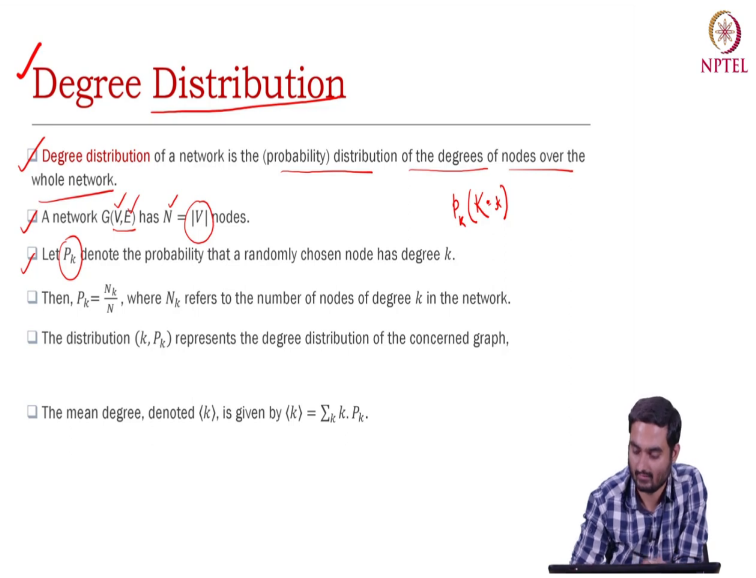
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 02
Lecture - 07
Lecture - 02

Hi everyone. Welcome back again. So, in the last lecture we have started discussing about network measures and we have you know discuss cases you know where we have seen you know different new cases where you know network measures plays an important role. Particularly, I think we stopped at discussing the degree of a node, right.

(Refer Slide Time: 00:53)



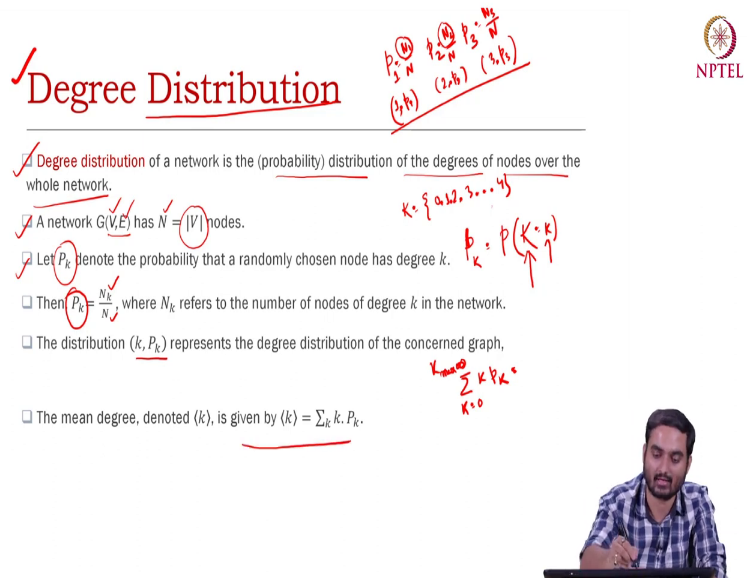
So, the degree of a node is a very important property. And today we will discuss how we can use it for you know further defining you know other kind of metrics. So, one such metric is, one such property is degree distribution, right. So, what is degree distribution? So, as the name suggests, you basically, here we basically you know draw the distribution of degrees of nodes.

And when it comes to distribution you know I hope you are aware of probability distribution. There are you know two kinds of variables, one is discrete random variable other is continuous random variable, right. So, in discrete random variables we basically relates to quantities like probability mass function. In continuous random variable we can relate it to you know probability density function, right. So, if you do not have you know ideas about these measures, right. I strongly suggest you guys to look at fundamentals of probabilities, right. So, we will get ideas about this, ok.

So, degree is essentially discrete variable discrete quantity because we cannot say that you know the degree is 2.553, right. A degree can be 1 or 2 or 3 or 4 and so on, right. Degree is always an integer, right positive integer, in fact, ok. So, degree distribution of a network is the probability distribution of the degrees of nodes over the whole network, ok. First, I will quantify what is degree distribution, then I will give an example, and then I will tell you why this is, so you know this is such an important metric, ok.

