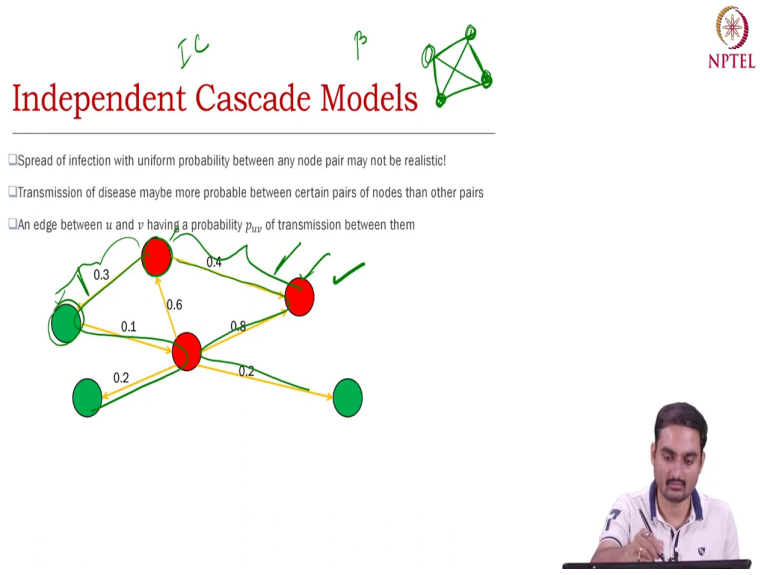
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 07
Lecture - 08

So, the epidemic spread model that we have discussed so far SIR, SIS, SIRS all these models. So, they assume that the graph is completely connected, I mean all the nodes are connected to, I mean each node is connected to the remaining nodes in the graph right. It is a completely connected graph and why we assume that? Because this is the simple you know representation of a population.

We assume that all the, I mean all the citizens in a population, they are friends of each other and that is the simple way and that is the worst possibility, I mean whatever you whatever prediction you make, you are making the prediction on the worst possible case, right.

(Refer Slide Time: 01:09)



Let us assume that you know the network structure and you also know that each node is infecting another node with certain probability ok. Let us say this is the graph ok and all these nodes are there. So, this node for example, infects this node with probability 0.3, this node infects this node with probability 0.4 and so on and so forth. And it is not the case that the sum of all the probabilities of the outgoing edges for a particular node would be 1, no.

So, with certain probability a node will infect another person, that node will infect another node with another with a different probability ok. If this is the setting, then how do we predict the cascade growth right? So, we will discuss a new model called IC model or Independent Cascade model ok.

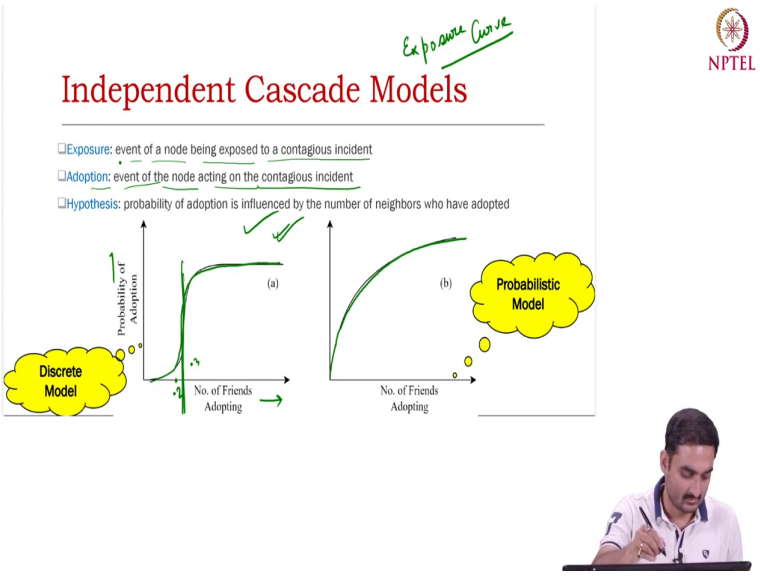
So, here the idea is that as I mentioned. The spread of infection with uniform probability between any node pair may not be realistic, in SIR SIS model we assume that the graph is completely connected and this guy is infecting this guy with same probability as infecting this guy. Similarly, this guy is also infecting this guy with the same probability.

