A Practical Model for Analyzing Long Tails

Kalevi Kilkki
Nokia Research Center
Helsinki, Finland
kalevi.kilkki@nokia.com

Abstract

This article provides a simple formula and a practical methodology for analyzing long tails. The same model can be applied to topics as diverse as books, search phrases, size of companies, and geographical distribution of populations. This article makes it possible for anyone to utilize the long tail concept not only as a general idea, but also as a tool to make realistic and useful analyses of real phenomena.

1. INTRODUCTION

Google “Long tail” and you get about 8,000,000 hits. Most of them refer to the concept utilized by Chris Anderson in his article[1] and book[2] about long tails. In essence, long tail refers to those numerous objects that have very limited popularity but that together form a significant share of the total volume[3]. The object could be a book, movie, piece of music, word, place, or almost anything else. Anderson’s book contains a lot of illustrative examples of long tails and eloquent writing without abstruse mathematical analysis – a viable formula for a best-seller[4] (in this article, “the long tail book” always refers to Anderson’s book). The approach, feasible as such, is problematic for those who want to make concrete inferences about long tails. The aim of this article is to make it possible for anyone to utilize the long tail concept not only as a general idea but also as a tool to make realistic and useful analysis of real phenomena.

A feasible analysis requires a mathematical model, enough raw data, and good understanding of the subject and the properties of the model. Thus not only is a formula necessary in this essay, in fact the selected formula forms the concrete basis for the whole analysis of long tails. Nevertheless, we have to

be aware of the fact that there cannot be any simple formula that is able to explain all the diverse phenomena that result in a long tail distribution.

Still, this essay is able to demonstrate that there is a lot of resemblance between many of those phenomena, which makes it possible to apply one analytical model. Further, the use of one formula makes it possible to systematically utilize the knowledge of the long tail phenomenon in general, particularly in cases in which there is very limited data available.

Thus we need a concise mathematical formula for analytical purposes. The formula applied in this essay to model all kinds of long tail is the following:

$$F(x) = \frac{\beta}{\left(\frac{N_{50}}{x}\right)^{\alpha} + 1}$$

Where $F(x)$ = the share of total volume covered by objects up to rank x

N_{50} = the number of objects that cover half of the whole volume

α = the factor that defines the form of the function

β = total volume

In this article, the above formula and its application to long tail data is called “the long tail model.” Of special note is that here we model primarily the cumulative distribution, not the “tail itself” (if we think of the tail itself consisting of the number of copies of objects in the rank order). The form of the real tail can be determined by the difference between two consecutive values of cumulative distribution ($F(x) – F(x-1)$). Thus, it is easy to calculate both all the cumulative values and the copies of an individual object. The only somewhat laborious task is to calculate the rank of an object if the number of copies for the object is known.

We can use a concrete example, book sales in 2004[5], to illustrate the basic characteristics of the formula. The sales are presented in Table 1 below.

Table 1. Book Sales in the US in 2004.

x = rank	Cumulative volume copies/year	Cumulative share based on real data	Cumulative share based on the model: F(x)
10	17396510	2.6%	2.7%
32	31194809	4.7%	4.6%
96	53447300	8.0%	7.8%
420	100379331	15.1%	15.1%
1187	152238166	22.9%	23.4%
24234	432238757	65.0%	65.1%
91242	581332371	87.4%	87.0%
294180	650880870	97.8%	103.7%
1242185	665227287	100.0%	118.7%

As with additional equations, the popularity of an article can be endangered by using complicated graphs. This poses a serious problem with exceedingly long tails. If we want to illustrate the popularity of all the books in a way that both the sales of the most popular book and the sales of the least popular book are discernible, we would need a graph with dimensions of 100x60 m. In this huge graph, the nearest point to the origin would be one meter from the origin, and the size of one book would be less than 0.1x0.1 mm. In a way, that huge graph would be extremely illustrative; we would get a concrete sense of the number of books and the number of copies. Figure 1 shows the graph on a sports field.

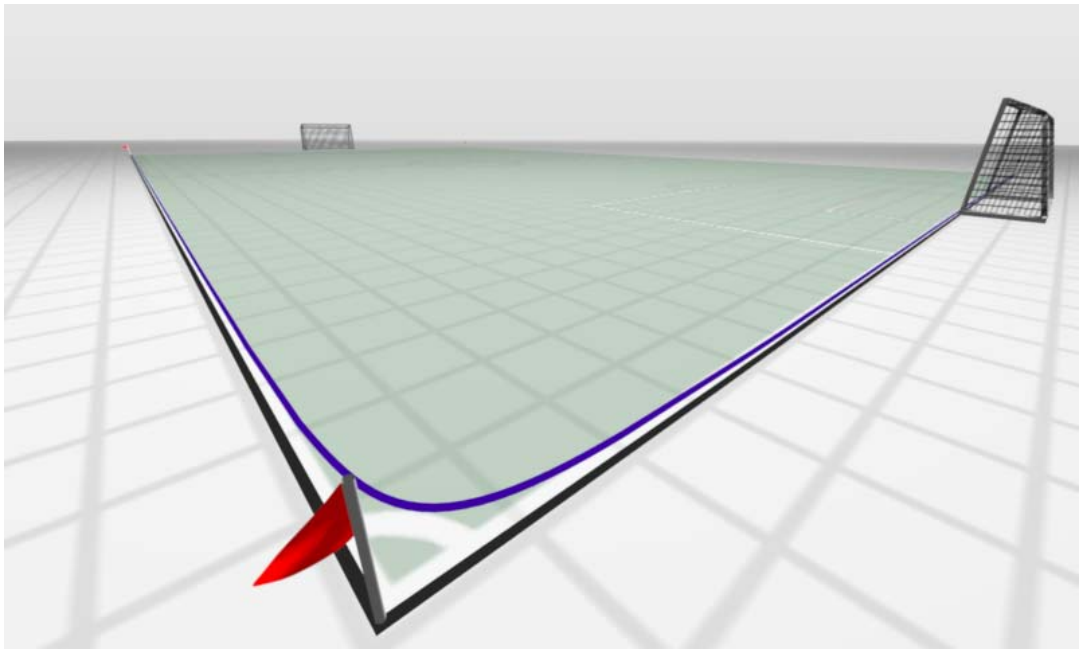

Figure 1. Books sales in the US as graph on a sports field.

Unfortunately, the illustration works only as a large graph, because graphed out on small paper gives us only two discernable lines, one on each axis. To solve this dilemma, we need to apply logarithmic scale. However, we also have to be aware of the fact that on a logarithmic scale the size of an area is not a direct indication of the real volume. To alleviate this problem we primarily use graphs that show cumulative volume rather than the volume of individual objects. The books sales statistics is illustrated in Figure 2.

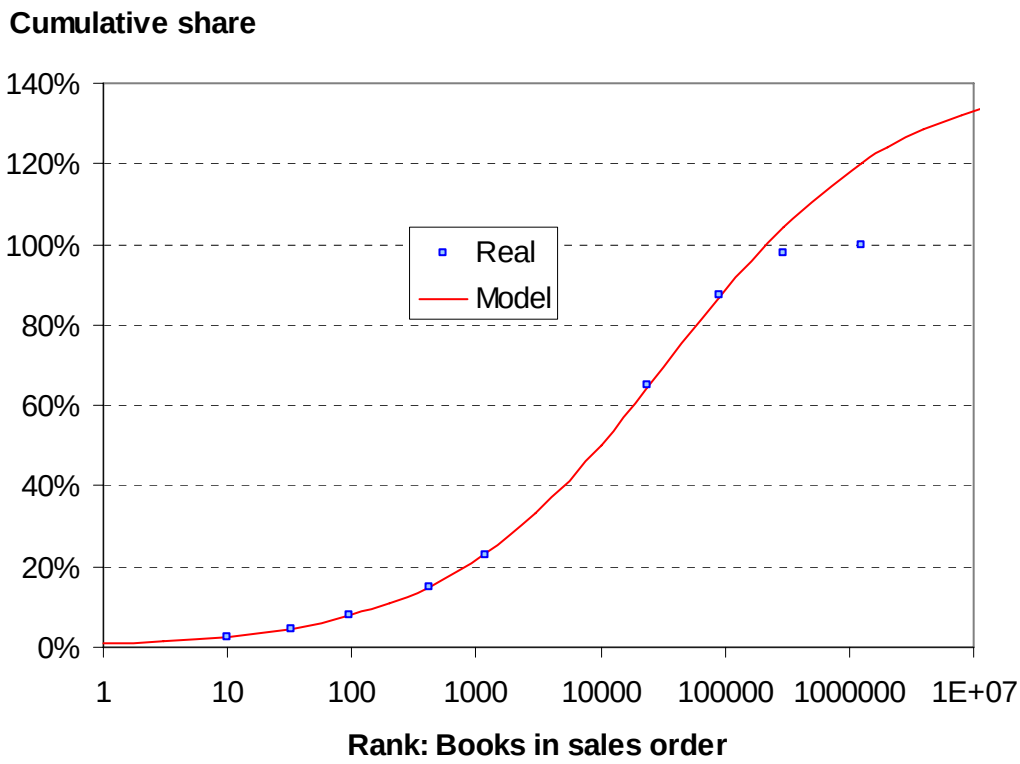

Figure 2. Cumulative share of book sales, 2004.