So, let us look at let us take a graph G(V,E), V is a set of nodes and E is a set of edges where the number of nodes is N, ok capital N. And let us denote P(k), ok P_k, P(k) as the probability that a randomly chosen node has degree k, ok. So, this P(k) is important. So, basically it is saying that, so p(k) is you know that degree is small k, right.

(Refer Slide Time: 03:34)



So, p(k) I mean you can write it, in in other way you can basically say that you know P_k=P(K=k). So, this is the random variable capital K and small k is one value of the random variable, right. In our case this random variable can take say 0, 1, 2, 3 and so on and so forth. So, degree can be 0, degree can be 1, 2, 3, so on. Degree can also be 0 by the way, remember this.

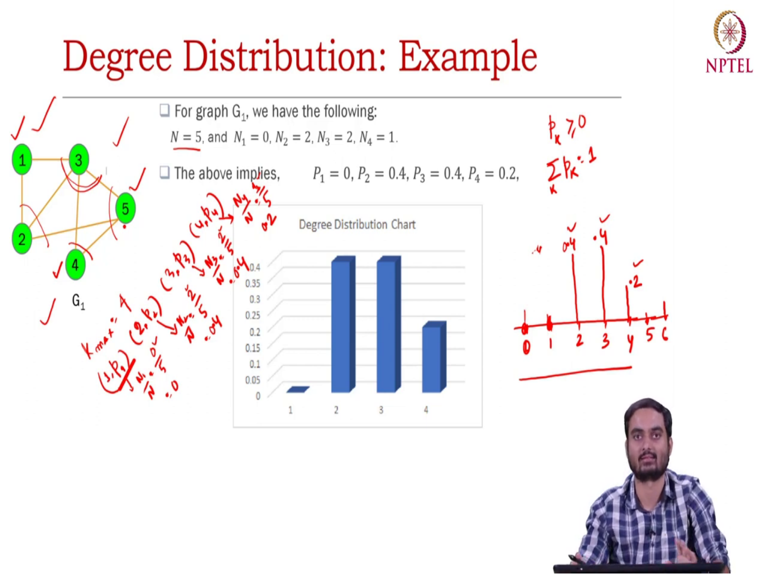
For example, if a node is disconnected, right. If the node does not have any edges, right incident on a on that particular node. So, degree can also be 0, ok. So, how do we define P(k)? So, P(k) or P_k is basically a fraction of N_k and N. What is N_k? N_k is a number of nodes of degree k, right. And what is N? N is the total number of nodes.

So, you are basically asking that how many number of, how many number of nodes are there whose degree is k, right and you basically divide it by the total number of nodes, right. That will give you P(k), ok. So, you can actually define P(1), P(2), P(3), and so on. So, P(1) would be N_1/N, P(2) would be N_2/N, right, P(3) would be N_3/N. What is N_1? N_1 is the number of nodes with degree 1. N_2, number of nodes with degree 2 and so on and so forth, right.

So, this (k,P(k)), right in our case (1,P(1)), right, (2,P(2)), and these are standard notations by the way, right. If you are not familiar with these notations please go back and read the basics of probability (3,P(3)), right. So, this represents the degree distribution of a particular; of a particular graph, ok.

So, of course, if you know these are different quantities, what is the mean degree? So, the mean degree would be summation of (k,P(k)), right, where k ranges for whatever 0 to infinity, right. Yeah of course, infinite number of degree that is not possible, but you basically go up till you know what about the maximum degree and that is possible. And you can say that 0 to, right, it can be k_{max} for example, ok. And that would give you the average degree or the mean degree of the graph, ok, mean degree of node in a graph, ok.

(Refer Slide Time: 06:23)



Now, why this is so important? Let us look at, first let us look at an example and then we discuss. So, let us take this G1, ok. This is a graph. And how many nodes are there? We have 5 nodes. So, N equals to 5, right. And let us look at and what are the degrees present in this graph. So, we see that you know; what is the max degree? What is k_{max}? Ok k_{max} is look at this one, this is 3 right, 4 right, 3 and 2, right. So, max degree is 4, ok.

