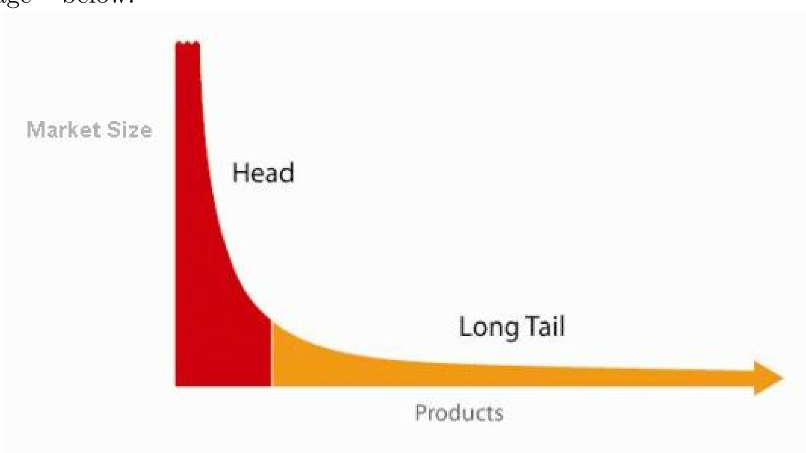
MS&E 235, Internet Commerce

Stanford University, Winter 2007-08

Instructor: Prof. Ashish Goel, Notes Scribed by Shrikrishna Shrin

Lecture 7a: Long Tails

Imagine a market where some class of products is being sold. We plot the i^{th} largest product on the X axis and the market size (Revenue/No of units sold/No of customers) on the Y axis. A long tail typically looks like the graph in the image [1] below.



The graph is said to have a long tail if it can be represented by an equation of the form $m(i) \approx a(b+i)^{-e}$ where e is called the exponent. Notice that the market size only decreases polynomially with increase in i.

For simplicity, let us assume that $m = i^{-e}$.

We can plot another curve with market size on the X axis and frequency (number of items with the corresponding market size) on the Y axis. We find that as market size increases, frequency decreases.

$$f = -\frac{di}{dm}$$

It is observed that long tails are created as a consequence of human endeavor or manufactured (ie. not natural) activities of human beings. Natural traits however do not exhibit the long tail phenomenon and tend to decay exponentially (of the form b^{-i}, notice that i is in the exponent in this case) instead.

A few quotes on how long tails are created:

> Long tails are a signature of human activity.

[1] Original image obtained from http://www.entrepreneurs-journey.com/images/long-tail-graph.gif

Long tails are a result of the tension between human individualism and social behavior.

For example, if customers were to walk into a shoe store, some customers would buy a Nike shoe because it is a popular shoe and many of his friends may have bought that shoe (social behavior). However, some people would prefer to buy a shoe that is not very common and it is this tension between social behavior and individualism that leads to the long tail. If all customers bought a shoe only based on social behavior, then everyone would buy the same shoe. On the other hand, if everyone bought a shoe only based on individualism, everyone would buy a different shoe (assuming there are enough choices).

Modeling the long tail phenomenon

Let us assume that customers arrive one by one at a music website. Every customer does two things.

1. Choose a previously heard song with probability proportional to the number of times it has been heard in the past (Social behavior)

2. Find a new obscure song and listen to it.

After t customers arrive, we have $2t$ hearings and t songs that have been heard.

Let $m_t(1), m_t(2), ...m_t(t)$ = the number of times song i has been heard after t steps. Probability that a customer at time $t+1$ listens to the first song is $m_t(1)/2t$.

Probability that $t+1^{th}$ customer chooses song $i (i \leq t) = m_t(i)/2t$.

Somewhat informally, $m_t(i)/2t$ can be thought of as the expected increase in volume of product i (song i) at time $t+1$.

Therefore, $m_{t+1}(i) = m_t(i) + m_t(i)/2t$.

