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

Chapter - 03
Lecture - 01

Welcome to the course on Social Network Analysis again and this is the 3rd chapter on Network Growth Models. So, in the previous chapter we have seen how to quantify a network, we have seen different matrix that we can use to quantify a node or an edge or a graph, right. We have seen 3 levels of quantification microscopic, macroscopic and mesoscopic level of quantifying nodes, edges and networks.

And we have also seen properties like degree distribution, right scale free properties and so on and so forth. So, when we analyze networks, right. So, the first important thing is to scrape data from some social network, right. Now what happens is that when we try scraping data there are multiple problem.

So, first problem is say you know there are limitations in the crawling process, right. Crawling sometimes takes a lot of time crawling also had has its limitations for example, you may not be able to crawl those nodes or those edges which are private for example, right protected in some ways. At the same time the current network social networks like Facebook, Twitter it is so huge, right.

And you may not be able to scrape the entire social network as a whole, right. You can you may scrape a partial network for example, 1 percent or whatever 0.5 percent of the network, right. So, and oftentimes since we do not know the we do not know the exact connections, right.

So, sometimes we end up having connections which are partial which may not be complete. Therefore, most of the analysis that we do that we do on the partial network may not be significant enough, right. So, what people generally do in that case in social network analysis? People generally you know create a graph synthetically, right it is kind of a synthetic network lab I mean the network that you generate in your lab.

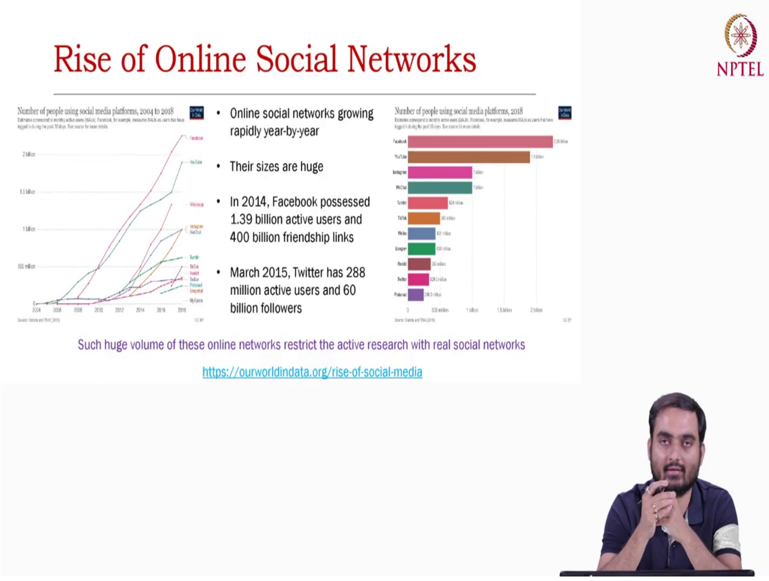
And then you know test all your models on the synthetic network, right. But we also need to remember that the synthetic network or the toy network that we develop or we did we that we generate should have properties that the real world networks exhibit, right. It should not be the case that your synthetic network looks completely different from the real world network.

Otherwise, you know there is no point in testing your method on a network. So, what is the idea? The idea is that say you have a synthetic network you test all your models it can be say community detection model, it can be link prediction model or whatever anomaly detection problem, right model. And you run those models on the synthetic network and see how these models perform, right.

You can also compare across models on different synthetic networks. And, then since you have control on the generation of the synthetic network, what you can do? You can say increase the size of the network, you can increase the number of nodes, number of edges, decrease and so on and then we test the scalability of your models on different in a different networks or different sizes, right.

So, you can do whole bunch of stuffs because, now you have control on the network generation process. So, you can create large network small network whatever, right partial network and so on.

