

Social Networks
Prof. S. R. S. Iyengar
Department of Computer Science
Indian Institute of Technology, Ropar

Rich Get Richer Phenomenon – 2
Lecture - 127
Rich Get Richer – The Long Tail

We have seen power law on networks. There is a different kind of power law that is in play in nature right and we are going to study more about it in this chapter in this particular lecture.

(Refer Slide Time: 00:22)



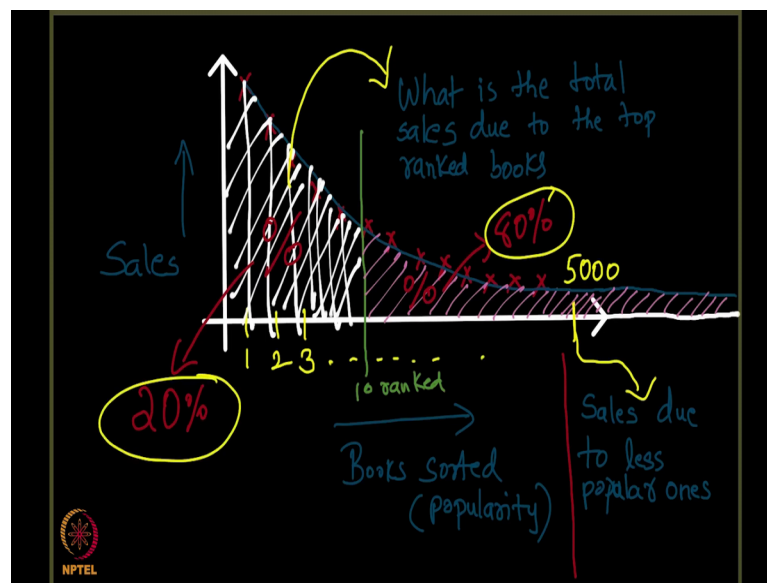
So, let me write more power law. So, this is mostly outside the realms of graphs and networks. So, let us go to a book store. So, assume you go to a book stores and you browse through some books right, go to a book store and you see a book 1 and there are several books in the shelf as you know.

I am not good at drawing; let me make some attempts to write books. So, here is a book store with a lot of books alright, books. There are different types of books, different genres ok. Now, there is a huge collection generally, as you would have observed there is generally huge collection of books and a question is why would a book store have such huge collections right.

There firstly, there are different genres of books as a huge collection. Why would anyone have something like this right? If I will were to run a book stores, I would simply have the best sellers right. I would simply have the best sellers. So, roughly let me assume that 10 percent of the books that are there in the book stores are best sellers, best sellers. So, why not just keep the 10 percent of it and discard the remaining 90 percent; it only makes sense you see right.

If I were the book shop owner, I would simply keep the best sellers and discard the rest why then do book stores any big book store that you visit they have a whole lot of books, why cannot they just keep the best sellers? Now, what is what is the obvious answer to this question? The obvious answer to this question is that we should probably ask this question in a different way.

(Refer Slide Time: 02:52)



I should plot this graph ok, let me just make it clear to you people slightly involved.

I will see the most popular the first most popular book and the second most popular, the third most popular book, so on and so forth. Assume, I have some 5000 such books. So, what is my x axis? My x axis is going to be popularity rank of the book. These are these are books sorted according to popularity, according to popularity fine. What is on the y axis? It is the sales of this books, the sales of this books. Obviously, the most first rank book will obviously, have very high sales, second rank book will have slightly less sales, third rank that is what you mean by ranking you say, correct.

Evidently there will be this power law like distribution correct, correct. So, on let me join these points and as you can see it will be a curve like this. Now, this looks like the same old theory of power law, but power law is slightly different as you know, it is about the distribution of the node degree correct. So, but the curve here resembles that. Now, let us get back to our question, let me go to the previous slide and let me help you remind let me see the question that I asked.

Why not have only the best sellers and discard the not so best sellers ok? I want to discard this and only retain this. Why because at my profit if my profit is mainly based on a best sellers, why then have this the remaining stuff 90 percent of the books are not best sellers. Let me just discard them. Now, what is so, wrong about this hypothesis of my book store business that I am going to start very soon of only the best sellers?

Well, that is where we need science rather smart accountancy or smart economics understanding of economics the situation. What we should do is think for a minute, what one should look at it is what is the amount of sales of the top few. Let say let me let me just take the first 10 which is let us say here; let us say first 5 or first 10 whatever that is. So, let us say first 10 ranked books; top 10 rank, the first 10 books I will see what is the sales of this.

A very small a very whatever I mean, it does not take a lot of time for you think. What is the answer for this Let me just write that down. So, that it seem to our minds. My question is what is the total profit that I might possibly make profit due to the top ranked books. What do I mean by this? I just mean what exactly I mean it is thought will tell you all I am trying to say here is what is this plus, this plus, this plus, this plus, this plus, this plus this. So, basically it is the area below this curve right.