So, all the edges, if you think of edges and the associated probabilities they are all same. The infection rate is same, so the infection rate is beta right, but in real world case as I mentioned transmission of disease may be more, may be more probable between certain pair of nodes, than the other pair of nodes, right. So, if this is the case then how do we deal with this?

Now, let us first understand how this model works right. So, say this guy is infected right and you look at you look at this pair first ok and with probability 0.4, you infect this guy right. Then you look at this node, this pair with probability 0.3 you infect this node ok. And why this is called independent cascade? Because this infection is independent of this infection, right.

So, if you look at a cascade through this path, this cascade is independent of this cascade for example, right.

(Refer Slide Time: 03:56)



So, now, let us look at the problems here. So, independent cascade model is simple, but if you think of all these probabilities associated with edges, all these probabilities are different. And how do we compute it? So, you can think of this probabilities as different parameters, that you can derive from the data set. So, you actually need to estimate a lot of parameters right, from the data.

So, what is the what is the simple solution. The simple solution is that you assume all the edges with same probability. So, all the edges have same probability of infecting others. So, if this is the assumption, then it is same as SIR model, then there is no difference the only difference is that it it is not a clique. Whereas, in SIR model we assume that the structure is clique, otherwise the remaining part would be same, right.

If all these numbers are same there is basically a single parameter beta, right. So, how do we you know how do we make this thing little better ok? So, we define something called exposure and adoption ok. So, exposure is basically a nodes neighbor exposes the exposes the node to the contagion, to the virus, right. So, exposure is basically event of a node being exposed to a contagious incident right and adoption is the event of the node acting on the contagious incident, right.

So, adoption is that the node actually acts as a contagion, it has already got infected and exposed is basically you know that you are exposed, but you are not infected so far right. So, what you do? We draw something called exposure curve ok exposure curve right. What is

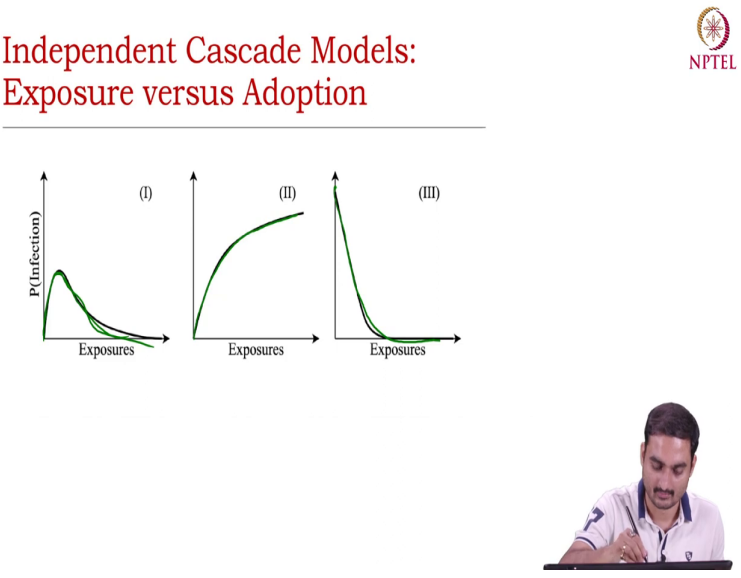
exposure curve? So, in the exposure curve we will basically look at the change of, the change of adoption right, the rate of change of adoption with the number of friends who are already, who have already adopted right.

So, the rate of change of infection with the number of infectious neighbors ok. For example, if you look at this curve right, x axis is the number of adoptions number of infections basically, number of friends who are already infected and y axis is the probability that particular node will be infected. Say this is 0.2, it means that if 20 percent of my neighbors are infected what is the probability that I will also be infected? I will adopt it, right.