As to the practical use of the long tail model, two important comments have to be made. First, the last two rows in Table 1 were omitted when fitting the parameters N_{50}, α, and β. Secondly, the total share (β) is not limited to 100 percent. These choices may appear debatable. The justification of the approach is based on the assumption that there is a latent demand for objects with very low popularity that cannot be satisfied because of the cost structure of current business model. Thus we may presume that the last rows in Table 1 do not represent the "true" demand but that they are "artificially" suppressed by the current business model. Of course, someone may still prefer to model the real data with all data points instead of modeling the "assumed true demand" expressed only by a part of the whole data. However, here I have made a deliberate decision to primarily model the middle part of the tail. Then in those cases in which either of the ends does not comply with the model, I seek additional explanations for discrepancy. The benefits of this approach are:

- The fitting of parameters is often essentially better when the end of the tail is omitted.

- In most cases, the middle part of the function behaves very regularly and in a similar way even in very diverse cases.
- We automatically obtain an estimation of the latent demand.

Still we have to be aware that the concept of an idealistic situation is problematic and hard to define exactly. Moreover, the deliberate omission of available data always needs valid reasoning made separately in every individual case.

Now we can return to our example of books. We have seven data points and three free parameters in the long tail model. The best fitting between the seven data points and the model is found with parameters $N_{50} = 30714$, $\alpha = 0.49$, and $\beta = 1.38$.

What is the real meaning of these parameters? As to parameter N_{50}, the most popular 30714 books would represent 50 percent of sales in an ideal situation, which includes the latent demand for books with low popularity. However, we shall make a clear distinction between the real data and idealistic situation: according to the same model, the most popular 9640 books represent 50 percent of the total sales *in reality*.

Parameter β describes the amount of latent demand. The model indicates that the latent demand for very low volume books is 38 percent of the current total volume. However, this result is highly indicative. The latent demand could be somewhat smaller or essentially larger. Even if we assume that the long tail model is, in principle, valid to estimate the latent demand, we can find an acceptable fitting between real data and the long tail model even when β is as large as 2. It is difficult to apply any formal statistical method to assess the reliability of the result, because data points are highly correlated with each other. On the contrary, the essential question is whether the form of the function in the middle part of the tail, in the first place, can give any reliable estimation about the behavior of the end of the tail. This reliability question cannot be answered by assessing any individual case; it requires the careful evaluation of various diverse cases.

Despite these cautious remarks, several issues support the claim that there is a considerable latent demand. First, the fitting of the model is excellent up to rank 100,000. Further, the turn above rank 100,000 is quite abrupt and atypical for a long tail phenomenon. Finally, the idea of latent demand for books with very low popularity is credible without any mathematical model (see, e.g., the discussion in the long tail book). The estimate for the latent demand is so large that it obviously provides a feasible business opportunity.

The long tail of books has demonstrated the basic structure of long tail analysis. First, we gather relevant data about the phenomenon. Secondly, we apply the long tail model to the available data. If there is a clear difference between the data and the model, we seek a realistic explanation for the difference. Finally, we use the model to make conclusions about the business potential for different players.

For people who are just generally interested in long tails, there is no compelling need for a deeper understanding of the theoretical part of the model. In contrast, for those who want to apply the model in real cases, a somewhat deeper understanding is certainly useful. Annex 1 provides an introduction to the long tail model and some guidance for applying the model in real cases and interpreting the results. The main part of this article consists of examples that illustrate the capabilities of the long tail model.

2. MORE ABOUT WRITINGS

Let us first apply the long tail model to Amazon's online book sales. The problem with Amazon, from the viewpoint of a formal analysis, is that Amazon reveals only limited information about its book sales. Basically, they give a rank for each book, but they do not tell how many copies are sold. Moreover, the rank is based on an unknown algorithm that takes into account both the most recent sales and the long-term sales. Therefore, an outsider is able to make only rough estimates about the sales[6]. The strength of the long tail model is in the opportunity of using also the knowledge about the real book sales and combining that information with the estimations of Amazon's sales. Thus as a starting point we have the model parameters for real sales: $N_{50} = 30714$, $\alpha = 0.49$, and $\beta = 1.38$.

There are two major matters that separate real book stores and online sales. Firstly, the online model is efficient even for books that sell only a couple of copies per year. Secondly, the online model can also cover second-hand books. Thus we may safely assume that N_{50} is larger and β is smaller with online stores than with ordinary book stores. Further, because N_{50} is larger, α might be somewhat smaller, though the difference is likely small.

Then if we combine estimates for Amazon sales with real sales models, we can assume the following parameters depicting Amazon's sales: N_{50} = 40 000, α = 0.48, and β = 1.2. Using these parameters, we obtain the results presented in Table 2 below.

Table 2. Estimation of Amazon's book sales

Books with rank	Share of sales
1 – 10	2.2%
11 – 100	4.2%
101 – 1 000	11%
1 001 – 10 000	23%
10 001 – 100 000	32%
> 100 000	27%

Because α is relatively small, there could be a separate business opportunity both for books with high popularity and for books with very low popularity (the effect of parameter α on business potential is discussed in Annex 1). A small selection of best sellers is a feasible approach in airport shops, for instance. Similarly, the business opportunity for those books that are not covered even by Amazon (20 percent as β is 1.2) represents a considerable opportunity for self publishing. However, it will be difficult for any player from those two special sectors to challenge players with a strong grip on the middle of the tail.

It could be anticipated that the popularity of magazines is similar to the popularity of books, although there are many more books than magazines. Here we use the circulation statistics provided by Audit

Bureau of Circulation[7]. The data can fitted to the long tail model by using parameters N_{50} = 82 and α = 0.72. In addition, parameter β is selected in a way that 200 magazines represent 66 percent of total circulation. The data and the long tail model are shown in Figure 3.

The correlation between reality and the model is so good that there is no need to explain any differences. Still, because the data is limited to 200 magazines, and we do not have data about the total circulation, we cannot be sure how well the model is able to predict the circulation of smaller magazines. In reality there might be a significant latent demand to be satisfied with web-based publications (part of the demand might already be satisfied).

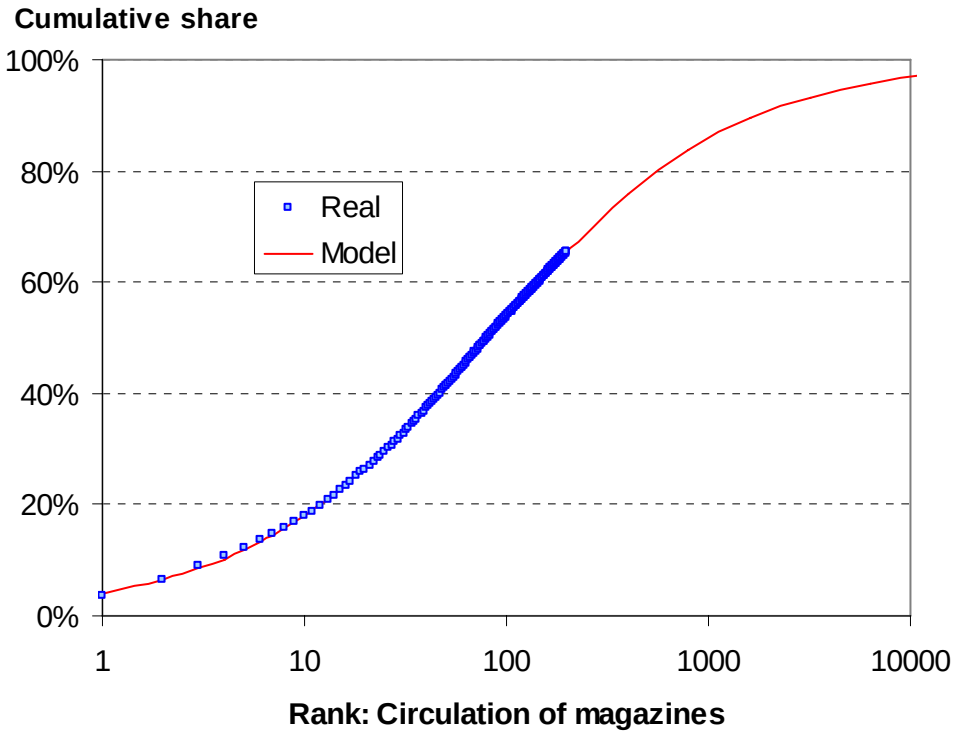

Figure 3. Newspapers by Largest Reported Circulation.

3. MOVIES

Let us next move to movies. Movies offer a possibility for an extensive analysis, because there are a lot of public statistics available over different periods (week, year, all time), in different markets (world, USA, smaller country), and based on different business models (movie theaters, rented movies, DVD sales). However, in most cases the statistics are limited to the most popular movies. Thus let us first take a case

in which we have complete statistics: All the movies presented in movie theaters in Finland in 2003[8]. In total, 223 different movies were presented with a total number of viewers of 7,632,717. The total audience ranged from 12 (!) to 614,097. Figure 4 shows the number of viewers per movie and the corresponding long tail model with parameters N_{50} = 26, α = 0.85, and β = 1.38.

