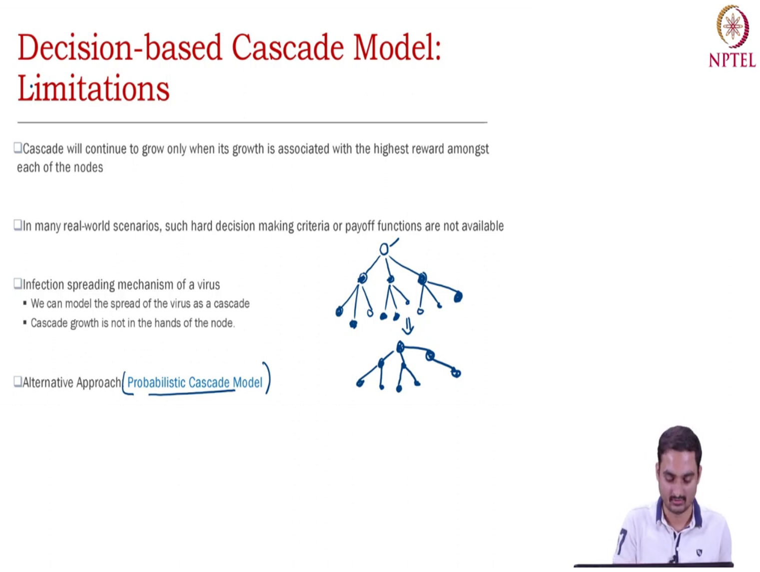
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 07
Lecture - 05

Ok. So we so, far we have looked at one type of cascade prediction model, cascade growth model which was decision based model. Now, we will look at another type of models called probabilistic cascade model.

(Refer Slide Time: 00:40)



Decision-based Cascade Model: Limitations

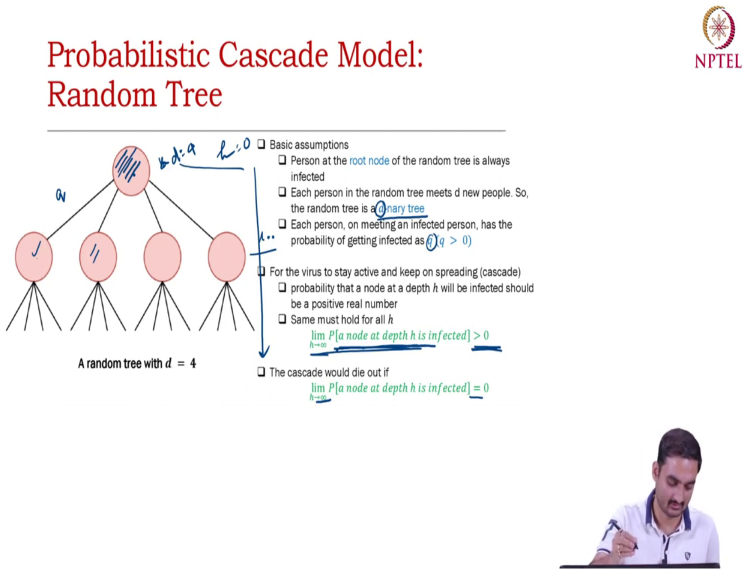
- ❑ Cascade will continue to grow only when its growth is associated with the highest reward amongst each of the nodes
- ❑ In many real-world scenarios, such hard decision making criteria or payoff functions are not available
- ❑ Infection spreading mechanism of a virus
  - We can model the spread of the virus as a cascade
  - Cascade growth is not in the hands of the node.
- ❑ Alternative Approach (Probabilistic Cascade Model)

So, what is the problem so, we discussed in the last lecture, what are the major problem in cascade decision based cascade model. The first problem was that you know the real contagion right for example, real epidemic spreading is not as simple as a decision based model, that you decide whether you will be infected or not and then you get infected right. It is more complex and more this is more of an unobserved event right.

So, and a lot of a lot of randomness is involved in this process. So, when we talk about you know a contagion spread right, we generally look at this kind of structure. We already mentioned, say this is the node and let us say the node has 3 neighbors. Out of these 3 neighbors, all these 3 neighbors have come to the come in close contact with this guy, but only these 2 neighbors have got infected.