So, we will actually take (1, P(1)), right. We will measure (2, P(2)); (3,P(3)) and (4,P(4)), ok. So, in order to measure P(1), what I need? I need N_1/N. What is N_1? Number of nodes with degree 1, right. How many nodes are there in this graph with degree with degree 1? 0. There is no node with degree 1, right. So, this is 0 by 5.

What is P(2)? P(2) is N_2/N. How many nodes are there with degree 2? This one and this one. So, there are two nodes, so this would be 2/5, right. What is P(3)? N_3/N. So, how many nodes are there with degree 3? We have one here and one here, so this would be 2/5. What is P(4)? P(4) would be N_4/N. How many nodes are there with degree 4? There is only one node with degree 4, right, so 1/5.

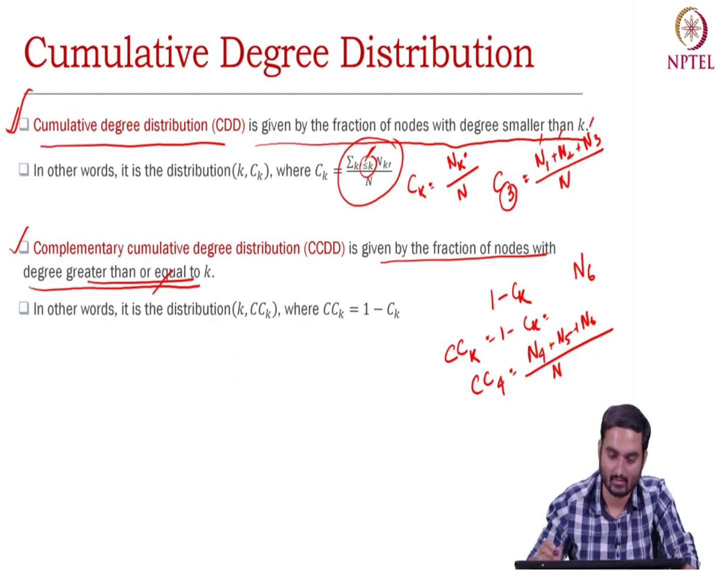
Now, you can also cross check it whether you have missed some nodes, missed out some nodes, so you can you can actually sum the numerators of all these fractions. So, 0 plus 2 plus 2 plus 1, so this would be 5. So, we check whether 5 nodes are there in this particular graph or not. There are 5 nodes, so it is ok.

So, once we get this quantity, so then P(1) is 0, P(2) is 2/5, right 0.4, right P(3) is also 0.4 and P(4) is 0.2, ok. So, then we can easily plot this. So, as you see here this is the distribution, so discrete distribution. So, you have 0, 1, 2, 3 and 4. So, there is nothing corresponding to 0. There is 0.4, right, so let us say this is 0.4, so this is bar graph, right. So, this is 0.4, right. This is also 0.4, right. So, this is this is 0, right because P(1) is 0, right. This is 0, this is 0, this is 0.4, this is 0.4, and this is 0.2, ok.

So, it is also satisfies the probability mass function you know constraints which basically says that you know all these P(i)'s, right P(k)'s in our case should be greater than equals to 0 and sum of P(k)'s right should be 1. You see here 0.4 plus 0.4 plus 0.2 this is 1, ok. So, therefore, this is a distribution, right, probability mass function, ok.

So, now, this is discrete. Right now it may happen that you know there is no node with degree 5, but there is again one node with degree 6, right. So, corresponding to 5, x axis at 5 the value would be 0, y would be 0. Again corresponding to say 6 you have some values, right. So, you see that at certain points you do not have any value. So, this is kind of a discrete in that sense, ok.

(Refer Slide Time: 10:40)



So, what we generally do discrete distribution, right, if you look at pictorially also this is difficult to compare. For example, you have two discrete distributions you can compare, but you know visually it would be difficult to compare. So, what we do? We basically convert it to a continuous you know function where instead of looking at the you know the probability mass function or probabilities are at every k, we basically measure something called cumulative degree distribution.