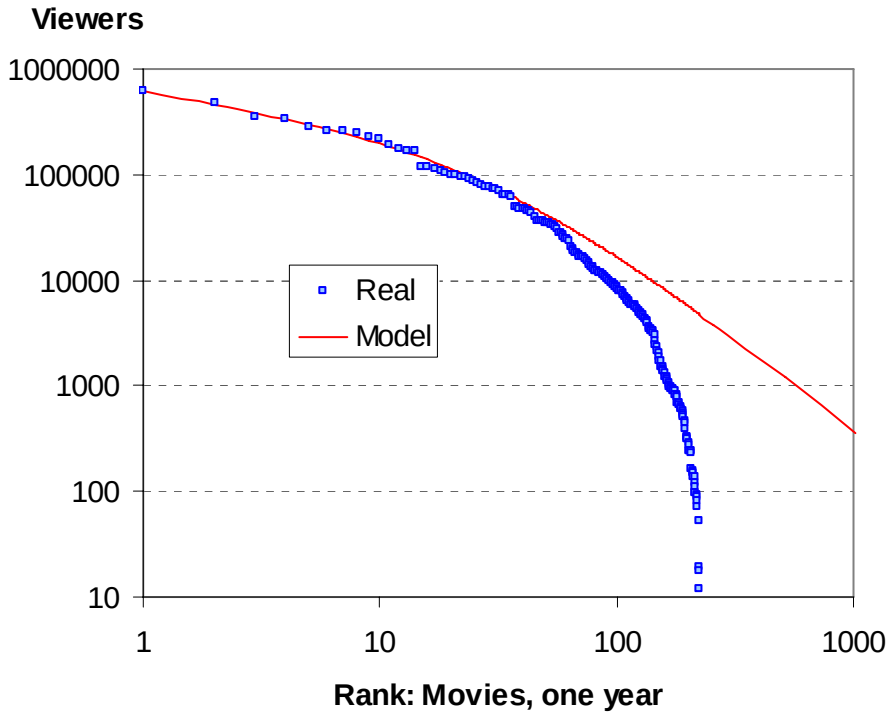

Figure 4. Movies presented in movie theaters in Finland, 2003.

The fitting of the model is excellent from rank 1 to rank 36; the model even describes almost exactly the number of viewers for the most popular movie. But when the rank exceeds 36, the reality and the model start to diverge. The same behavior is illustrated on page 128 of the long tail book for Hollywood movies. Anderson uses the figure to illustrate the latent demand for those films that are not presented in local movie theaters. He does not, however, give any estimation about the size of the latent demand.

Now we can fit the figures shown in the long tail book to the long tail model. The result is N_{50} = 56, α = 0.82, and β = 1.60. In this case our model describes the curve correctly up to rank of 70. Because the market in Finland is smaller and more homogeneous than in the US, it is quite understandable that parameter N_{50} and the total number of movies are smaller in Finland. What is more interesting is that the

estimate for β is clearly larger in the US. If we rely on the long tail model, this result indicates that the huge US movie market is suppressed more strongly than the small market in Finland. Note that in Finland the majority of the movies presented in movie theaters are the same as in the US, except that there are some Finnish movies (22 percent of sales in 2003) that usually have a limited audience outside Finland.

We may speculate that the huge marketing budget required to get market attention is one of the key factors that artificially limits the offerings in movie theaters. Moreover, it might be that the US market does not properly satisfy the needs of minorities. Surely there are other relevant aspects to explain the abrupt drop in viewers above a certain rank, as discussed by Chris Anderson.

Finally, let us apply the long tail model to the all time box office sales in the US[9] shown in Figure 5. The result is N_{50} = 1715, and α = 0.77. In addition, the fitting is done in a way that the movies up to rank 1000 represent 40 percent of total sales. Once again, the fitting is extremely good except that we are still waiting for the truly best movie; according to the long tail model, the most popular movie should have sales of $805 million instead of $600 million, which is the box office sales of Titanic. Perhaps some new movie will solve this discrepancy. As to the other end of the tail, we shall be somewhat cautious because the short-term statistics imply that the end of the tail could be relatively short. Hence, the real curve likely takes a sharper turn somewhere between ranks 2000 and 5000 than what the long tail model predicts.

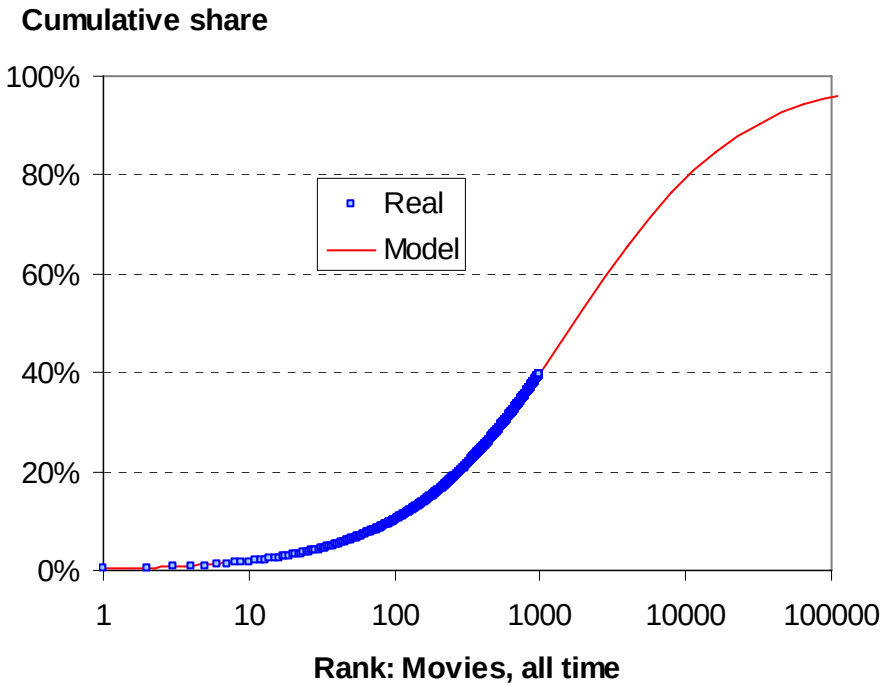

Figure 5. Box office sales in the US.

Parameter α seems to be relatively large with movies, typically around 0.80. In addition, the real length of the tail is clearly shorter than the tail predicted by the model (in which the popularity of a movie is not limited due to marketing budget and other external issues). Therefore, now when the cost of producing and selling a movie on DVD is diminishing, there is a considerable business opportunity for all those movies that do not fit with the movie theater business model.

4. MUSIC

Another important sector of culture is music. It also offers a good opportunity for long tail analysis. In this essay we use the information from last.fm[10]. Let us first take the artist ranking and look how that distribution fits to the long tail model. The real data and the long model (N_{50} = 550, α = 0.84) are shown in Figure 6. According to the model, the most popular 400 artists represent 43 percent of the whole volume. The largest difference between reality and the model is, quite typically, for the most popular artist (12 percent), while for all other artists the difference is less than 7 percent. However, the available data includes only 400 artists, which does not allow robust conclusions about characteristics of the whole tail.

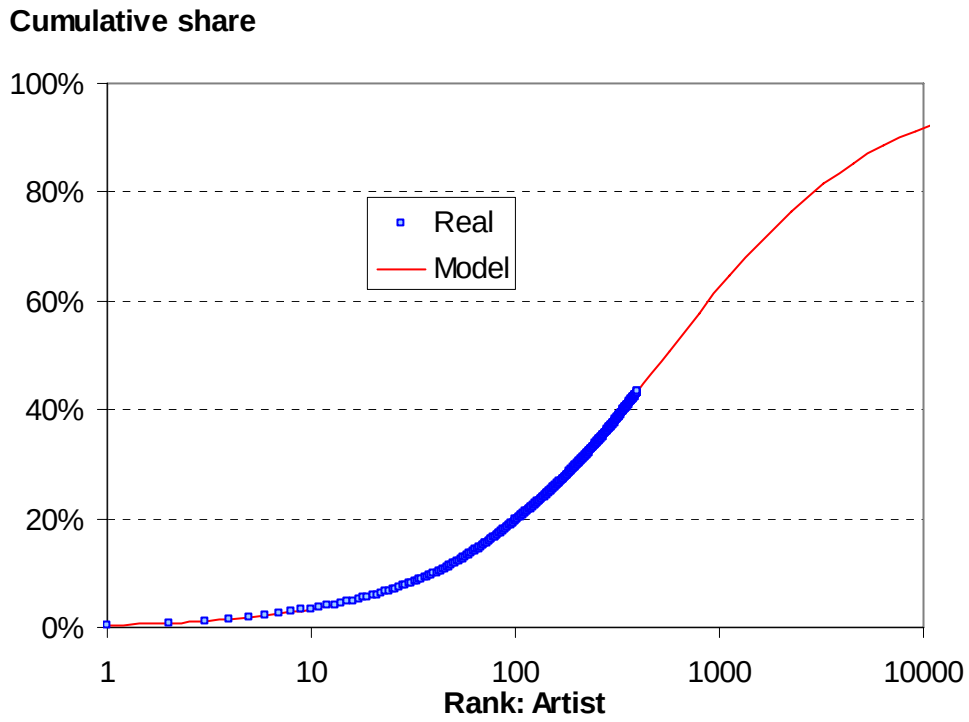

Figure 6. Top Artists for the week ending 3 Sep 2006 at last.fm, column "Reach"[11].