Similarly, right similarly say each of these neighbors again has 3 neighbors and only some of them have got infected right and so on and so forth. So, from this tree like structure you can basically get a filtered tree like this ok, based on only the nodes which have got infected ok, like this right. So, our task would be to understand whether this kind of spread will go on and on right, whether they will stop eventually or it will not stop ok.

(Refer Slide Time: 02:38)



Probabilistic Cascade Model: Random Tree

- ☐ Basic assumptions
  - ☐ Person at the root node of the random tree is always infected
  - ☐ Each person in the random tree meets d new people. So, the random tree is a d-nary tree
  - ☐ Each person, on meeting an infected person, has the probability of getting infected as q ($q > 0$)
- ☐ For the virus to stay active and keep on spreading (cascade)
  - ☐ probability that a node at a depth h will be infected should be a positive real number
  - ☐ Same must hold for all h

$$\lim_{h\to\infty} P[a\ node\ at\ depth\ h\ is\ infected] > 0$$

- ☐ The cascade would die out if

$$\lim_{h\to\infty} P[a\ node\ at\ depth\ h\ is\ infected] = 0$$

A random tree with $d = 4$

So, to understand this, we basically use a setting which is called a random tree right. What is a random tree? A random tree is a tree like structure where and the depth is infinite, infinite depth and we assume that each node has same number of neighbors right. So, in this case we assume that each node has 4 neighbors, d equals to 4.

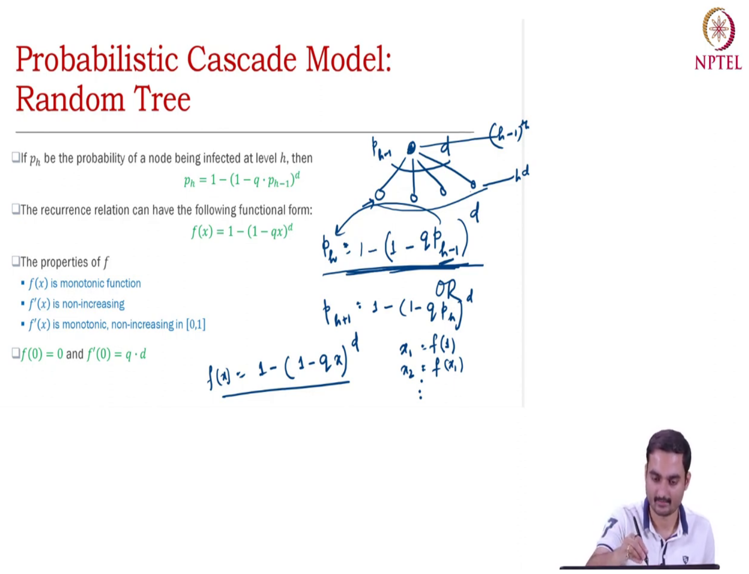
So, we assume a d-nary tree right. A node, each node has d neighbors and we also assume that the first node, the root node has already got infected and the probability that a person will infect another person is basically q right. So, with probability q this guy will infect this guy, with probability q this guy will infect this guy and so on and so forth.

So, we with 1 minus q probability, the neighbor will not be infected ok. And, what we are interesting interested in understanding is that for which value of this d and q right, the virus will spread will the this spreading process will go on and on, that the epidemic will last forever ok. So, that is the probability that we will try to measure in terms of q and t right.

So, essentially if you look at this depth right, depth h right, h equals to 0, h equals to 1 and so on. At every depth, if at least 1 node gets infected, then it means that this cascade will go on and on right. So, we will measure that what is the probability that at least 1 node gets infected at depth h ok and in the limiting condition h would tend to infinity.

So, we will basically measure this probability as h tends to infinity, what is the probability that at least 1 node at depth h will be infected? Ok. So, what is the probability, what is the case wow I mean what is the chance that this probability would be greater than 0 as h tends to infinity? If it is greater than 0, then the epidemic will go on and on right. And, if it is 0 in the asymptotic limit, then it will stop eventually, the epidemic will stop eventually right. So, we will calculate this probability alright.