(Refer Slide Time: 04:02)



Rise of Online Social Networks

- Online social networks growing rapidly year-by-year
- Their sizes are huge
- In 2014, Facebook possessed 1.39 billion active users and 400 billion friendship links
- March 2015, Twitter has 288 million active users and 60 billion followers

Such huge volume of these online networks restrict the active research with real social networks

https://ourworldindata.org/rise-of-social-media

And then when we do that and when we compare models on this synthetic networks. Then we choose the best model for example, and then and then you say then you say that look this is the best model for link prediction and you use this model for production, right.

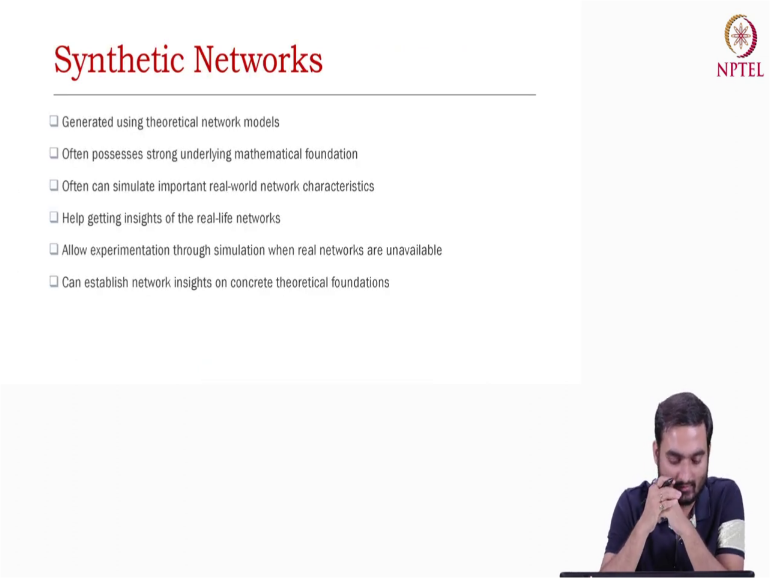
So, you give this model to somebody who is doing link prediction task and he can readily use this model for the production, ok. So, we have a lot of flexibility in terms of in terms of the size, right or different variations of a network which we may not have when you scrape data set for example.

When you scrape a data set you know how do you know that I mean how do you sync the network, when you sync the network when we reduce the size of the network you need to remove some nodes, how do we remove nodes? You cannot remove nodes randomly.

Because when you start removing nodes randomly the network may look I mean the network may you know start exhibiting different properties that the original network may not have, right. So, there should be a systematic way to control the network and synthetic network is one way of systematically you know controlling the generation process.

So, there are basically 2 reasons the first reason is that we may not be able to deal with such a gigantic real world network. Therefore, we can generate synthetic networks and test our models on the synthetic network this is number 1. Number 2 is that since we have control on the entire network generation process, we can change the you know process accordingly. And then we can create small network, big network or you know semi large network and so on, ok.

(Refer Slide Time: 05:55)



Synthetic Networks

❑ Generated using theoretical network models
❑ Often possesses strong underlying mathematical foundation
❑ Often can simulate important real-world network characteristics
❑ Help getting insights of the real-life networks
❑ Allow experimentation through simulation when real networks are unavailable
❑ Can establish network insights on concrete theoretical foundations

So, and you know all these models that we generally develop, right. They in the initial stage they may not be that scalable, right. So, when we test this non scalable models on the massive networks a massive real world network, right. It would take huge amount of time for example, right.

So, if it if these models take huge amount of time, then how do you compare across different models? Right. That would take ages, ok. Therefore, you know we use synthetic networks and then test the efficacy of these models on synthetic networks and then we finalize a particular model and then send it for production, ok.

So, synthetic networks generated using theoretical network models this is very important. So, we will discuss in this particular chapter a series of network models, network generation models that we use for synthetic network generation, right. Again, I am repeating it is very important to remember that the synthetic networks should exhibit real world properties, it should not look like a random network, right.