Rewriting the above equation we get,

$$\frac{m_{t+1}(i) - m_t(i)}{t+1-t} = \frac{m_t(i)}{2t}$$

On taking the limit, the above equation becomes:

$$\frac{dm_t(i)}{dt} = m_t(i)/2t$$

On rearranging the above equation we get:

$$\frac{dm_t(i)}{m_t(i)} = \frac{dt}{2t}$$

On integrating with limits i to t ($m_t(i)$ is not defined for $t < i$) we get:

$$\ln(m_t(i)) - \ln(m_i(i)) = \frac{1}{2} * (\ln(t) - \ln(i))$$

Now, $\ln(m_i(i))$ is 0 as at $t = i$, $m_i(i) = 1$.
Therefore, we get:

$$m_t(i) = \sqrt{\frac{t}{i}}$$

This equation is the power law. Suppose customers were to choose a new product with probability p and an old product with probability $1 - p$ then the only change in the power law will be the value of the exponent.

Now we can obtain the frequency as:

$$f = -\frac{di}{dm} = -\frac{1}{\frac{dm}{di}}$$

$$= \frac{\frac{2}{3}}{\sqrt{t}\frac{1}{i^{\frac{3}{2}}}}$$

$$\approx \frac{i^{\frac{3}{2}}}{\sqrt{t}}$$

Substituting for i we get,

$$\approx \frac{(\frac{\sqrt{t}}{m})^3}{\sqrt{t}}$$

$$\approx \frac{t}{m^3}$$

As m increases, f of selling the item decreases. However, the volume of sale can still be significant.

Determining the median

The median is that point at which half the market lies in the tail.

Now, we can derive that market size $= 2t$ as follows:

$$m_i(t) \approx \frac{\sqrt{t}}{\sqrt{i}}$$

Integrating from limits $i = 1$ to $i = t$ we get,

$$\int_1^t m_i(t)\, di = \sqrt{t} \int_1^t \frac{1}{\sqrt{i}}\, di$$

$$= 2\sqrt{t}(\sqrt{t} - 1) \approx 2t$$

Let M be the total market size and med be a value such that $\int_1^{med} m_i(t)\, di = \frac{M}{2} = t$ $(M = 2t)$.

(Half the market lies below this product and half the market lies after this product (ie. in the tail)).

Therefore, $2\sqrt{t}\sqrt{med} \approx t$.

Which implies $med \approx \frac{t}{4}$.

The top 25% of the products only has half the mass. The remaining 75% contains the other half.

General form of the long tail

The general form of a long tail is:

$$m_i(t) \approx ai^{-e}$$

When $e < 1$, $med \to \inf$ as $t \to \inf$.

When $1 < e < 2$, med is bounded, expected product rank $\to \inf$.

When $2 < e < 3$, med is bounded, expected product rank is bounded but standard deviation of product rank $\to \inf$.

These are the three classic power law distributions.

Let us say, $m_i(t) = a(\frac{t}{i})^e (e > 1)$.

$$M = a \int_1^t (\frac{t}{i})^e \, di$$

$$= \frac{at^e}{e-1}(1 - t^{1-e})$$

(Now as $t \to \inf$, $t^{1-e} \to 0$.)

$$\approx \frac{at^e}{e-1}$$

Therefore $M = \frac{at^e}{e-1}$.

$$\int_1^{med} m_i(t) \, di = \frac{M}{2} = \frac{1}{2}\frac{at^e}{e-1}$$

$$\frac{at^e}{e-1}(1 - med^{1-e}) \approx (1/2)\frac{at^e}{e-1}$$

Canceling $\frac{at^e}{e-1}$ on both sides we get,

$$1 - med^{1-e} \approx 1/2$$

Therefore, we find that the expected product rank is inf but median is fixed when $1 < e < 2$.

Note on Zipfian distribution

If a new product is chosen with probability $1/t$ and an old product with probability $1-(1/t)$ and $e = 1$, the resulting distribution is called a Zipfian distribution. Lots of real life distributions have been found to follow the Zipfian distribution.