Next we can take another dimension of classifying music, namely tags. Tags are chosen by users to depict artists, albums, and tracks. In this case the model gives the following parameters: $N_{50} = 55$, $\alpha = 0.70$, and the most popular 100 tags represent 60 percent of the whole volume of tags as illustrated in Figure 7. In this case the long tail model is able to predict, not only the middle of the tail, but also the popularity of tags with the highest rank. It seems that the fitting is the best for those cases in which the availability of different objects is not limited due to a fixed cost.

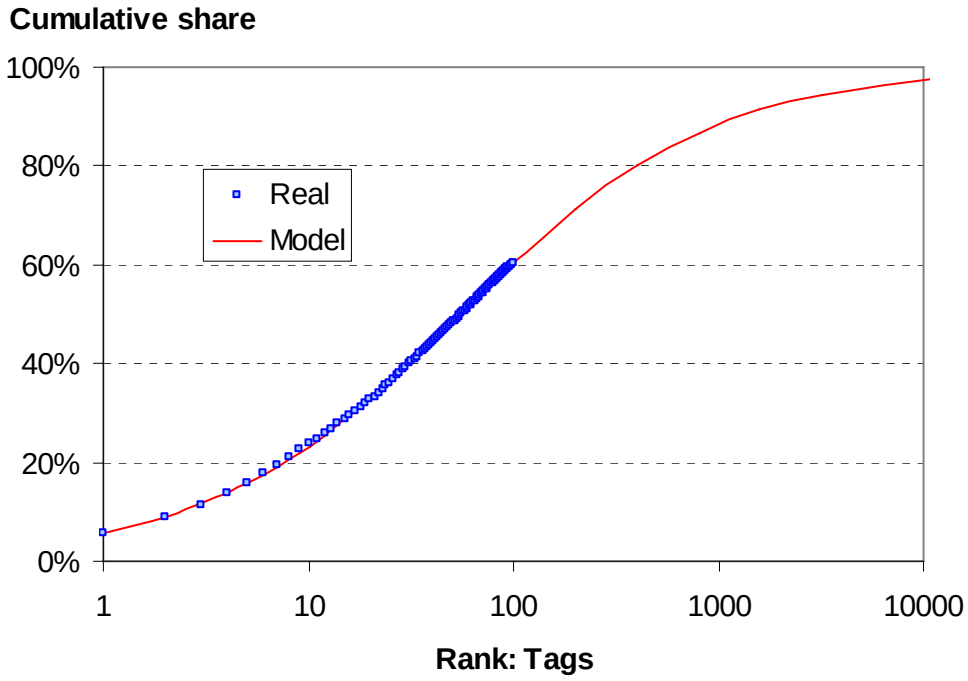

Figure 7. Most Popular Tags at last.fm[12].

The tag distribution is interesting from the viewpoint of commercial radio-stations. Because N_{50} is small and the value for parameter α is moderate, the most popular tag ("rock") covers as much as 5.5 percent of the total volume. In contrast, although tags with a rank above 100 seem to represent together 40 percent of the total volume, the popularity of each of those tags is less than 3 percent of the interest of rock. That is hardly a real business opportunity for a conventional radio station. In order to utilize the end of the tail, the business model has to somehow combine a significant amount of tags with low popularity. This is a common dilemma for the business based on the end of the tail. In the case of radio stations, the solution could be something like a virtual station, in which the listeners create their own play lists with the help of the virtual radio station.

5. WORDS, STRINGS AND PHRASES

Books, movies and music are typical examples of objects produced by professionals usually with the explicit purpose of earning income. Next we consider some issues with which the popularity emerges in a different manner. Wiktionary provides interesting information about the frequency of words[13]. The most common words in TV and in movie scripts are shown in Figure 8. The parameters of the long tail

model are N_{50} = 63, α = 0.68, β = 1.00. Indeed, only 63 words (you, I, to, … , one) are needed to make up half of everything said on TV.

In this case the model works pretty well through the whole tail – the only thing needing an explanation is the lack of discrepancy. We may argue that this just is an example of a case in which there is no obvious cost of using a rare object, except possible misunderstanding, and thus a natural form for long tail ensues. Actually, the biggest error of the long tail model is that it predicts a larger difference between the two first words. Thus, in theory, *you* should be more popular and *I* should be less popular than what they are in reality.

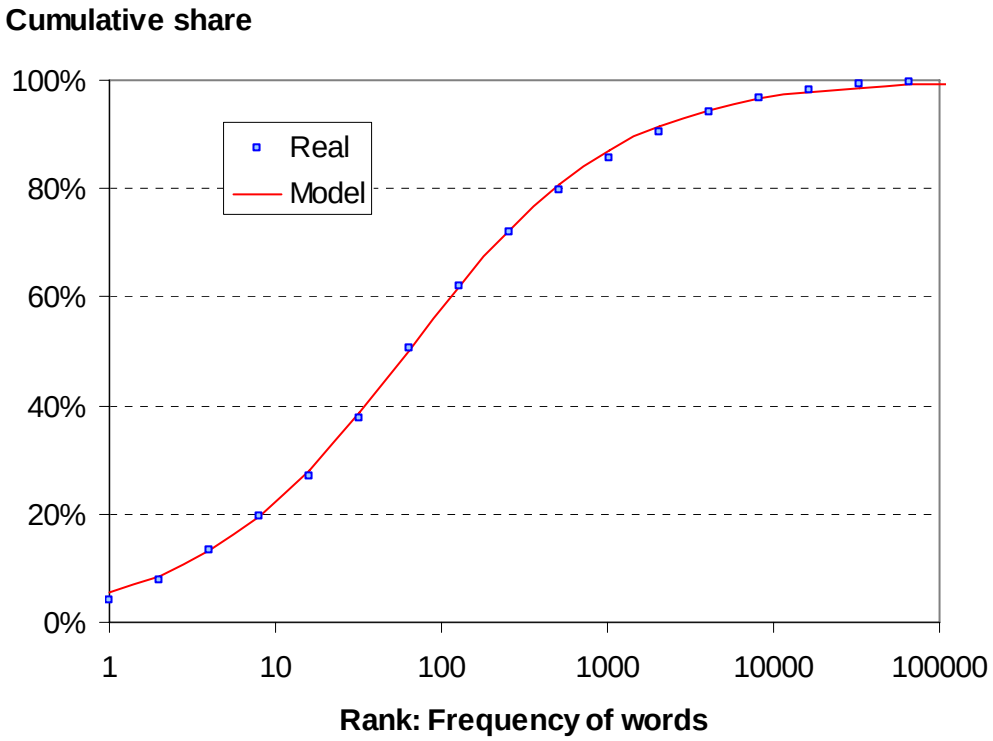

Figure 8. Most common words (TV and movie scripts).

An attractive way to assess the popularity of anything is to use a search engine. There are basically two possibilities: first, to measure how often a certain string is searched, and secondly to rely on the number of hits given by search engines. The first approach likely gives more useful information, but also is more difficult to gather. Richard Wiggins has provided some concise, but useful, data about the popularity of

search phrases in a university environment[14]. The statistics are shown in Figure 9 together with a long tail approximation with parameters N_{50} = 434, α = 0.53, β = 1.00.

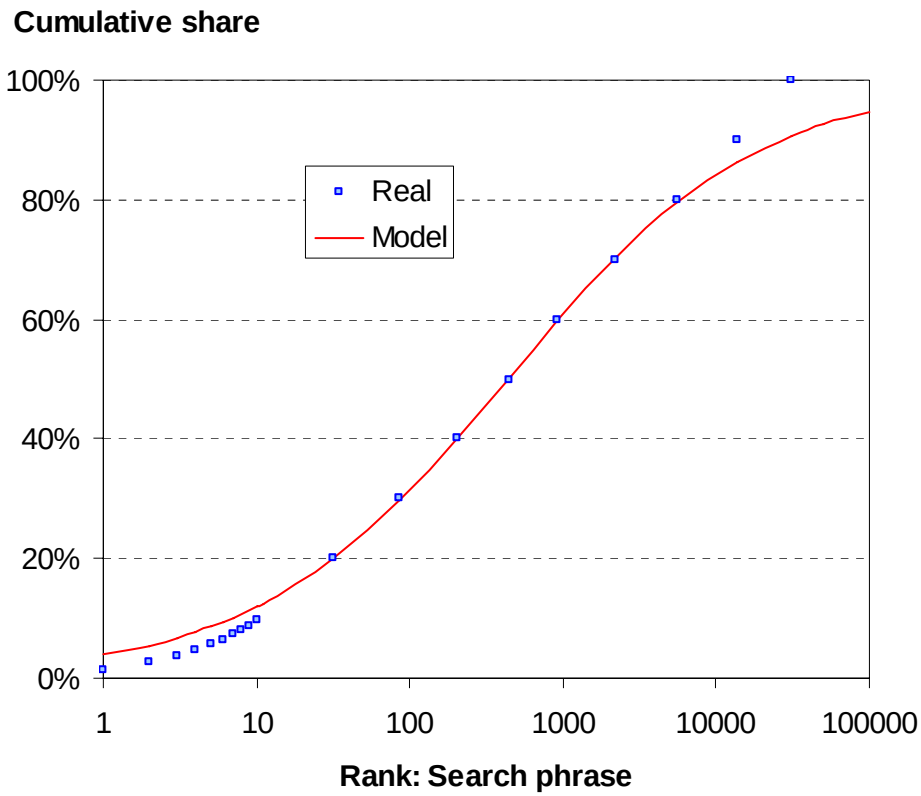

Figure 9. Unique search phrases at Michigan State University.