Because otherwise there is no point in generating a you know real world like synthetic network, ok. So, this underlying mathematical foundations is very important. So, when we draw a synthetic network, we need to understand how real world networks are generated. If we come up with a mathematical model to understand how a real world network is generated, right.

We can use the same mathematical model to create a synthetic network, right. For example, if we know that when a node comes in into the system, it is going to join, it is going to be connected with a node of highest degree for example, right. If we know that ok this process is followed when real world network grows, right. We can actually you know mimic the same process to create a synthetic network, ok.

But what is the exact underlying mathematical process, which is generally followed for you know the network growth process? Ok. So, the mathematical modelling is very vital and it can actually differ across different models, right. So, we will discuss you know random network model, we will discuss preferential attachment model and so on and so forth.

And basically over times people tried you know tried to come up with sophisticated models to mimic a real world network. Now what do I mean by a real world network mimicking a real world network? It means that you know in the synthetic network the properties.

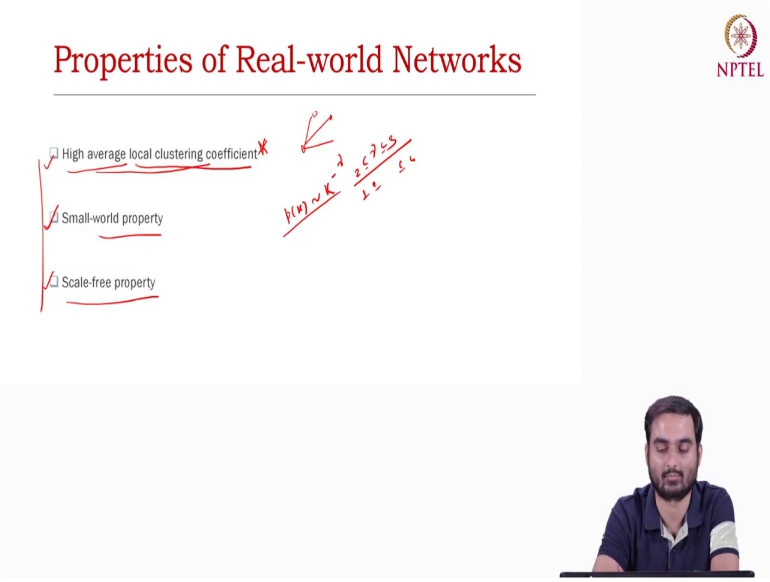
For example, you know the clustering coefficient degree distribution, right path length, right diameter all these network properties that are generally observed in a real world network those properties should also be preserved in a synthetic network, ok. That is very important. So, we know over the years people tried coming up with different you know different growth models to mimic how a real world network grows over time.

And they actually test the efficacy of their growth models you know by checking whether a particular property of a network is actually preserved, right or remains same across you know synthetic network and real world network. For example, you generate a network and then you test the degree distribution of the network.

If it its follows a power law then you say that ok you know it its looks like a you know real world network because real world network also has power law degree distribution, ok. So, power law is one property. There are many such properties and all these network growth models basically try to satisfy at least a subset of these properties that a real world network exhibits, ok.

So, we will see you know over this particular chapter that you know how the field of network growth models evolved over time. And, how more sophisticated models you know came into the picture and they actually tried to you know mimic real world properties, ok.

(Refer Slide Time: 10:23)



Properties of Real-world Networks

- □ High average local clustering coefficient*
- □ Small-world property
- □ Scale-free property

So, when we say that you know the synthetic network essentially mimics real world properties. Now what are the real world properties that we generally you know look at? We generally look at these 3 types of these 3 types, right. So, basically people studied a lot of such real networks and then they realized that you know all these real networks they basically exhibit high average local clustering coefficient.

So, local clustering coefficient is something that we discussed in the last lecture. If you remember it basically says that you know for a given node you look at their neighbours. So, and then you create an induced subgraph of neighbours and then you look at you know the actual number of edges among the neighbours and you divide it by the possible number of edges among the neighbours. That is the local clustering coefficient and then you basically take an average across all the nodes, right.