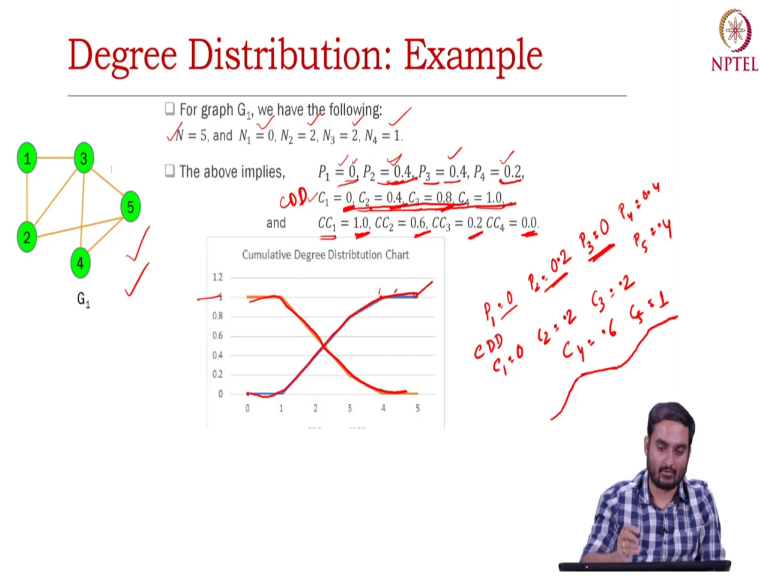
So, again this is a very standard term in the probability. This is called CDD, right cumulative degree distribution. What is this? So, cumulative degree distribution is given by the fraction of nodes with degree smaller than k, ok. So, Ck is Nk' by k, right as you see here, right where k is less than equals to k, where k' is less than equals to k.

What does it mean? Say, say C3, right, so C3 would be N1 plus N2 plus N3 divided by N, right. So, these are the nodes whose degree is less than equals to 3, ok. This is cumulative degree distribution. So, why this is useful? We will take an example we will see. Similarly, the complementary of that CDD is something called CCDD, right, complementary cumulative degree distribution.

So, this is essentially 1 minus Ck. Meaning that is the fraction of nodes with degree greater than or equals to k. Now, if we if we take equals to here then it would be greater than only greater than k, if we do not take, if we do not take equals to here then it would be greater than equals to k. So, you can adjust it, right. I hope you understand.

So, this complementary cumulative degree distribution CCk is 1 minus Ck which is basically which is basically you know say let us say CC4, this would be and say there are max N6, right, there are nodes max degree of a node is 6. So, it would be if we take greater than equals to, then it would be N4 plus N5 plus N6 by N, ok. So, you see that this is just a complementary of this one, right.

(Refer Slide Time: 13:27)



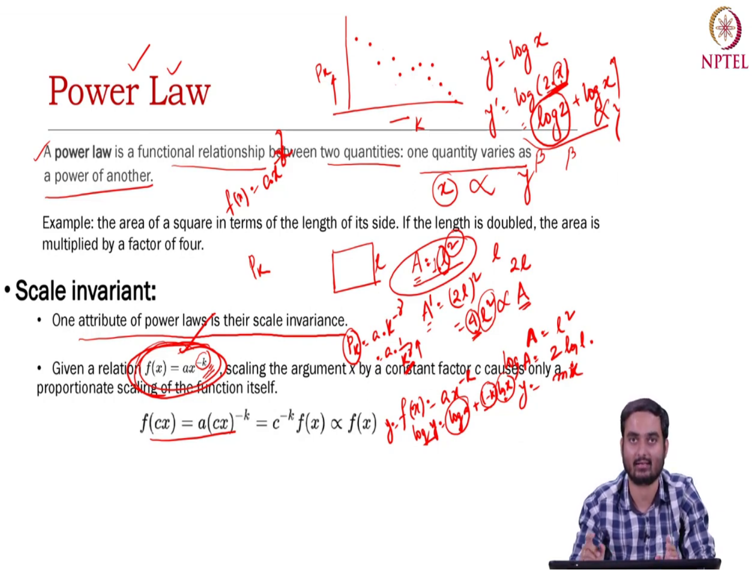
Now, why this is important? Now, let us take an example, ok. Let us take this graph, right. We have already seen this one. And N is 5, N1 = 0, N2 = 2, N3 = 2, N4 = 1; therefore, you have seen P(1) is 0, right P(2) is 0.4, P(3) is 0.4, and P(4) is 0.2, right. So, cumulative degree distribution, right CDD. So, C1 would be 0, right C2 would be this this plus this which is 0.4, C3 would be; C3 would be this plus this I mean P(3) plus P(2) plus P(1) which is 0.8 and C4 would be 0.2 I mean P(4) plus P(3) plus P(2) plus P(1) of course, this would be 1, ok.