There are two anomalies to be explained: One at the base of the tail and another at the end of the tail. The first one is more difficult to explain: Why don't the 10 first phrases comply with the middle part of the tail? It should be noted, however, that the base of the tail could be changed significantly even by one very popular phrase. Thus we may speculate that the difference is just a coincidence caused by the randomness of the phenomenon without any specific reason. However, because the "missing phrase" should be about four times more popular than the current number one, we need to admit that in this case the long tail model is unable to accurately describe the base of the tail.

In contrast, the other anomaly at the end of the tail has a credible explanation. Almost half of the search phrases are used only once. Thus those phrases (together with some of the phrases that have been searched only a couple of times) represent objects with much lower true frequency. The sampling

process just randomly picks up some of the objects from the vast quantity of possible phrases that form the long end of the tail. If we take this phenomenon into account, the real curve and the model match almost perfectly with each other.

Unfortunately, relevant information about the popularity of search phrases is rarely available. The other possibility is to rely on the number of hits declared by search engines. But does Google or some other search engine provide a feasible and reliable way to assess the real popularity of anything? What a search engine obviously does is that it gives an estimate about the commonness of a string of characters on all available web pages.

Now we can utilize the long tail model by conducting tests with different strings by checking how well the real results comply with the model. Because it is impossible to test all possible phrases or strings, we need to limit the search somehow. Here we concentrate on character strings containing only Latin alphabets. Figure 10 shows the strings with over 1 billion hits according to Google (I have probably missed some of them), and the corresponding long tail model with parameters $N_{50} = 993$, $\alpha = 0.67$.

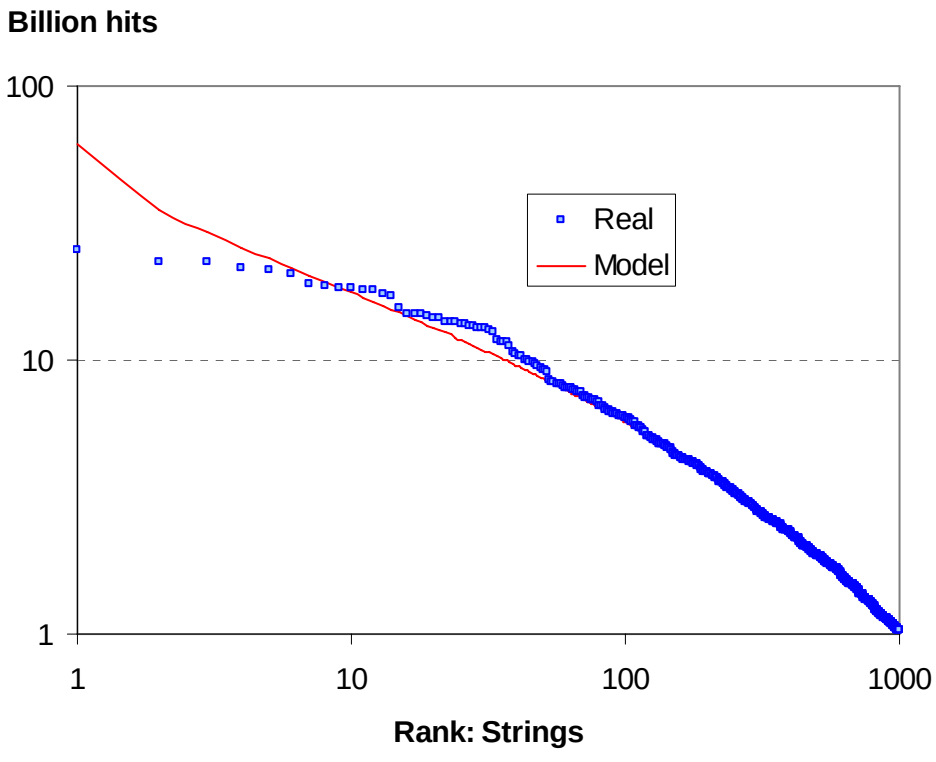

Figure 10. The most popular strings according to Google[15].

The fitting was done for the range from 50 and 1000, because the long tail model cannot appropriately describe the popularity of strings that produce more than 9 billion hits. In contrast, the number of hits seems to provide a consistent measure of popularity in the region from one billion to nine billion hits. Other search engines may behave somewhat differently. For instance, MSN Search gives a systematically smaller number of hits and appears to comply better with the long tail model up to the highest ranks.

The other end of the tail is somewhat trickier to analyze. One possibility is to generate random strings and observe how the results behave compared to the long tail model. The simple scheme adopted here was to generate 250 random numbers between 100,000 and 1 million and to put them in three different search engines (Google[16], Yahoo[17] and MSN Search[18]). Then those 250 numbers were put in the order of hits. The results are presented in Figure 11.

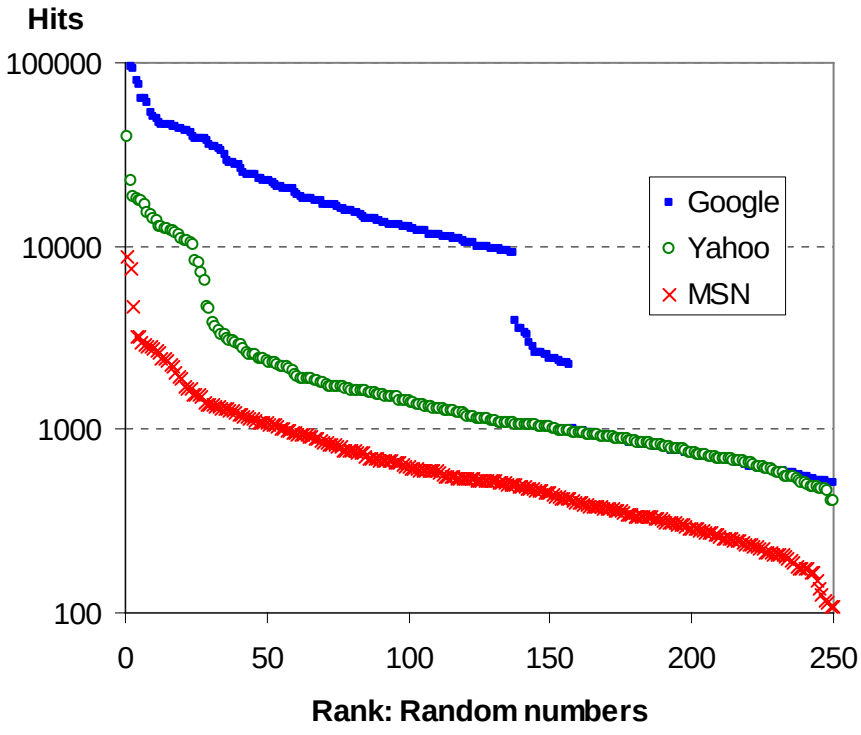

Figure 11. Two hundred and fifty random numbers between 100,000 and 1 million; vertical axis: results by Google, Yahoo and MSN Search.

With Google there are two obvious discontinuities, first between 1000 and 2100 hits, and secondly between 4000 and 9000 hits. We also conducted a similar experiment with random words with four letters. The result was similar but not exactly the same: The caps were smaller and the difference between “above 10000” and “below 1000” was smaller. Yahoo indicates similar but not as pronounced behavior in the region between 3000 and 10000 hits. The behavior of MSN Search is in this respect smoother.

It is hard to imagine any reason why there should be any discontinuity in the number of hits. Thus it is likely that the steps in Figure 11 are created by the algorithms used by search engines. Based on our randomized experiments, it seems that some algorithms work differently above 10000 and below 1000. Below 1000 the number of hits probably gives a realistic estimation of real pages to be found in the Internet. Above 10 000 hits the results do not necessarily have as concrete meaning. Yet, based on experiments with random strings with various lengths, the number of hits seems to behave quite smoothly between 10,000 and nine billion hits.