And it turned out that for real world network surprisingly the average local clustering coefficient is pretty high, ok. Meaning that even if say a node has very high degree for example, degree 100 or degree 200, right. So, as the number of as that you know degree increases the denominator of the clustering coefficient will also increase, the denominator basically says that you know what is the total number of possible number of edges that can be formed across neighbours this is NC 2 so, DC 2, D is a degree, right.

So, even if the you know the degree increases, right clustering coefficient remains almost same, right. And with the massive network it was shown that you know on a network of say 1 million nodes, right the clustering coefficient is still around 0.12, 0.13. 0.12, 0.13 is actually pretty good number, right. When we look at the nodes with very high degree and so on.

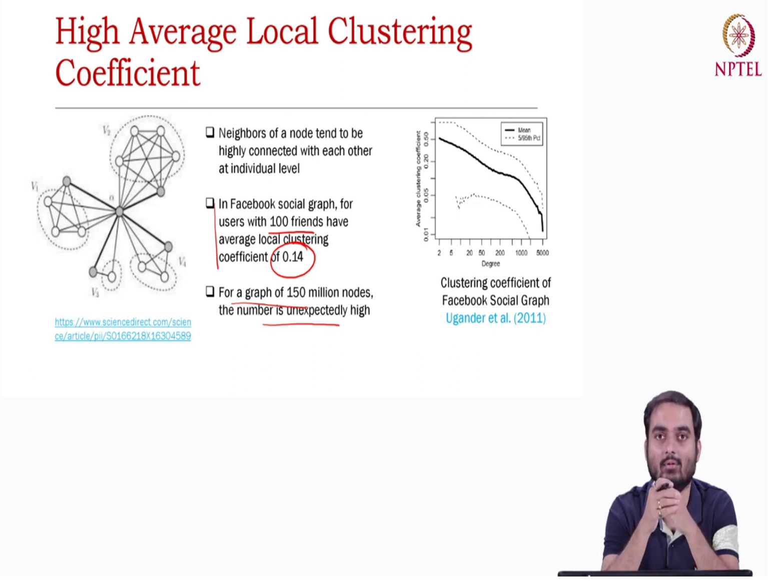
So, this property should be preserved in the synthetic network, ok. The second property that was observed across real world networks is a small world property, ok. I will discuss what is small world property, but if you remember in the last lecture we briefly discussed what is small world property.

Basically, says that you know if you take any pair of nodes from a network the distance is very small, the shortest path distance between any pair of nodes in a network is very small and how small it is we will discuss today. So, small world property and then we have this scale free property.

So, scale free property is again basically the degree distribution should follow a power law, right. And power law itself is scale free. So, when you plot the degree distribution of a synthetic network it should follow a power law, ok. And the kind of power law that we are expecting, right is basically K to the power minus gamma where gamma ranges between 2 to 3, right.

So, if your synthetic network follows this kind of property where gamma ranges between say if not 2 to 3, but at least 1 to 4 for example. Then you say that, ok you know it has a scale free property. So, we will try to see which network growth model actually follows all of these 3 properties, ok.

(Refer Slide Time: 14:02)



High Average Local Clustering Coefficient

- ❑ Neighbors of a node tend to be highly connected with each other at individual level
- ❑ In Facebook social graph, for users with 100 friends have average local clustering coefficient of 0.14
- ❑ For a graph of 150 million nodes, the number is unexpectedly high