(Refer Slide Time: 05:33)



Probabilistic Cascade Model: Random Tree

- ☐ If p_h be the probability of a node being infected at level h, then
$$p_h = 1 - (1 - q \cdot p_{h-1})^d$$
- ☐ The recurrence relation can have the following functional form:
$$f(x) = 1 - (1 - qx)^d$$
- ☐ The properties of f
  - $f(x)$ is monotonic function
  - $f'(x)$ is non-increasing
  - $f'(x)$ is monotonic, non-increasing in [0,1]
- ☐ $f(0) = 0$ and $f'(0) = q \cdot d$

$$p_h = 1 - (1 - q p_{h-1})^d$$

OR

$$p_{h+1} = 1 - (1 - q p_h)^d$$

$$f(x) = 1 - (1 - qx)^d$$

$x_1 = f(1)$
$x_2 = f(x_1)$
$\vdots$

So, let us try to understand. So, each node has d children right and the probability that the probability that say this is h minus 1th depth and this is hth depth ok and we will measure this p h. The probability that at least 1 node is infected at depth h, at least 1 node. How do you calculate it? So, what is p h minus 1? p h minus 1 is the probability that at least 1 node is infected at depth h minus 1. Let us say this is the node ok.

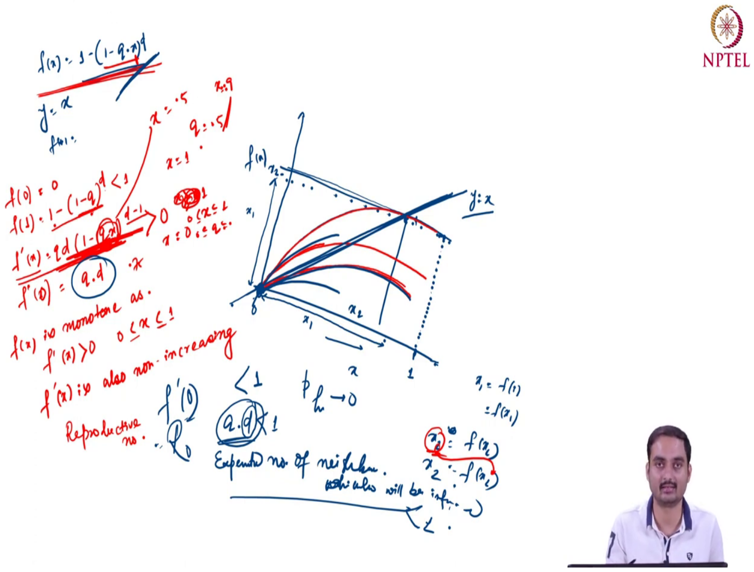
So, this node will infect each of its neighbors with probability q. So, with probability q, it will infect one of its neighbors. So, q times p h minus 1 is the probability that at depth h minus 1, there is at least 1 infected node. And, depth h that infected node will affect will infect another node with probability q. What is 1 minus this? This is the probability that node say I am talking about this node, this node will not be infected ok; to the power d that is.

So, this quantity is none of this d neighbors will be infected, 1 minus this is the probability that at least 1 neighbor will get infected. So, this is kind of an or function. We discuss similar kind of functions where, if you remember we discussed this thing in the big clamp formulation in the link prediction chapter ok. So, this probability which basically indicating that at depth h at least 1 nodes will be infected due to at least 1 node which has got infected at depth h minus 1 ok.

Now so, how do we this is kind of a recursive function, you see here right because in the next stage we will use it to calculate p of h plus 1. So, now how do we solve this problem? So, if you think of it carefully, this is essentially f of x equals to 1 minus right say f of so, you start from this root node and this root node is already infected.

So, the probability is 1. So, then you put 1 here, this is your f 1, this would give x 1. In the next iteration, you put f of x 1, you will get x 2 and so on and so forth right and so on and so forth. So, this we need to now calculate. So, let me erase it or let me let me see alright.

(Refer Slide Time: 09:28)



So, f of x is 1 minus 1 minus q times x to the power t. How do you solve this? Ok. This is kind so, there is no closed from solution right. So, how do we solve this? We discussed in the network growth model chapter right, Barabasi Albert model right, that we use it in a graphical form ok. So, there are essentially two things, one is y equals to x or and the other is whatever right f x is this 1 ok.