For instance, if the number of hits for objects A and B are 1 million and 50 million, respectively, that seems to justify the statement that object A is fifty times more popular than object B (in some respect). On the contrary, the situation is vaguer if the number of hits for objects C and D are 500 and 25000, respectively. In that case object C might be eight times more popular than object D, in the case of Google, whereas the ratio might be something else with another search engine. As a rough estimate, the following parameters could be used to assess when the number of hits is above 10000:

- Google: $N_{50} = 993$, $\alpha = 0.67$, total volume = 6170 billion hits.
- Yahoo: $N_{50} = 720$, $\alpha = 0.74$, total volume = 2650 billion hits.
- MSN Search: $N_{50} = 530$, $\alpha = 0.68$, total volume = 470 billion hits.

These estimations are primarily valid for English pages and words, whereas the models may somewhat underestimate the share of pages written in other languages.

Another problem with search engines lies in the systematic bias of the results. This phenomenon is illustrated in Table 3 below.

Table 3. Rank of words according to Google.

Word	Rank	Word	Rank	Comment
www	1	the	2	www is The abbreviation.
search	15	with	16	Search is The verb.
information	32	an	33	Information is The term.
copyright	38	news	39	More copyright than news - this is really sad news.
policy	56	do	58	What does policy do?
internet	94	world	98	Internet is bigger than world.
technology	110	over	111	If technology is over over, what is over?
php	123	people	124	What on earth is this PHP that is more popular than people?
documentation	563	friend	563	Documentation might be important, but is it like a friend?
portal	615	love	615	Portal is as important as love - does anyone agree?
webmaster	662	value	662	But who is the webmaster?
electronics	985	error	985	Electronics is as common as error - does anyone disagree?

Based on these examples, it is apparent that terms related to web technology are highly overrated compared to ordinary words. The presumable reason is that the algorithm used to estimate the number of hits gives a lot of weight to those strings that appear on the main page of a web portal, particularly on the links to other pages. Therefore, the number of hits might be appropriate for comparing the popularity of terms related to the same topic, whereas it is quite questionable to compare terms related to totally different topics. If we take the 1000 most popular words on TV and the 1000 most popular terms according to Google, only about 43 percent of them are the same. The general context defines which of the lists gives a better estimation about the real popularity of a term.

As a summary, search engines could be used to assess the popularity of a term, as long as the following issues are taken into account:

- The terms used on the main page of Internet sites obtain a disproportionately large number of hits. Illustrative examples are the abbreviations php, rss, and htm.
- The algorithm used by a search engine may produce somewhat strange behavior, as illustrated in Figure 11. The results seem to be relatively consistent both below 1000 hits and in the region from 10,000 to 9 billion hits.
- There are quite a lot of variations from day to day, or even during a shorter period. For a smaller number of hits, the results seem to be more stable.
- The results from different search engines are not directly comparable.

6. NAMES

One obvious candidate for a long tail is names. Names also provide a practical topic because extensive statistics are available from different sources. Here we use the data from the U.S. Census Bureau[19]. Figure 12 shows the distribution of surnames up to rank of 88799. The parameters of the long tail model are N_{50} = 1711, α = 0.52, and β = 1.00.

The main differences between the reality and the model occur, once again, at the ends of the tail. At the base of the tail, the most common surname (Smith) should be, according to the long tail model, about 100 percent more common than what it is in reality. From the second name (Johnson) up to the rank of 40000, the model is accurate. Then the very end of the tail seems to be thinner than what the model predicts. This difference may stem from the fact that the unit for a surname is not an individual person, but rather a family or kindred. Thus the difference might be explained by using a similar reasoning as with search phrases.

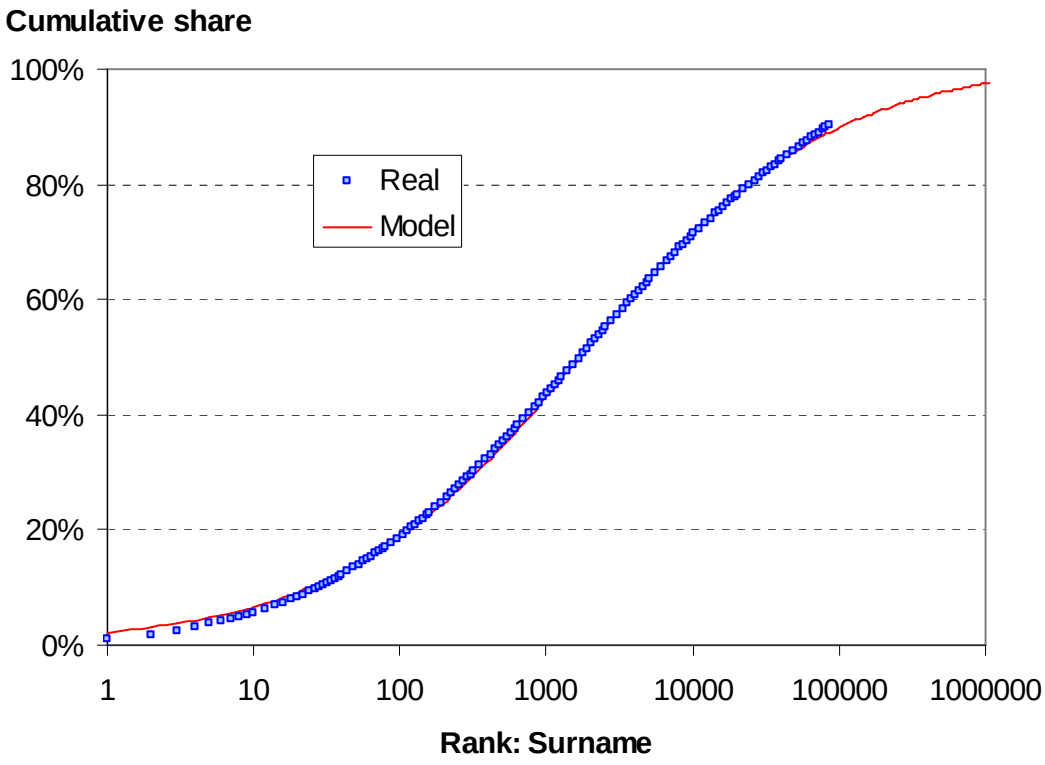

Figure 12. Most common surnames in the United States.

We can also look at first names, for instance, female first names in the US. Mary is by far the most common female name with a popularity of 2.6 percent[20]. The difference between Mary and the second most popular (Patricia, 1.1 percent) is so large that our model clearly underestimates the difference. But from the second name up to the rank of 600 the long tail model works properly with parameters N_{50} = 137, α = 0.83, and β = 1, as shown in Figure 13. The dissimilarity at the end of the tail is so significant that it certainly calls for a credible explanation. So why are the names above the rank of 4275 used 87 percent more often than what the long tail model predicts? A possible answer is that the first half of the tail (up to the rank of 139) mostly consists of English names, while at the end of the tail a significant number of the names originate from other cultures. Those foreign names may considerably lengthen the tail of names.

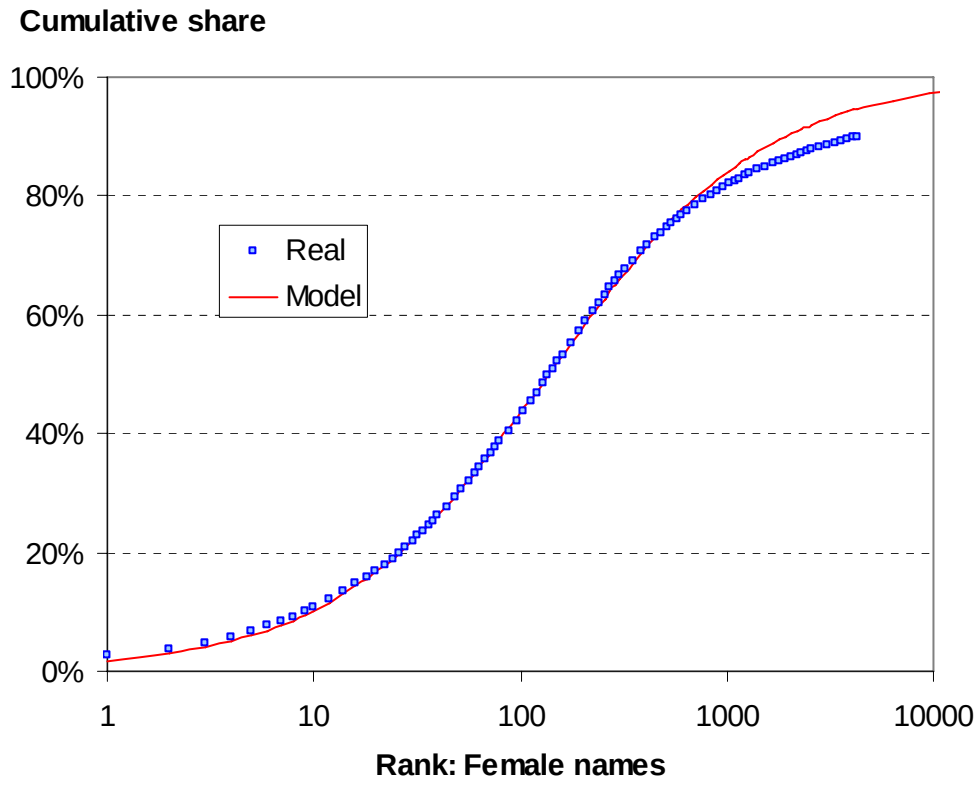

Figure 13. The most common female first names in the United States.