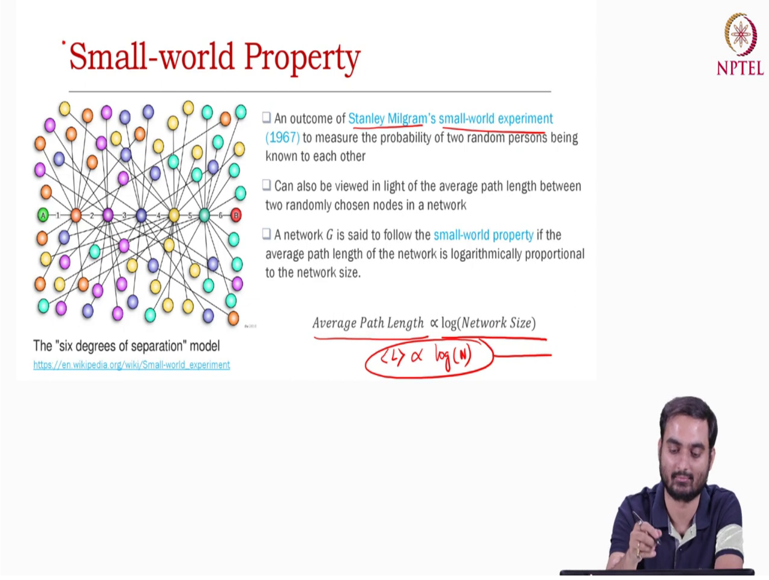
https://www.sciencedirect.com/science/article/pii/S0166218X16304589

Clustering coefficient of Facebook Social Graph
Ugander et al. (2011)

So, high average local clustering coefficient we discussed. It turned out that you know massive Facebook network with say with users having average 100 friends, the clustering coefficient is still 0.14 which is pretty large, ok. In fact, a graph with 150 million nodes the expected number of the you know the local clustering coefficient is unexpectedly is very high.

So, which may not look like you know trivial observation because we basically you know we can assume that as the network size grows, the average clustering coefficient may decrease, right naturally. But yeah it actually decreases, but it is not like you know it decreases massively, it is not like it I mean the clustering coefficient distribution you know does not follow the power law property, ok.

(Refer Slide Time: 14:58)



Small-world Property

- ❑ An outcome of Stanley Milgram's small-world experiment (1967) to measure the probability of two random persons being known to each other
- ❑ Can also be viewed in light of the average path length between two randomly chosen nodes in a network
- ❑ A network G is said to follow the small-world property if the average path length of the network is logarithmically proportional to the network size.

$$\underline{Average\ Path\ Length} \propto \underline{\log(Network\ Size)}$$

$$\boxed{\langle L \rangle \propto \log(N)}$$

The "six degrees of separation" model

https://en.wikipedia.org/wiki/Small-world_experiment

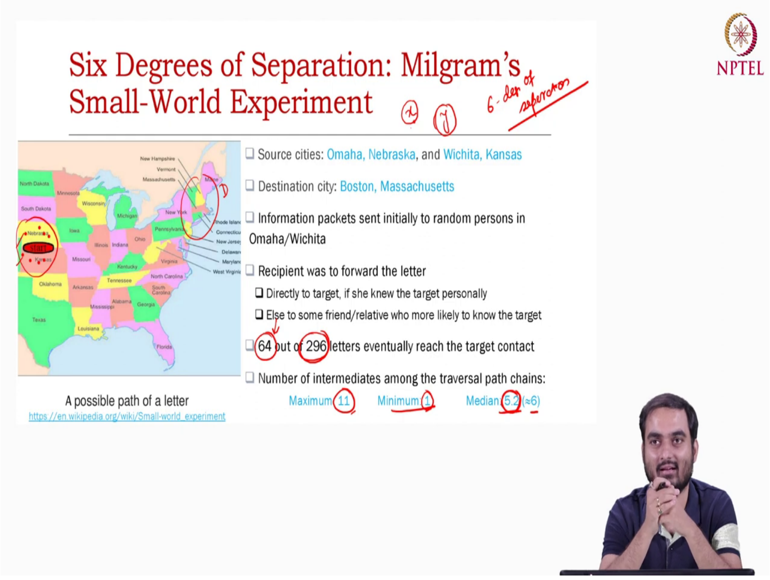
So, small world property lets discuss about the small world phenomena, right. So, the name suggest the world is small, ok. So, what does it mean? It basically means that you know if you send a if you want to send a letter, ok to the president of US, right. The distance that this letter will traverse, right the distance is pretty small, right. And how small it is?.