Similarly, you know cumulative, I mean the complimentary 1 would be 1 minus C1 which is 1, 1 minus 0, 1 minus 0.4, 0.6, 1 minus 0.8, 0.2, 1 minus 1, 0, ok. So, if you plot this one you see that in case of CDD it is growing, right from 0 to 1, right. So, for this graph the CDD would look like, this is 1, ok. Of course, then if this curve looks like this, then then the complementary one would be the mirror image of mirror image of this 1, right. So, it starts from 1 and it would stop at 0. So, this would look like this. Exactly the mirror image of the other one.

And this is continuous, why this is continuous? Because let us say you do not have, say you have some values P1 some value, right say let us say P1 is 0, P2 is you know 0.2, P3 is 0, right P4 is 0.4 and P5 is 0.4, ok something like this. So, you see that if you plot it as a simple mass function you will see that there is 0, right.

Then, if you take cumulative degree distribution, then C1 would be 0, C2 would be 0.2, this plus this, right C3 would be again 0.2, 0 plus 0.2 plus 0, C4 would be, right 0.6 and C5 would be 1. So, there will be no you know no break, right in between. So, you will see that it would look like this 0, then 0.2, right 0.2, 0.2, 0.6 and 1, something like this, something like this, ok. We will see a plateau here, I mean a horizontal line here, similarly CCDD, ok.

(Refer Slide Time: 16:29)



So, now, this function, right the degree distribution; if you plot degree distribution, right say you have graph, you have a huge graph and you plot the degree distribution. So, this is k and this is Pk, right and you see some points, right like this. At every k you have a Pk. I mean of course, this you would not, it would not look like this, right. It would essentially look like this, right like this. So, you will have point like this, right something like this, ok. For every k you have a point.

Now, if you take real networks, biological network social network does not matter, you will see that it follows a particular property which is called power law. What is power law? Power law is a particular function for example, we have logarithmic function, we have heard about logarithmic function, we have heard about linear function for example. Power law is also another type of function, right.

A power law is a functional relationship between two quantities, we have two quantities x and y, right and power law indicates a particular relation between x and y, right. The relation is one quantity varies as a power of another, right. Say one quantity is x another is y. So, x will vary as the power of y. So, the power meaning say x to the power beta, where beta is a constant, ok. This is called power law.

Let us take a simple example of a power law. Let us take you know area of a square. You have a square, right and the length of every side is let us say l, right. So, what is the area? Area of a square would be l squared, right. You see that A varies with the power of l, ok. Now, this is a power law. Now, this power law has a nice property. Remember, I have not connected the relationship between power law and degree distribution. Let us only you know focus on what is power law function. This is a power law function, ok.

Now, this has a nice property, ok. The property is say you know double the size, double the length of a side. So, earlier it was l, now this is 2l. What would the impact on the area of area of the square? Right. So, it would be; so, then the new area A'would be (2 * l) squared which would be 4 * l squared, right. So, 4 is a constant.

So, you see that this is proportional to A, A' is a proof is proportional to A, why? This is just a scale you know just it basically scales up the value, earlier it was 1 times l square, now this is 4 times l square. But the functional form, right that one is a power of another that remains same, ok. This property is called scale invariant property. One attribute of a power law is that it is scale invariant, ok.