A somewhat unexpected result is the clear difference between surnames, male names, and female names. Parameter α is low for surnames (0.52), moderate for male names (0.75) and large for female names (0.82). Moreover, the ends of the tail seem to behave differently in the case of surnames (shorter than expected) and first names (longer than expected). An interesting research question would be to study whether these differences are systematic among different countries and cultures.

7. BUSINESS

The total revenue is an evident indication of the popularity of the products of a company. Thus it is quite natural that the same long tail model used to assess the popularity of other objects is applicable also to the size of companies. Figure 14 shows the size of companies according to the data provided by CNNMoney.com[21]. The parameters in the long tail model are $N_{50} = 482$ and $\alpha = 0.71$.

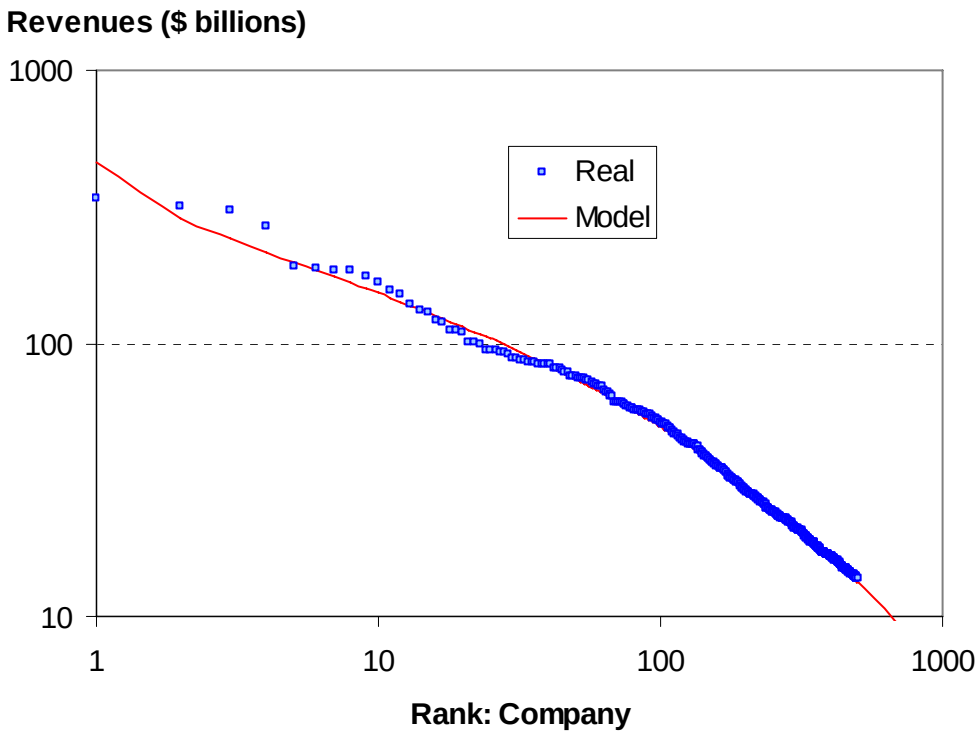

Figure 14. The revenues of the 500 largest companies in 2005.

A couple of remarks can be made. First, although the overall fitting is good, the model cannot accurately predict the revenues of the largest companies. Particularly, the largest company should be 36 percent larger, whereas the companies with rank 3 and 4 should be clearly smaller. We may even argue that in 2005 the four largest companies were artificially similar in size. According to the long tail model, the largest company should be about 61 percent larger than the second largest company. Secondly, it should be stressed that the available statistics are too limited to accurately estimate the end of the tail. Even though the best fitting indicates that the 500 companies represent 51 percent of the total volume, that share could be anything between 40 percent and 60 percent. Respectively, N_{50} may vary between 300 and 1000.

Similarly, in the case of the 500 biggest companies in Finland a good fitting is obtained if we assume that $N_{50} = 37$, $\alpha = 0.63$, and the largest 500 companies represent 83 percent of the total revenue. In this case the long tail model works well even with the largest company, Nokia, as shown in Figure 15. Hence we may argue that the size of Nokia was natural in 2004 compared to other companies and to the Finnish

economy. In addition, the revenues of companies between ranks 3 and 9 seem to be somewhat larger than expected, whereas the long tail model is accurate with smaller companies.

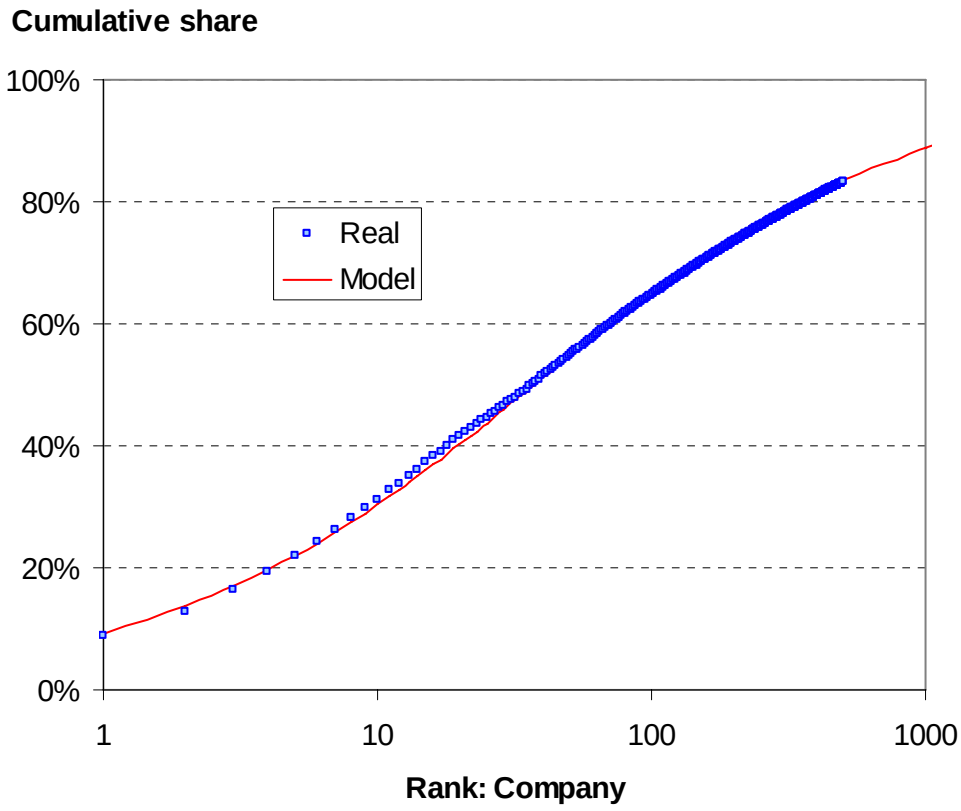

Figure 15. The revenues of the 500 largest companies in Finland, 2004[22].

8. POPULATION

As the last topic, let us consider a matter that seems to be quite different compared to the popularity of books, music and movies: geographical distribution of populations. Yet, your place of residence is a matter of preference, if anything. Moreover, we may speculate that similar processes are used when people choose tangible objects and how they select their place of residence. Figure 16 shows how people are distributed in Finland over different square kilometers.

In the densest square kilometer, there are about 20,000 inhabitants, while there are no permanent residents in about 60 percent of the square kilometers. An excellent fitting from rank 1 to rank 1300 can be found by parameters N_{50} = 1314 and α = 0.78. The fitting is somewhat worse between ranks 1300

and 20000. In reality there are about 35 percent more square kilometers in which the number of inhabitants is between 20 and 50 than what the long tail model predicts. Whether Finland is a special case in this respect needs further studies. Then, as to the end of the curve, the difference is once again easy to explain by the fact that there is no square kilometer with a fraction of inhabitants.