So, it turned out there was a very interesting social you know social science experiment you know way back 1967, 1968 by Stanley Milgram, right. And it turned out that the shortest path distance between any pair of nodes in a network is 6, ok. So, I mean it sounds very unrealistic, right.

So, it basically says that if we think of this world, right as a social network think of this user-user interaction network, right or a person-person interaction network, right. In the entire world if you choose any pair of users any pair of you know individuals the distance is less than equals to 6, ok.

It sounds little unrealistic, but I will show you that it is not that unrealistic if we do it mathematically, if we try to calculate these things mathematically, ok. But let us try to understand this experiment done by Stanley Milgram, right in 1967 and this experiment is called small world experiment.

(Refer Slide Time: 16:40)



Six Degrees of Separation: Milgram's Small-World Experiment

A possible path of a letter
https://en.wikipedia.org/wiki/Small-world_experiment

❑ Source cities: Omaha, Nebraska, and Wichita, Kansas

❑ Destination city: Boston, Massachusetts

❑ Information packets sent initially to random persons in Omaha/Wichita

❑ Recipient was to forward the letter
  - ❑ Directly to target, if she knew the target personally
  - ❑ Else to some friend/relative who more likely to know the target

❑ 64 out of 296 letters eventually reach the target contact

❑ Number of intermediates among the traversal path chains:
Maximum: 11 Minimum: 1 Median: 5.2 (≈6)

So, what you know what he has what he had done. He basically you know chose two extreme cities I mean 2 cities which are far apart from each other in USA. One is Nebraska, right Kansas and Nebraska and other is Boston, ok. Nebraska is here and Boston is here, ok. 2 parts of 2 different regions of United States, right.

And then he chose some individuals here, right. So, this is your source Nebraska and Boston is your destination, ok. And then you choose some users some individuals here, ok. And then to each individual you give a letter, ok and in the letter it is mentioned that the destination is in Boston, right.

So, the task of that individual is to send that letter to the destination, the destination is in Boston address in Boston, right. And how the individual can send this letter individual can send this I mean if the individual knows the destination address directly. So, he can just go and drop the letter to the destination, ok.

If the individual does not know the destination properly, what he can do? He can basically choose a friend who the individual thinks may know the destination better, ok. Say this individual is x ok and so x if x does not know the destination. So, x will choose another friend. So, y should be the friend of x, right.

So, x will choose one of his friends y, which x thinks that you the friend y knows the destination better than x, ok. So, if I do not know the destination I will basically pass the letter to somebody who I feel that knows the destination better, ok. So, when the letter comes to y will again see the destination address. And if y knows the destination y will directly send it otherwise y will again choose one of his friends who y assumes that knows the destination address better, ok.

So, this is the process. So, therefore, you see that you know when the letter passes and when you pass a letter from when a letter is passed from x to y, right. x will write you know x will basically write his address and pass it to y and then again when y passes it to z. So, y will write the y will write his address and pass it to z, ok.

This is the, this is the strategy, ok. And then they and you know around 296 such letters were given to individuals in Nebraska and Kansas. And it turned out that and say when a letter when the letter moves from x to y then y to z you can easily see the number of hops and you keep on counting the number of hops for letters, right.

Now, it turned out that there was at least one person one individual who knew the destination address directly and he directly sent the letter to the destination. Therefore, the minimum hop that was required to send the letter was 1 and the maximum hop that was required, right to send the letter to the destination is 11, ok. Of course, I mean out of 296 only 64 letter reached the destination, right.