So, if you look at the decision based model that we discussed earlier and this probabilistic models, for the decision based model you will see this kind of exposure curve, is this kind of S structure curve, which is quite expected because we know that in the decision based process model there is something called social pressure, with 40 percent or 50 percent infections of my neighbors, I will be infected right.

When majority of my neighbors have adopted something, I will adopt it. So, you see that a certain peak is here right, around 0.3, 0.4. Whereas, in case of probabilistic model, it basically increases like this, right. So, it means that there is no there is no concept of social pressure, it all depends on the probability P, whatever beta the infection probability right, it gradually increases.

(Refer Slide Time: 08:29)

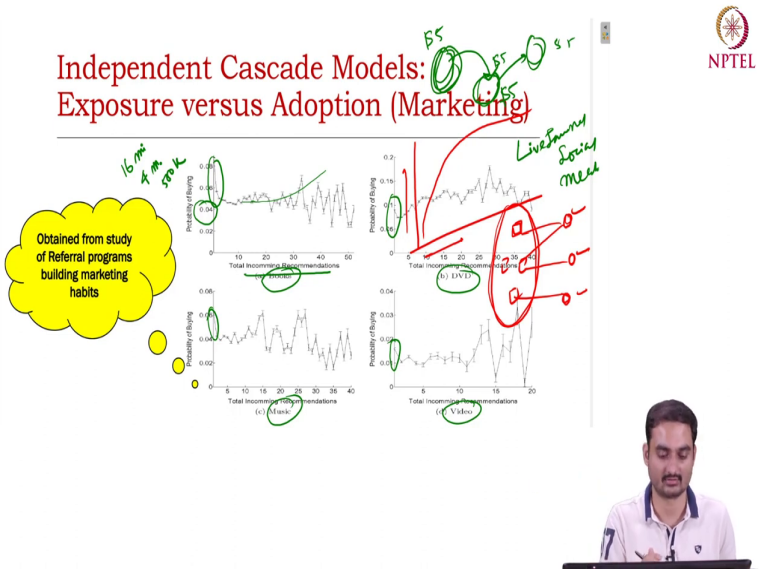


If you look at this exposure curve in other cases. For example, if you look at I mean you can think of different types of exposure curves, one type would be like this; monotonically increasing then decreasing right. These are basically cases where say you are infected, you have adopted a particular music right. So, you have suddenly started listening to a music and then you then you do not feel that attractive and then you do not feel that music that attractive and then you gradually you know leave ok.

Another type of exposure curve could be this one gradual increase over time. And other type of cases would be you know peak at the beginning and then a decrease, like this. So, it means that as soon as the product comes right, you immediately buy it irrespective of what your neighbors are doing.

Say for example, you are fond of apple right, so as soon as any apple product comes right comes to the market you immediately buy it, irrespective of what your neighbors will think about it or not. And then gradually you know that interest decays, with the increase of the number of exposed number of adopted neighbors right or whatever exposed or adopted neighbors, ok.

(Refer Slide Time: 10:00)



Now, let us look at you know some sort of viral marketing referral based marketing ok, say there are three persons right, if you if this person refers a particular product to this guy this guy will get 5 dollars, this guy will also get 5 dollars right. And if this guy refers the product to this guy right, he will get 5 dollars, the other guy will also get 5 dollars.

So, this is also kind of a viral marketing a cascade, but this is more of a referral kind of cascade, where there is an incentive associated with every purchase or every recommendation ok.

So, there was this experiment where around 60 million such recommended recommendations were collected and around 4 million people were there around 500 k products where were considered right. And if you look at this exposure curve right, number of purchasing versus number of incoming recommendation. So, number of incoming recommendation is basically number of neighbors who have already adopted it right, is also exposure curve. So, you see that say in case of books or DVD music videos, the patterns is more or less same.