I mean whatever is the area that will denote what will that denote that will denote the total profit total sales rather not necessarily profit I think I should remove the word profit here, it is not profit it is the sales. What is the total sales due to the top ranked books? It is going to be this curve correct and definitely this is going to tell me what is the sales due to the top ranked books and the remaining stuff which is this will tell me what is my sales, what is my sales so on; this will go on as you know right.

This will extend, this does not stop here it just extends to all 5000 books. What does this denote? This denotes this denotes what? Sales, sales due to a line here; sales due to due

to less popular books, less popular books. So, what I am trying to say here? All I am trying to say here is I am looking at my sales due to my popular products versus sales because of my niche product. If this sales is very high; now please note these are the sales because of best sellers. Just because they are best sellers and just because they sell in big numbers does not mean that they make most of my complete sales.

So, a good question for you all to ask which I am sure this is running in your mind already is what percentage of my sales is this and what percentage is this. To get your clear understanding I think you people should rather not see the curve this way which is slightly misleading you must know that the curve here the blue line extends like this, so does the white line here, the y x axis and then the area below the curve also extends like this.

So, good question to for one to ask here is what percentage is this ok, what percentage is this and what percentage is this right. Just in case, this was only 20 percent; you know what I am going to where I am getting at right now. This will be 80 percent. Now, you know very well that I cannot have my book stores with just the popular products, I need the niche products. So, niche means, the less popular ones.

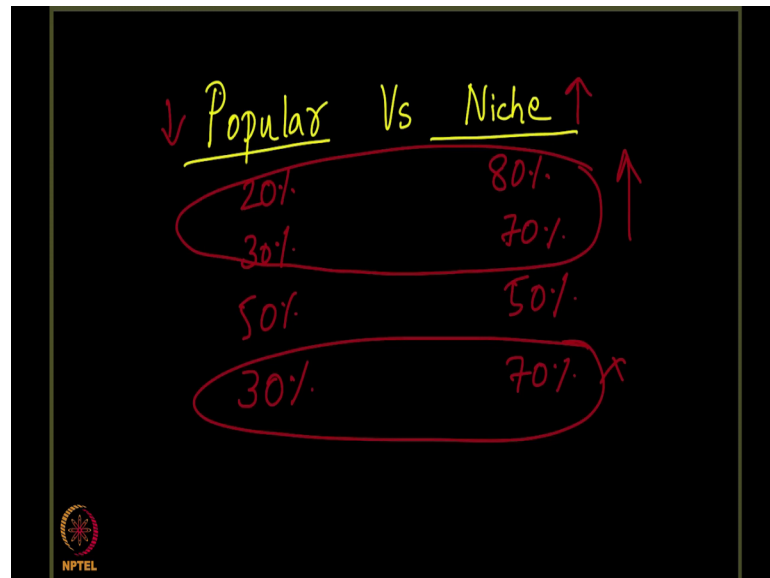
So, what are all here? As you can see let us say the past best sellers might be here, the best seller will be a best seller for a few months and then it will come here right or may be less popular books. Science books unfortunately are less popular because not many people read it. So, more scientific books are all here or technical books all here, novels and all the masala stuff is here, although you see that this may not be a big chunk of your sales, it can just be 20 percent and this can just be 80 percent.

So, the point here is you may want to concentrate on this more than this although your sales is mainly because of the bestselling book. So, you think if you write a curve like this and then see you will realize that it may not be. So, one thing that you almost observed is this. Do you think a book here continues to stay here? It will probably come somewhere here very soon right. Every best seller has it is shelf life after while it will stop being a best seller and it will become a niche product, it will become a classic right.

So, it is very interesting to see how this curve continuous to stay like this where more popular stuff keeps coming here and less popular stuff with time goes to this part of the

curve. And what is more interesting is what exactly is the percentage? 20 percent, 80 percent is what I am saying it could be different.

(Refer Slide Time: 11:36)