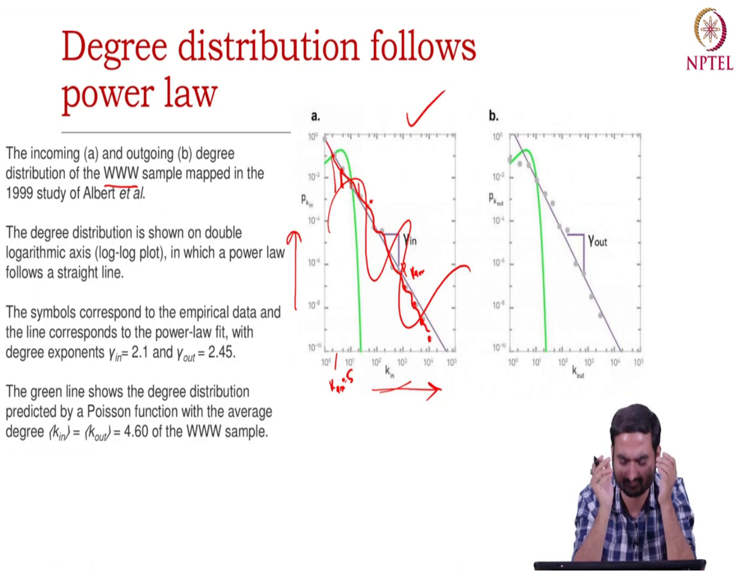
Let us take a function which is not a power law. Let us take y equals to log of x, ok. Now, what you do? You double the value of x, now this would be log of 2 time x. So, this would be log 2 plus log x, right. If you think of it from you know, you see that this is not scale invariant because when you scale this thing up, right the value of x, right. What happens is that you get a new quantity, but this is addition. This is not, it is not scaling up or scaling down log x. Now, this is not proportional to y, ok. So, that is the beauty of power law.

Now, why this is important in this context? It turned out, so people analyzed lots of such social networks, right biological networks also, and it turned out that the degree distribution this P(k), right P_k it actually follows power law, ok. Something like this, f(x) is a to the power a times x to the power minus k or minus gamma is basically say minus gamma, right.

You see that this is also power law, but here the power is negative that is, ok, right. In this this case, the power was positive now this is power this is negative, right. What it is saying that Pk would be a times k to the power minus some constant gamma, ok. So, it can also be written as a times 1 by k to the power gamma.

What does it mean? It means that as k increases, right; so, Pk is now inversely proportional to k, as k increases Pk value will decrease. So, as the degree increases degree of a node increases the fraction of nodes with that degree will decrease, ok. So, it and I mean why suddenly I got this this formula? Now, I suddenly got, I mean this is not something that that came magically, right.

(Refer Slide Time: 22:54)



You know let us say you know there are many such graphs which have been analyzed manually, right meaning that you know people essentially you know computed the degree, then they plotted the degree distribution and then fit the appropriate function which describes the distribution, right.

For example, this is an example. So, this is worldwide web network, WWW network, ok and this is directed graph. So, therefore, we have both in degree and out degree, in degree distribution and out degree distribution. Let us say in degree distribution, ok. So, this is k in degree of node and this is Pk in the fraction of nodes with certain in degree. And you see this dots right, so this dots are basically is Pk in values at every Pk, at every k in. Let us say k in is 5 and this is and let us say this is k in 5 equals to 5 and we have some values Pk in. So, these are dots.

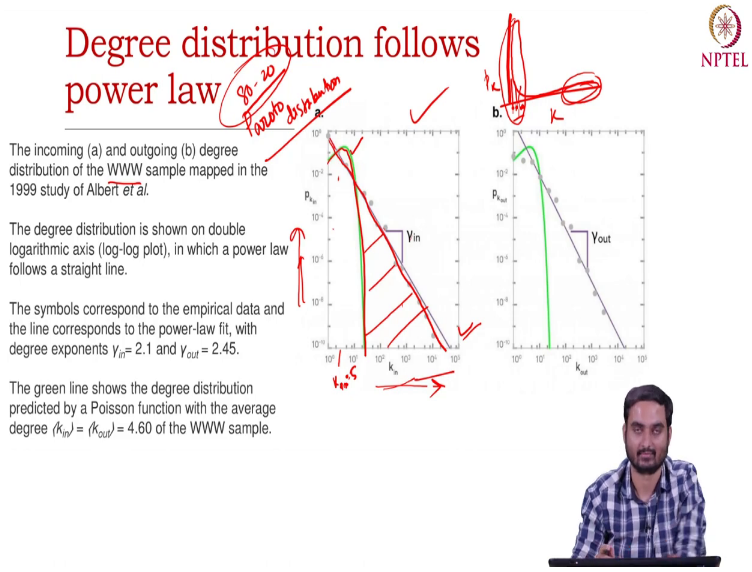
Now, which particular function, right which particular function best fit best fits all these dots? Right. So, people have you know tried with multiple functions which can fit this one, you can take log, you can take exponent, right does not matter, and, but how do you know that which function fits these data points properly, right.