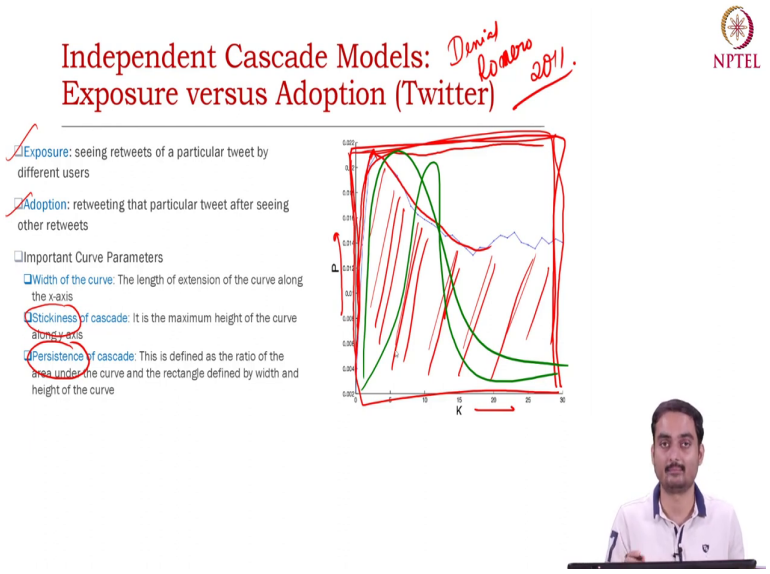
In fact, you see that in certain cases, so you know I mean always there is a peak at the beginning right, they are essentially fans of these movies or whatever books and DVDs. So, they would immediately buy it and with the increase of recommendation it will almost remain same right. There is not no such increase. In fact, some cases it will decrease right and if you look at the probability very very less 0.04, 0.606 and so on and so forth. So, these numbers are very less ok.

It basically indicates that this referral business does not help much in most of the cases ok, particularly for the data set that this experiment for this experiment was collected ok. Similarly, there was another experiment where people looked at this live journal social media, similar to Facebook, where say there is a group, there is a fan club and nodes are there ok, they are their friends.

So, this rectangular users they are already part of the club right and these circular users they are not part of the curve this group. So, what is the probability that this these guys will also be a part of the community group right. And it depends on the number of recommendations and here the recommendations are basically proportional to the number of users, who are number of neighbors who are already part of this social group, right.

So, if you plot the exposure curve for this one, you will also see the same pattern, x axis is the number of recommendations, number of neighbors who are already part of the group and y axis is the probability of joining that particular group. You will see that this kind of curve exists right. Some sort of diminishing return kind of initially increased and then after certain point it will remain constant, ok.

(Refer Slide Time: 14:15)



So, this was quite interesting. Now, if you look at this one. So, this is again same sort of experiment on a twitter data set right and this was done by Daniel Romero, one of my good friends from University of Michigan and this was published in 2011 ok. And in this paper, they collected Twitter data set and they looked at number of exposures and number of adoptions. Exposures basically seeing retweets of a particular tweet by different users, adoption retweeting that particular tweet after seeing the other retweets, right.

If you look at this parameter here. So, K is a exposure and P is the probability right, probability of adopting it. K is if you look at it here right. So, the curve increases, the value increases and then decreases after certain point, ok. So, now, they tried to understand the important aspect, different important aspects of this curve ok, right. And they actually defined a two quantities, one is called the persistence of a cascade this is a cascade behavior, the persistence of a cascade and the stickiness of a cascade.

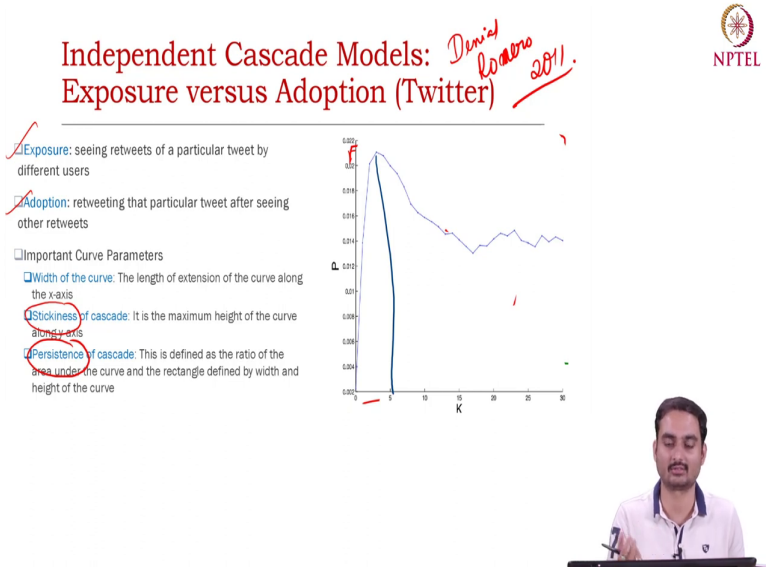
What is the persistence? Persistence is essentially the area under this curve right and it is it is basically the ratio of the total area ok, the total area and remember the total area is considered based on the maximum peak ok. So, based on the maximum peak, you draw a you draw a square right, you get square or rectangle whatever you get the area and you also get the area under this curve ok.