So, this is a y equals to x curve, this is x and this is f x ok and let us assume that this curve looks like this ok. What is x? x is the probability, ranges between 0 to 1 ok. So, this is 1 ok, what would be f 1? If we look at it here, this is 1, f 1 would be this one. This is f 1. So, now f 1 is x 1 right. So, in the next stage f of x 1 right. So, f of x 1 would be say this is this is your value right, this is x 1. I look at here x 1.

So, this is the same length ok x 1 and corresponding and say corresponding value this one would be this one say for example, right. Similarly, right similarly the next one would be x 2 and so on and so forth ok. So, we will keep on doing things keep on doing the this thing, until and unless we reach here, this point. What does this point indicate? At this point, if x int intersects y equals to x, what does it indicate?.

It indicates that so, say this is x 2 right. So, f of x 2 you get x 2 ok and you get x 2, you again put the same value within ff you get x 2 and so on and so forth. So, you kind of get stuck ok. So, you kind of get stuck, if you intersect at this point right. But, if you intersect at this point and what is this? This is x equals to 0 right and that would be your of that would be a solution. You basically want that you know that this probability value p h right, who tends to 0 for the epidemic to die out right.

So, what would be your target? Your target would be to come up with a function which would look like this. It would only intersect at x equals to 0 ok. So, let us try to understand the behavior of this particular curve ok. So, let us look at it. So, f of 0 I am trying to understand, because I drew this curve randomly. Now, let us try to understand the shape of this curve.

So, f of 0 is 0 right, because this would be 0, this would be 1, 1 minus 1 is 0; f of 1 would be 1 minus 1 minus q to the power d at x equals to 1 right. Now, think about it; what is q? q is the probability. So, 1 minus q would be less than 1 right to the power d would be less than 1. So, this would be less than 1, 1 minus this would be less than 1. So, this would be less than 1 right. Now, let us take the derivative f dash x, this would be sorry this would be d times right.

What is f dash is f dash 0? This is at x equals to 0, this would be q times d ok. So, now, let us see let us look at this one. So, if x is greater than 0 right or greater than equals to 1 right. So, what is x? x is the first probability right from where you started right and then the you get this one, then it will again come here and so on. So, x is also probability right. So, x ranges between 0 to 1, q also is a probability, q ranges between 0 to 1.

So, this would be less than 1 right, 1 minus this of course, I am not I am not considering the case when both x and q would be 0 ok, it is greater than 0. So, 1 minus this would be less than 1. This to the power d minus 1 would be less than 1, but greater than 0 right. So, the entire thing will be greater than 0. So, this would be greater than 0 ok. So, if the derivative is greater than 0, then this function is monotone.

So, f of x is monotone as f dash x is greater than 0 right yeah, that is sufficient to show that this is monotone ok, within the entire range of x ok. So, it is monotone right. Now, let us see what is the other property. So, f dash x is also non-increasing. What does it mean? Say look at this one ok, say x equals to with the increase of x. So, let us x let us say x equals to 0.5 and x equals to 0.9 ok.

x equals to 0.5 and let us say q equals to some number ok, let us say q equals to 0.5 ok. So, when x equals to 0.5 and q equals to 0.5, this would be 0.25 ok, 1 minus this would be 0.75 right. And, when x equals to 0.9, this would be 0.45 right and this would be 0.45 and this would be 0.55.

