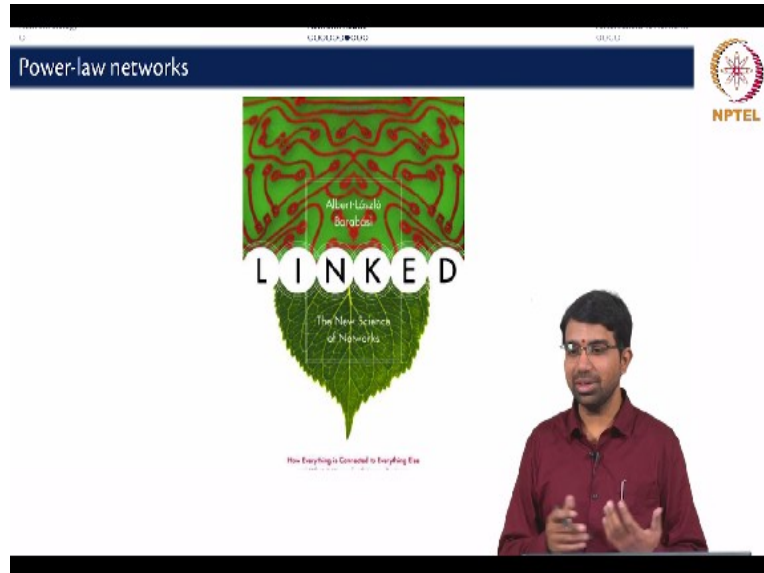
So, there is a, there is a useful tool that was released by Aaron Clauset for MATLAB. I think the tool is something like pl fit or something like that which essentially tells you if you have a power-law network. It will give you a p value for the confidence in your statistics and so on. So, you find the networks generated by this model have a power-law degree distribution with γ equals 3 and you find that scale-free networks with $2 < \gamma < 3$ which is commonly observed in biological domain are ultra-small with a characteristic path length that is very, very small, right.

So, it is $\log \log N$ which is significantly smaller than that of random networks, okay. So, you can imagine, right, if you have a network like this, it is kind of very easy to most all these nodes connect to a hub and from this hub maybe you also have connections like this, right. So, from this hub, it is very easy to go to this hub or any other node.

So, your $L_{\text{Barabasi-Albert}} \sim \log \log N$ which is a very small value. Well, it depends, right. Hubs are

very important nodes. So, you want to target hubs. So, they have higher centrality measures obviously they have very high degree and so on. We will look at some of these properties may be in the afternoon.

(Refer Slide Time: 20:49)



So, Barabasi has a very good book which is called LINKED. So, the other book says that every network in real life is, you know, can be seen as a small-world network and Barabasi would argue that most real-world networks are power-law networks and reality lies somewhere in between, okay. So, welcome back.

So, we were looking at network models and the third network model that we were looking at was the power-law network model and this was a very nice book that was written on the power-law networks and I think you should try to take a look at both these books, both, SIX DEGREES by Duncan Watts and LINKED by Barabasi.

(Refer Slide Time: 21:30)

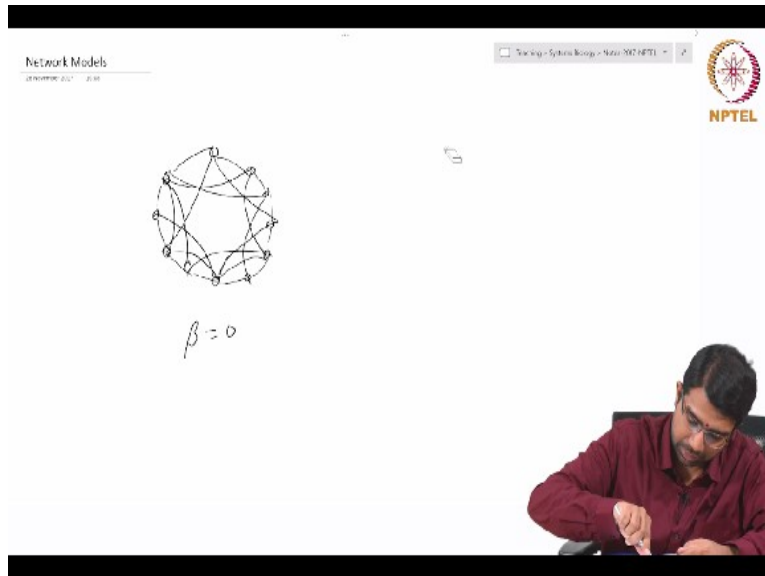
Small-world	Power-law
egalitarian	hubs and resources
no growth	preferential attachment
local clusters and distant links	power law
under attack: deteriorates	under random attack: stable
no particular targets	under targeted attack: weak
Small average node-to-node distance	



So, to compare small-world and power-law. So, small-world one would call it as egalitarian, right. So, it is uniform distribution of resources in some sense whereas power-law has hubs and resources, right. So, there are certain hub nodes which have a very high degree of connectivity and there are many nodes which have very low degree of connectivity and small-world networks don't have the concept of network growth.