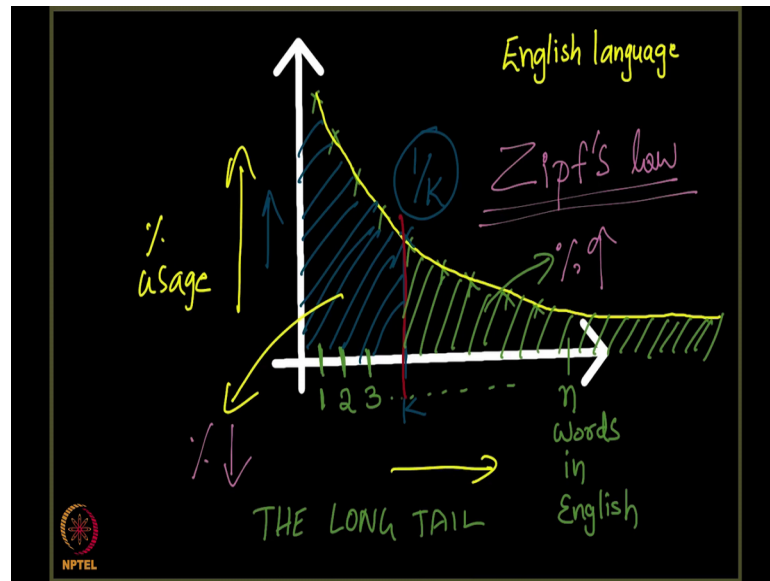


So, in your business you may have to see what is the popular products; popular versus the niche ones. You should have a good understanding of this ratio. It could be 80 percent; I am sorry just a minute whatever I said 20 percent and 80 percent or this could be 30 percent and 70 percent or it could be 50 percent and 50 percent it is always observed that something like this never happens, something like this never happens.

It is slightly on the higher side; the niche is on the higher side, popular is slightly on the lower side right. So, this speaks in volumes on why when you go to a book store you have all sorts of products. You cannot just keep the best selling once right. This is generally true no matter what business you are thinking of.

Let us say the kind of audio videos that people watch on YouTube, YouTube cannot afford to keep only the most watched video saying that people who watch this videos less rather I mean videos which are being watched less should just be removed right. So, the big chunk of what people watch is probably indeed the niche ones is probably the niche ones and less so of the popular once ok.

(Refer Slide Time: 13:22)



So, this has been there historically, in sciences people have observed this property in many many works of science.

One such thing is a very interesting fact which is let me try plotting and tell you what is this. If you look at the word frequency, let us not get lost. I just started off with this book store example and then I showed you this phenomena that happens and then I told you that popular sales, sales of popular products or sort of less in ratio as compared to the niche products and I also observed we also observed that a product that is there as a popular product might become a niche product.

It is interesting to see how this curve is preserved throughout in sales right ok. So, this is not this very similar to what happens in general with nodes as well. A node that is popular may not continue to be popular with time. It might become less popular and it might go this side right. What kind of questions do you have in mind? In fact, one can ask several research questions on this topic. So, I just right to me over email in case you want some cool questions on this topic. There are several nice very nice questions that one can attempt that the world does not know rather ok.

So, as trying to plot something very interesting about English language; about your English language, ok. If you take up the most popular word in English which is the most frequently used word in English call it the first ranked word and the second most

frequently used word in English call it the second ranked word and the third most frequently used English used word in English call it third rank and so go on.

And write down all the words let us say there are n words in English, you write n units on x axis English and then you see what is the number of times what is the proportion of times that you use this word. The first ranked word how many times, what percentage is it is usage. For example, in a document, if you think the word `the` is most frequently used I will ask what percentage of the document is it frequently used. So, the plot looks again similar to this is and then if you join this points you will get a power law like that is not a good way to write a curve ok.

And you will again get the curve like this and this curve is known to be known to follow this function $1 \text{ over } k$. So, if this is K , this will be $1 \text{ over } K$ and of course, I have used scale y axis. So, this might appear to be like a this distance seems the same as this distance, but do not mind my drawing the fact is this turns out to be $1 \text{ by } K$, if you look at the distribution of words in English language based on their ranking sorted based on their ranking and this is popularly called the Zipf's law. You can you can take a look at it on the internet.

It is a very popular law found several decades ago and this again says that the amount of frequently used words in English, same old theory back again if you see. That are there are there is this part of the curve which talks about the frequently used words in English, these are the words that are very frequently used rather popular words and this side is the sort of less popular words, but if you see this tail goes on like this ok; this is a big percentage this a big percentage.

I mean the area here is way bigger compared to the area here; this is a smaller percentage. Again this is observed in words to know. So, coming for coming getting back to where we started from, it is not just true in books book stores like business. It is also true in English language know that is surprising. It is observed in many places right and this is called the fat tail phenomena rather the long tail not the fat it is called the long tail. What is the long tail mean here? I am sure you all of guessed what I am trying to say here.

By this I mean this part of the curve this part of the curve is it goes on like this. It is start from here; it goes on like this given that this tail is longer do not have this part of the

curve looks resembles the tail of an animal that is probably the reason why they call it the long tail.

Since, it is long, it acts to a whole lot of the area as compared to the popular part of the curve right. So, that is why this is called the long tail phenomena. What is long tail phenomena? Long tail phenomena is nothing but the fact that; so, the popular products are less in number although they are popular and less popular products are more in number although they are less popular the sort of dominate the space.