So, the persistence is the ratio of the value which indicates the area under the curve and the total area. So, if the ratio is high, it means that the curve would look like this. It is persistent

it, did not and if the ratio is low it basically means that something like this right, after certain point or whatever. After certain point it will decrease right. And that is the persistence right, higher the persistence higher the chance that the event will persist forever, right.

Stickiness is another metric which captures the maximum peak, the maximum height of the curve and in this case the maximum height as you see here is this one, is this one right.

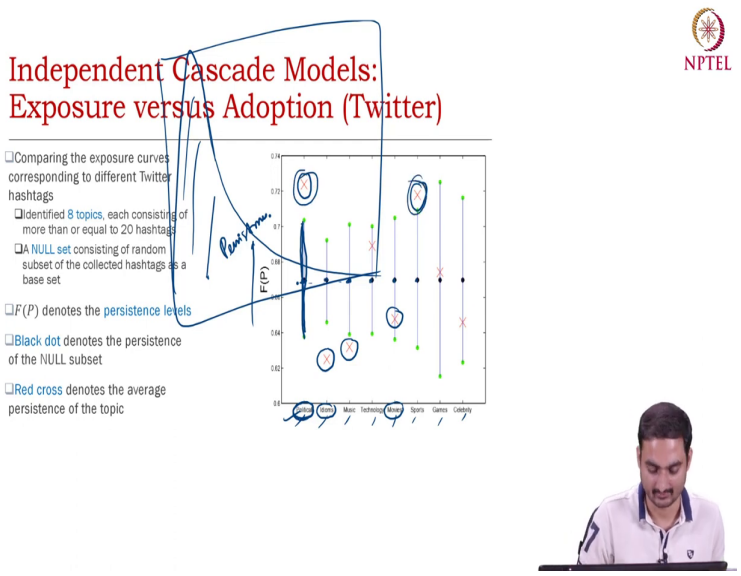
(Refer Slide Time: 17:28)



It basically indicates that the probability of usage as the maximum effective exposure right. When you have maximum exposure what is the what is the probability of that of that exposure? What is the probability of that usage? Right.

So, they plotted the values of persistence and exposure for different events, for different twitter events, for different I mean events which are whatever events or topics which are there on twitter, right.

(Refer Slide Time: 18:00)



So, now, these are events, political idioms, music, technology, movies, sports, games, and celebrity. So, you divide tweets into these different buckets ok and y axis is basically persistence ok.