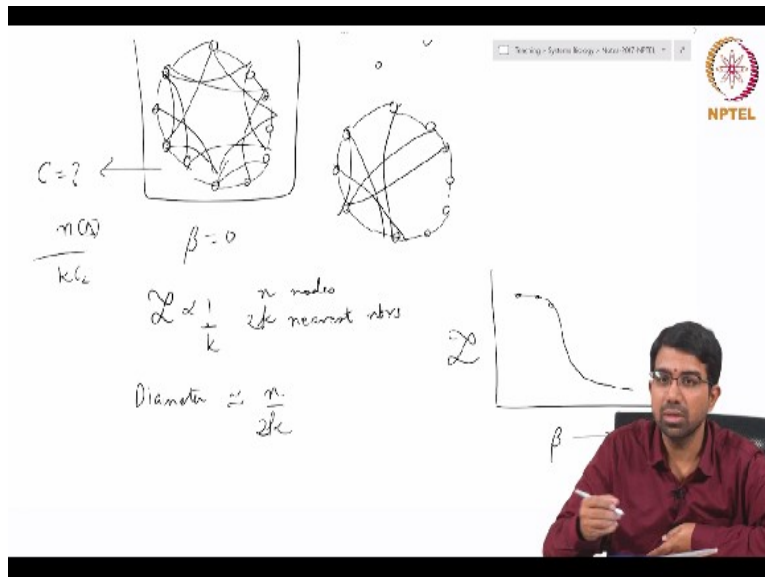
Whereas in power-law, there is growth by preferential attachment and you have local clusters and distant links in the small-world and you have distant links as we saw in the regular lattice. We will take a closure look at the regular lattice again. And what happens under attack? That is something that we need to look at. So, we will look at that and in all, in both networks, you have small average node-to-node distance.

(Refer Slide Time: 22:32)



So, before we look at biological networks, let's just refresh some concepts. So, how does the small-world network look like? Something like this, right. This is a regular lattice where every node is connected to and so on. I may have missed a few edges but let's look at a regular lattice like this. So, this is something we say no rewiring, right.

(Refer Slide Time: 23:46)



And then you have an intermediate lattice. I don't know why the thickness is changing. You can start rewiring it. You have some of these nodes but you also have some nodes that are essentially across the table. You have some nodes that really grow. So, what happens is when you introduce the first of these links, you see a massive drop in characteristic path length. What is the characteristic path length here?

Let's say you have n nodes connected to $2k$ nearest neighbours. What is the characteristic path length? It will be roughly of the order of n/k , $L \propto n/k$ right or what is the diameter first? The farthest nodes are diametrically opposite and you need to do, you can hop k nodes in one shot on one direction and how many k 's do we need to hop to get to the diametrically opposite, right?

So, it will actually be $n/2k$, right. What about characteristic path length? It will also be some inversely proportional to k for sure, right. So, but what happens is, when you start rewiring, you have a precipitous fall in characteristic path length. These are somethings that you may want to study. You can do simulations to understand some of these things, right and better still, let us ask this questions for this lattice that is here.

“Professor - student conversation” Well, you need a few links, right. So, it really depends upon your β step size, right but if you have like even the first few links, it actually even doesn't worry too much about the value of β , the number of links even. So, you do the first few rewirings, there is a precipitous fall in the characteristic path length or even diameter.

So, what is the clustering coefficient? Can you guess? Let's solve this. What is clustering coefficient again? If a node has k neighbours, kC_2 is the denominator, numerator is number of triangles that intersect at that point or the number of edges between the k neighbours.

(Refer Slide Time: 27:09)

Recap

Topics covered

- ▶ Small-World (Watts-Strogatz) Networks
- ▶ Power law Networks

In the next video ...

- ▶ Centrality-lethality hypothesis
- ▶ Assortativity

I hope you had a good introduction to small-world networks today. So, the small-world networks was first published in 1998 and very recently we had a 20th anniversary celebration of that work with a very interesting paper on the application of small-world networks and so on. And we also looked at power-law networks today. And in the next video, we will start approaching biology and look at something known as the centrality-lethality hypothesis and also the concept of assortativity.