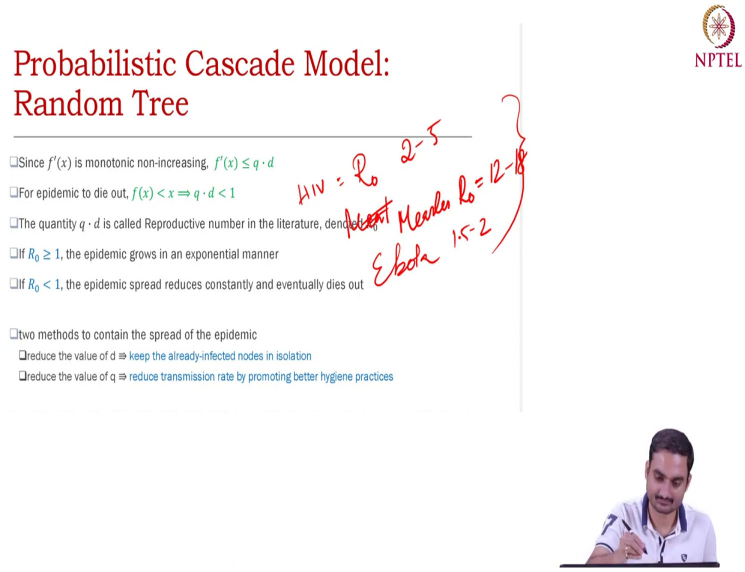
So, as x increases this will decrease which is very obvious. So, this will this is non-increasing so, the derivative the slope is non-increasing ok. So, the slope is non-increasing and the slope is also monotone. So, you can expect like this right, but it also it may also look like this. So, what is the what is the condition by which we can say that this curve would be below this point, would be below this one? Ok.

What is the condition? The condition would be at f dash 0 at this point right, thus the slope should be less than 1; think about it. At this point, the slope of y equals to x is 1. The slope of any line or any curve above this would be greater than 1 and the slope of any point would be less than 1, I mean any line which is below y equals to x would be less than 1 the slope. So, at 0 the slope is q times d.

So, q times d would be less than 1, in order to in order to in order to you know draw this kind of curve which will intersect only at this point right. And, what is this q times d? Think about it is very simple notation. d is the number of neighbors and q is the is basically the infection rate right. So, this is expected number of neighbors which or who will be infected. If this number is less than 1, then the epidemic will die out ok.

Now, this number q times d, this is called R 0 ok. So, this is called reproductive number, this is called reproductive number. So, for m e is for any epidemic, this reproductive number should be less than 1 in order and you can basically come up with policy strategy. So, that this would eventually be less than 1 so, that it would die out ok alright. So, let us look at and example case right.

(Refer Slide Time: 21:20)



Probabilistic Cascade Model: Random Tree

- ☐Since $f'(x)$ is monotonic non-increasing, $f'(x) \leq q \cdot d$
- ☐For epidemic to die out, $f(x) < x \Rightarrow q \cdot d < 1$
- ☐The quantity $q \cdot d$ is called Reproductive number in the literature, denoted R_0
- ☐If $R_0 \geq 1$, the epidemic grows in an exponential manner
- ☐If $R_0 < 1$, the epidemic spread reduces constantly and eventually dies out

- ☐two methods to contain the spread of the epidemic
  - ☐reduce the value of d ⇛ keep the already-infected nodes in isolation
  - ☐reduce the value of q ⇛ reduce transmission rate by promoting better hygiene practices

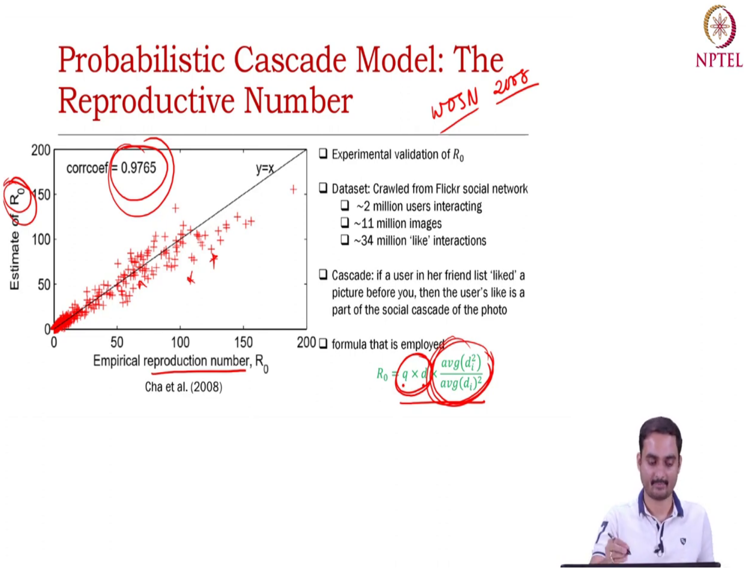
HIV : R_0 2-5

~~Next~~ Measles R_0 = 12-18

Ebola 1.5-2

So, and if you actually tally right, if you look at the real world cases this R 0 value for HIV was around 2 to 5, says for measles for measles R 0 was between 12 to 18. Think about the amount of infectiousness right. For Ebola, it was around 1.5 to 2. These are all scientifically shown ok.