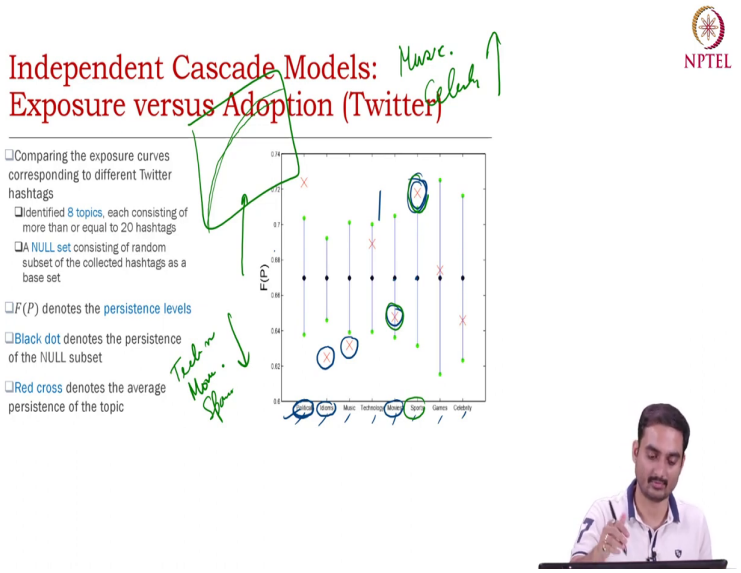
So, let us look at one event say political ok. So, they also define something called the null model. So, in null model you basically randomly choose a subset of hashtags, irrespective of the topic right and then you measure the persistence. That would give you the random model or null model. And whenever you compare any statistical test measure in statistical test you always do this with respect to a null model, right.

So, and this null model value, the persistence value for the null model is same across all the events. So, you see this number this dots this dots are the persistent values of the null model ok. And this bar graph or the error bar right this indicates the standard deviation, because you have to do this null model experiment multiple times because it is a random process right, you get a standard deviation and this is basically the mean ok.

Now, for political event, the persistence is quite high compared to the null model. For idioms persistent persistence is quite low, for music persistence is quite low, for movie also persistence is quite low, sports persistence is high; what does it mean? It means topics like music, movie right. So, these things generally get viral immediately after release therefore, you see a peak like this and then after certain point it will deteriorate, the popularity will deteriorate right cascade size deteriorate.

So, therefore, you see that the total area right, if you look at the area under the curve and the total area the ratio would be would be less right. You see here, whereas, topics like politics right sports. So, politics and sports these are topics which always you know last forever, run forever right. Therefore, you see the persistent values are quite high ok. Similarly, if you it is not drawn here in this particular slide.

(Refer Slide Time: 21:00)

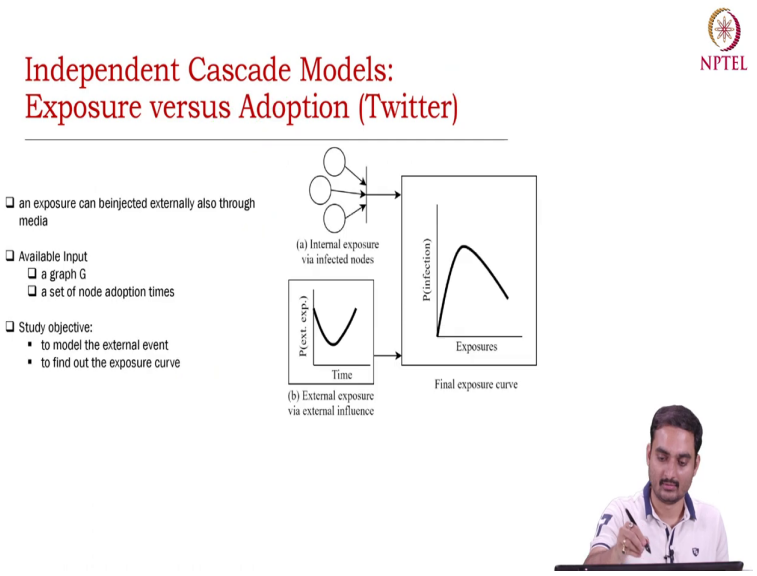


But similarly if you look at the stickiness right, you will see that the same plot y axis would be stickiness and x axis would be different events.

You will see for topics like technology right, technology, movie right, sports stickiness is quite low, stickiness is quite low ok. But for movie you see persistence is also quite low, for sports persistence is high, but stickiness is low; what does it mean? For it means that this exposure curve would possibly behave like, this ok, behave like this and cases like music right celebrity, the stickiness value would be quite high ok.