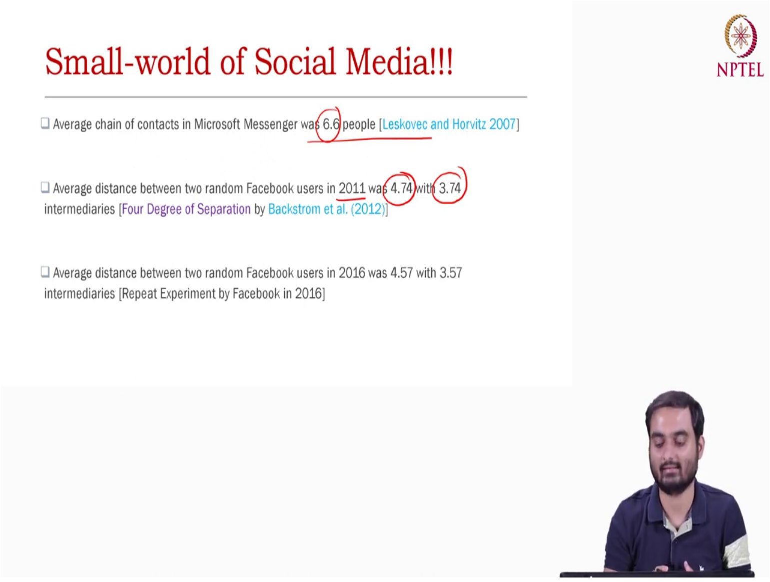
And in out of the 64 cases the minimum was 1 the maximum was 11 and the median was 5.2, ok. So, and from this number 5.2 Milgram you know invented coined the term 6 degree of separation. 6 degree of separation it is a very famous term. 6 degrees of separation, meaning that if you choose any two individuals, right from this world, right. There are there would be actually 6 hops between these 2 individuals ok.

So, average path length average shortest path length between any pair of nodes in a network is 6. Now why 5.2 is approximated to 6 why not 5, I do not know [laugher]. There is no reason I mean in Milgram thought that, ok I mean we should always take the upper limit not the lower limit. Therefore, he suggested this 6 as the shortest path distance, right.

I mean Milgram was very famous for all this you know very surprising interesting experiments. Of course, he was a social scientist. So, this was one of the experiments that for which he was very famous. And of course, people then started shouting because you know this number is very insignificant only 64 letters were actually sent, 64 letters reached the destination.

How can you come up with such a strong claim based on only 64 letters, right 64 samples, right.

(Refer Slide Time: 22:14)



Small-world of Social Media!!!

□ Average chain of contacts in Microsoft Messenger was 6.6 people [Leskovec and Horvitz 2007]

□ Average distance between two random Facebook users in 2011 was 4.74 with 3.74 intermediaries [Four Degree of Separation by Backstrom et al. (2012)]

□ Average distance between two random Facebook users in 2016 was 4.57 with 3.57 intermediaries [Repeat Experiment by Facebook in 2016]

But nevertheless, I mean later on people actually tried to come up with you know even better measures and so, on for example, you know there was an experiment done by Leskovec and his team in 2007 and the same experiment the small world experiment was performed on Microsoft messenger.

So, right Microsoft also had you know its own messenger kind of service, like we have Facebook messenger these days. And the same kind of experiments were actually conducted and it turned out that the actually the number is 6.6, which is quite same as the one that Milgram suggested with you know only a tiny amount of samples.

The same experiment was again repeated in 2011, right. But this time on the Facebook network because Facebook that time became already popular and this experiment the repetition happened 2 times, in 2011 it was happened. And it turned out that the average distance is actually 4.74 in the on the Facebook network, right. In 2012 the same experiment was the same experiment was repeated and the average you know distance turned out to be 3.74.