(Refer Slide Time: 21:14)



Probabilistic Cascade Model: The Reproductive Number

WOSN 2008

Cha et al. (2008)

- ☐ Experimental validation of R_0
- ☐ Dataset: Crawled from Flickr social network
  - ☐ ~2 million users interacting
  - ☐ ~11 million images
  - ☐ ~34 million 'like' interactions
- ☐ Cascade: if a user in her friend list 'liked' a picture before you, then the user's like is a part of the social cascade of the photo
- ☐ formula that is employed

$$R_0 = q \times d \times \frac{avg(d_i^2)}{avg(d_i)^2}$$

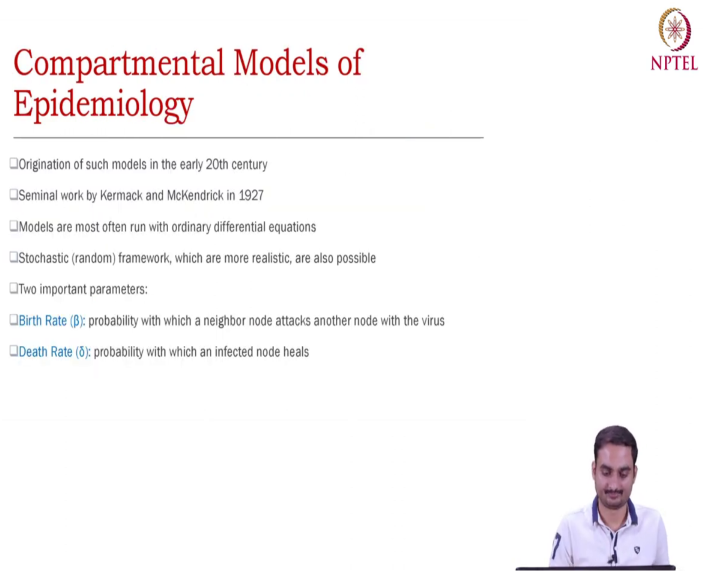
So, if you look at a real world online activity right, online spread; there is a nice paper ok of this you know that paper was published in ACM WOSN conference WOSN conference in 2008 which basically looked at how you know images on Flickr; so, Flickr is a famous online media site to upload images right, photos.

How photos spread on social media right and can we correlate the and can we come up with a reproductive number for this kind of spread of photo sharing behavior on social media. And, what they did they calculated the empirical reproductive reproduction number right, empirically right. So, for so, empirically how can you do? So, for every node you basically look at how many neighbors will are infected and so on and so forth.

And, you get an estimation using the theoretical value R 0 which is q times d right and it turned out that the correlation is really-really quite high 0.97 right. You see that and all these points right this cross points, these are different data points and you fetal you your fetal line and you see that the correlation is quite high 0.97 ok.

However, they use this formula, they used R 0 as q times d times some factor. So, this factor basically takes care of the skewed nature of degree distribution, the cascade distribution. This is some this is not that important, but they all they basically took into account this q times d factor to calculate the estimated value of R 0 ok.

(Refer Slide Time: 24:07)



Compartmental Models of Epidemiology

- Origination of such models in the early 20th century
- Seminal work by Kermack and McKendrick in 1927
- Models are most often run with ordinary differential equations
- Stochastic (random) framework, which are more realistic, are also possible
- Two important parameters:
- Birth Rate (β): probability with which a neighbor node attacks another node with the virus
- Death Rate (δ): probability with which an infected node heals

So, we stop here. In the next lecture, we will discuss about some other epidemic models, SIR model, SIS model. These kind of models which are primarily taken from standard biological epidemic, I mean standard epidemic growth model and those models are quite dated.

Because, but those models are quite important because when whenever we you know whenever you try to come up with the new model for capturing the spread of any disease right in epidemic, we start off by using those models and then we come up with sophisticated ways ok. So, we will discuss those models in the next lecture.

Thank you.