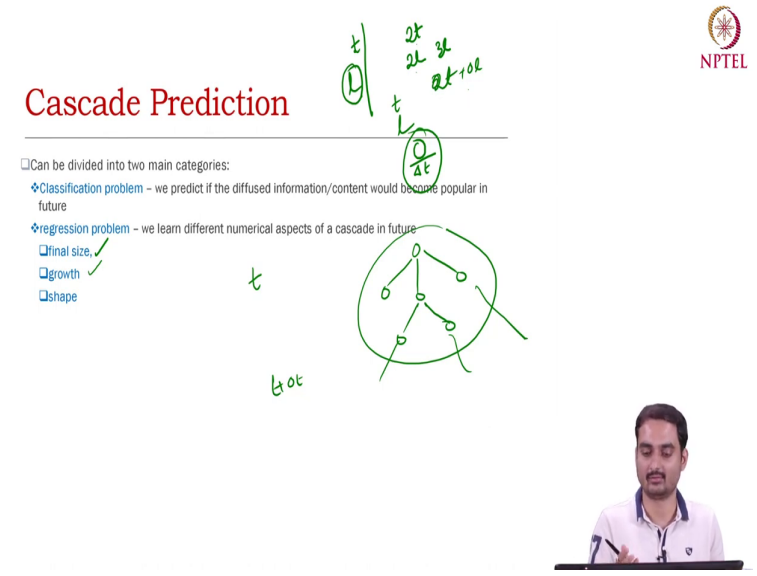
So, if you want to know more about this one you can either refer to the textbook that I mentioned the beginning or you can look at the original paper right, where this was published, alright.

(Refer Slide Time: 22:11)



Let us proceed. So, this kind of exposure curve right you would see in. In fact, you would also look at the expected number of the expected time when a particular node gets infected or gets or adopts a particular product or particular ideology and that you can plot with respect to the number of exposures and so on. Exposed neighbors and so on, right.

(Refer Slide Time: 22:42)



So, what task that you can basically model in case of cascade prediction? You can also you can think of it as a classification problem, you can say that ok right given time t you have a cascade depth of say l right, with time 2 t right whether the length this l th l would be 2 l or 3

l or not right. You can think of it as a classification problem, that given that at t you have the length l, length of the cascade l at 2 t whether it would be 2 l or more than 2 l or not right.

It is a classification problem it can be 2 l or 3 l or whatever you can you can basically run different classifiers, right.

Or you can think of it as a regression problem, you can say that ok at t you have a cascade length of l right. At 2 t or whatever at t plus delta t, what would be the cascade length? Ok, that would be that would be a regression problem right. So, you can predict the cascade length and the you can also predict the growth, the velocity of the cascade right, say you see that within delta t time what is the cascade growth and can we predict the cascade growth over time?

So, this is basically velocity ok, you can predict velocity, you can predict the length of the cascade, you can also predict the shape of the cascade. Meaning, if you think of a tree like structure right. So, what would be the properties of the tree? What would be the properties of the cascade tree after delta after t plus delta t time period? At time t you have a tree, you have this kind of tree.

After t plus delta tree this tree will grow, right. So, it would be very difficult to predict the actual tree, but you may be able to predict different properties of the tree, for example, the depth right the breadth or number of branches and so on and so forth. And from the property from this predicted property values of the tree, you can think of the shape of the tree ok.

So, this brings us to the end of this chapter. So, what we have learned so far a quick recap, we have basically learned two different types of models, decision based models and probabilistic models. In decision based models we have seen you know different cases where right, where you basically model it in terms of the social pressure right, then you then in the probabilistic model we basically look at epidemic spreading model. So, node will infect another node with certain probability right.

And this is different from the decision based model, decision based model is mostly useful for say understanding how products are being adopted right, how strategies are being adopted. Whereas, probabilistic models are useful for contagion spreading, epidemic spreading right.

We have seen models like SIR, SIS, SEIZ and so on, right. We have also discussed independent cascade model and applications of these models in different real world scenario on social network, such as beamer prediction, beamer propagation prediction and then also understanding exposure curve for different topics, different events, ok.

So, hope you enjoyed this particular chapter. This was one of the applications related chapter that we wanted to discuss. In the next chapter we will start discussing on anomaly detection, graph outlier detection right, we will see another application of graph right for social network analysis.

Thank you.