You can take the error, the error between let us say; let us say you have a function like this, ok. And then you basically see the distance from this point to this line. So, this is the distance, this is the distance, distance, right. You sum all the distance of these distances are basically errors, right. So, the best fitting curve would look like this, would look like this, right which would basically be kind of you know over fitted curve.

But if you think of standard probability distribution, right say you take you know binomial distribution or positive distribution and so on and so forth, then you see that the best fitting curve would be a power low curve, ok. And the beauty about again this power law is that if you take log in both the sides, let us say, let us say the square again the area of a square, right say A equals to l square, if you take log, log of A would be 2 log of l, right. What is this? This is basically y equals to mx, right y equals to mx. y equals to mx is what? This is basically an equation of a line, right.

Let us see this one, right f of f of, right f of x let us see this one, right f of x is a into x to the power minus k, right. If we take log say this is y, y equals to this one, if you take log log of y is log of a plus, right minus of k log of x. What is this? This is y equals to mx plus C, where this is C, this is m, and this is x. So, in the log-log scale, this power law looks like a curve, look looks like a straight line, ok.

(Refer Slide Time: 26:40)



So, this plot you see that both x axis and y axis are in log-log scale and if you fit a line, right, if you fit a line you see that this line best fits all the data points. What is this green line indicating? This green line indicates the degree distribution the best fit of a Poisson function. If you take a Poisson function then it looks like this.

Now, you see that this is a worst fit, right. Of course, this is not the worst fit, but this is not as good as the power law, right because you see a lot of distance, lot of noise from the points. So, many such networks similar to this world wide web network have been studied, and it turned out that the best fitting curve which best describes the power, the degree distribution is basically a power law distribution, ok.

And the beauty of this power law, I mean I can take a whole lecture on what is power law, right say and what are the implications of power law. For example, you know this power law essentially says that I mean if you do not plot it in a log-log scale, if you plot it in normal scale, it would actually look like this, a skewed distribution, right k P k, right.

It what does it mean it means that you have a lot of mass here and you have very small, very few negligible mass here. What does it mean? It means that most of the nodes, say if you think of these points, right this case. So, k could be 0, 1, 2, 3, but if you look at here, k could be say 100, 200, right something like this. Basically says that most of the nodes in a graph have less degree, lesser degree, and there are very few nodes, right, there are very few nose nodes which have you know higher degree, right.

You can also relate it to a social network. There are there are very few users who have a lot of followers, there are very few celebrities on social network, but there are a lot of users who are actually normal users like us who do not have much followers for example, who do not have many followers, who do not, who do who are not considered as celebrities and so on and so forth. So, this is power law.

And if a network follows the power law degree distribution you can blindly say that this network has a lot of nodes with you know low degree, a small degree and you have very few nodes with a lot of degrees, right. This power law is also called as Pareto distribution. So, if you look at you know advance probability theory you may heard about, you may have heard about this Pareto distribution. It all it also says the same thing, right. In fact, the Pareto distribution or power law you know this kind of function appears everywhere, right.

It basically, this this concept was borrowed from social science, right where you know it was said that you know 80 percent wealth of a society you know are with 20 percent of populations, right and rest 20 percentage of wealth is with 80 percent of population. This is also called 80-20 law. This 80-20 law is very famous in social science, right. It is also said that, in another context that 80 percent profits are caused by 20 percent of employees in an industry, right and rest 20 percent profits are caused by 80 percent as 80-20 law is very famous.

I will talk about power law, you know the interpretation of power law in the next chapter. We will talk about you know scale free network, right and other types of network models. We will discuss more about this in the next chapter, right. So, this is all about degree distribution. In the next part of this chapter, we will discuss other properties of a network. We will look at you know centralities, we will look at clustering coefficients, and so on and so forth. So, till then stay with me.

Thanks.