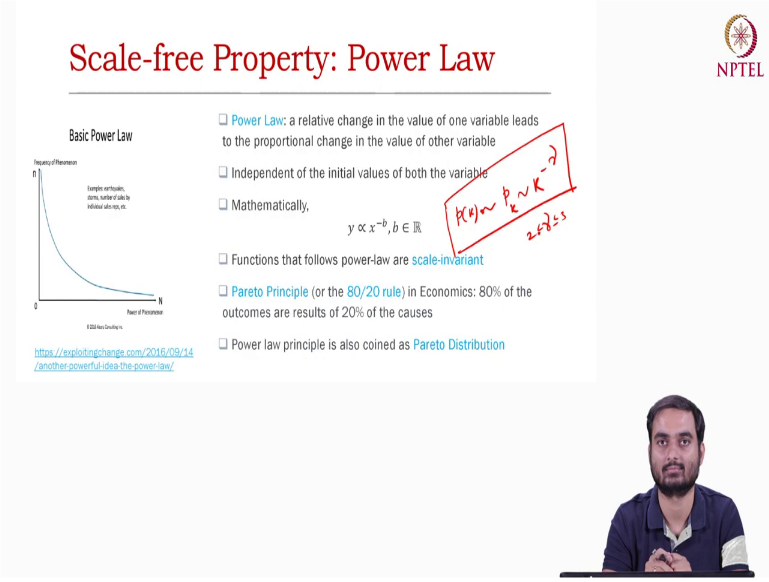
So, as you see over the time as the size of the network increases, right. It is also obvious that the distance between individuals should also decrease and it is actually happening from 6.6 to 4.74 to 3.74 and so on, right. So, in fact you can also do the similar experiment on the Facebook graph on the facebook social media.

There is an app there is a very interesting app and through that app you can check what is the degree of separation, right. Of your user account you know from the other users present on Facebook. Actually, I mean I did this experiment I tested it to check my degree of separation it turned out around you know turned out to be 3. 3.74 3.75. You can also check your degree of separation, right.

So, this small world phenomena is basically you know very famous and it basically says that you know the average path length of any pair of nodes in the network is it should be proportional to the log of the size of the network, ok. So, average path length is denoted by this one mod of L.

It is proportional to the log of N where N is a number of nodes in the network, ok. So, if you prove that your synthetic network follows this kind of property, right. Therefore, you can basically say that, ok I mean one property is now ticked, right. So, at least with respect to one property of real world network your synthetic network behaves quite similarly, ok.

(Refer Slide Time: 25:21)



Scale-free Property: Power Law

Basic Power Law

Frequency of Phenomenon
n
Examples: earthquakes, storms, number of sales by individual sales reps, etc.
0
N
Power of Phenomenon
© 2016 Akaru Consulting Inc.

https://exploitingchange.com/2016/09/14/another-powerful-idea-the-power-law/

- ❑ Power Law: a relative change in the value of one variable leads to the proportional change in the value of other variable
- ❑ Independent of the initial values of both the variable
- ❑ Mathematically,

$$y \propto x^{-b}, b \in \mathbb{R}$$

- ❑ Functions that follows power-law are scale-invariant
- ❑ Pareto Principle (or the 80/20 rule) in Economics: 80% of the outcomes are results of 20% of the causes
- ❑ Power law principle is also coined as Pareto Distribution

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

$2 \leq \gamma \leq 3$

NPTEL

So, that was small world property the third property is scale free property we have already discussed earlier. So, scale free property basically says the degree distribution should follow power law, right it means that. So, if K is a degree. So, the this p K right it is also denoted by this one it should be K to the power minus whatever gamma or lambda where gamma ranges between 2 to 3 or you know 2 to 4, right.

So, if you prove that your synthetic network actually follows this kind of power law degree distribution, you can say that it resembles with the scale free property that the real world network exhibits, ok. So, we stop here in the next lecture we will discuss the simplest you know network growth model called random growth model random network growth model, ok.

That is the simplest one and although it is a very simple method, but we will see that you know it basically follows some of the real world properties that we have been talking about ok.

Thank you.