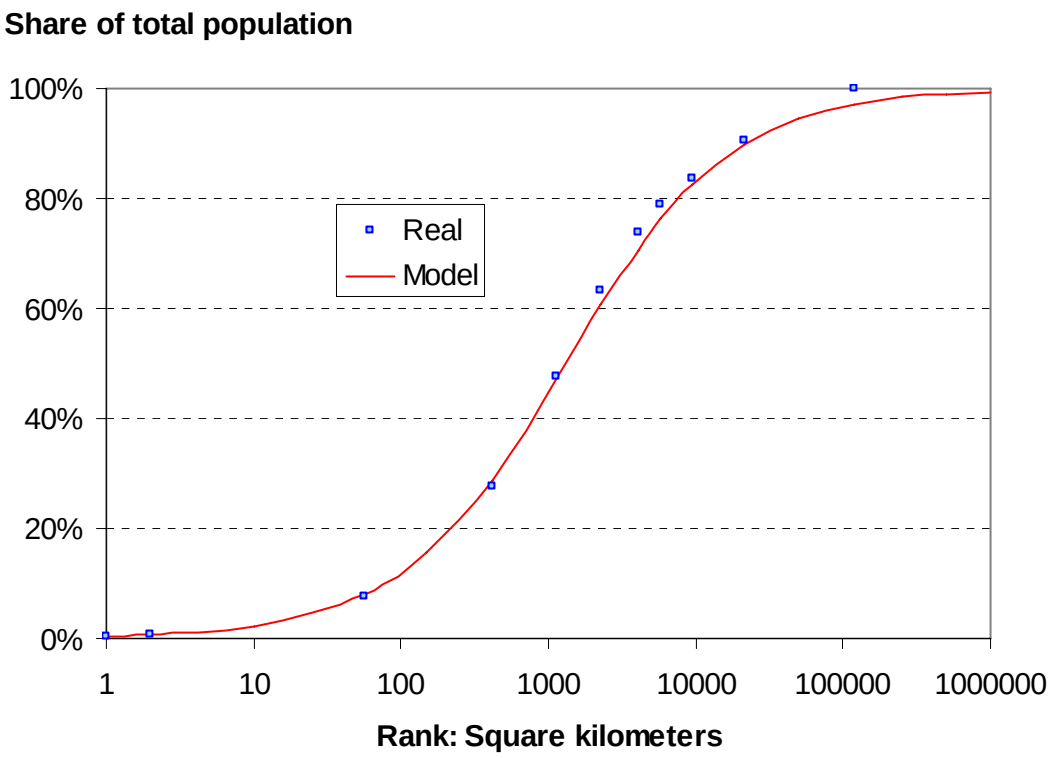

Figure 16. Square kilometers in Finland in order of population density[23].

9. CONCLUSIONS

This essay offers a dozen of examples of phenomenon, from books to square kilometers, that manifest themselves with a long tail of popularity. The long tail distributions are so similar that there is an obvious opportunity to model them by a single function. The main requirement for the function is that the cumulative distribution should generate a smooth S-shape when the x-axis is logarithmic.

An S-curve can be described basically by two parameters, one to define the turning point, and another to define the steepness of the curve. In our model the corresponding parameters are N_{50} (turning point) and α (steepness). Long tails differ in both respects. The really long tail of books can be modeled by parameters $N_{50} = 40000$ and $\alpha = 0.48$, whereas the much shorter tail of movies viewed during a year can

be modeled by parameters N_{50} = 25 and α = 0.85. The tail is naturally much longer for all time box offices sales than for sales during one year; still the form of the function (that is, parameter α) is relatively constant. More extensive studies are required to assess whether a specific type of object produces always a similar long tail, or whether the result varies among different countries and cultures.

As to the accuracy of the model, in many cases there are discrepancies that call for explanations. First, some anomalies could be explained by pure random variations, particularly with the objects with the highest ranks. Secondly, the abrupt end of the tail often is caused by the fact that in reality the size of the object is finite (e.g., one book), while the long tail function continues to eternity with ever smaller objects. Thirdly, the current environment may artificially shorten the tail. For instance, the business model of movie theaters significantly favors the most popular movies compared to an ideal distribution channel that can effectively distribute movies with a small audience. Fourthly, the effect of minorities (e.g. languages other than English) may considerably lengthen the end of the tail but are invisible in the base of the tail. Finally, in some cases there is no apparent explanation for the difference. To explain those unclear cases, we need more studies and better understanding. Nevertheless, the long tail model introduced in this article forms a robust basis both for understanding and analyzing diverse long tail phenomena.

Annex 1. Some notes on the application of the long tail model

This annex provides additional information about the selected long tail function and how to apply it in real cases. The long tail function applied in this article,

$$F(x) = \frac{\beta}{\left(\frac{N_{50}}{x}\right)^{\alpha} + 1}$$

is one of the numerous possibilities to model long tails. It has several advantages, like the limited number of parameters, simple mathematical operations, and easiness of use. Particularly, if we want to calculate the frequency for a given rank, we only need to calculate the difference between $F(x)$ and $F(x-1)$:

$$f(x) = F(x) - F(x-1)$$

Now an observant reader may notice that *f*(1) is a special case, because the cumulative function is undefined for x = 0. How should we handle this problem? The approach adopted here is to assume that the volume of the most popular object is just the value of the cumulative function at point x = 1[24]. As a result, the volume for the most popular object is essentially larger than that of the second object, when both N_{50} and α are small. Still, as the cases in this essay clearly demonstrate, quite often the model works pretty well even for the object with the highest rank.

The cumulative distribution can be divided into three parts: the base of the tail, the middle of the tail, and the end of the tail. In this essay we use the following formal definitions:

The boundary between the base and the middle of the tail is $x_{bm} = N_{50}^{2/3}$
The boundary between the middle and the end of the tail is $x_{me} = x_{bm}^{2} = N_{50}^{4/3}$

For instance, if N_{50} = 1000, the base of the tail covers objects with a rank up to 100, the middle covers objects with a rank from 101 to 10,000, and the end covers objects with ranks over 10000. On a logarithmic scale, the cumulative function is divided into three sections of similar size.

While parameters N_{50} and β have relatively concrete meaning, parameter α is quite abstract. Firstly, it should be noted that α shall not exceed 1, because larger values pose serious problems with the smallest ranks. In all the cases evaluated during the study, parameter α varies between 0.45 and 0.95, in a way that small values of α are more likely when N_{50} is large.

What happens when N_{50} is kept constant but the value of α is changed? The situation is illustrated in Figures A1, A2 and A3 when N_{50} = 512. Figure A1 shows the cumulative volume as a function of rank. In Figure A1 parameter α defines the steepness of slope in the middle part of the function. Figure A1 also presents the cumulative distribution for a power-law function in which:

$$f(x) = ax^{k}$$

Parameter *k* is selected in a way that half of the total volume is covered by objects up to the rank of 512. In all the cases evaluated during the study, the results fall between the two extreme cases (α = 0.45 and α = 0.95). Therefore, it is apparent that a power-law function is not suitable for modeling long tails. However, it should be noted that the long tail formula used in this article approaches a power law function when the rank (*x*) is large enough.

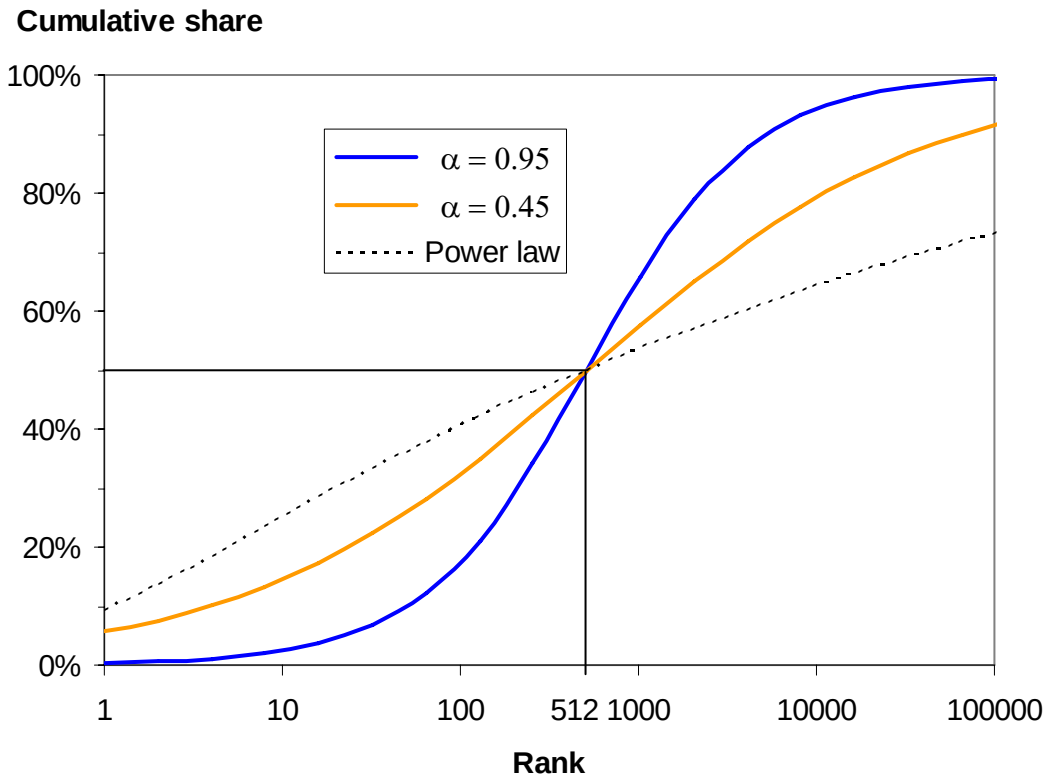

Figure A1. Cumulative share as a function of rank for different values of α. In addition, the cumulative function of a power law function is shown.

Figure A2 illustrates the volume of an individual rank when the total volume of all objects is ten million copies. Figure A2 clearly demonstrates that with a large α, the end of the tail is much thinner than with a small α. Correspondingly, a large α means that even the most popular object has a relatively low popularity.

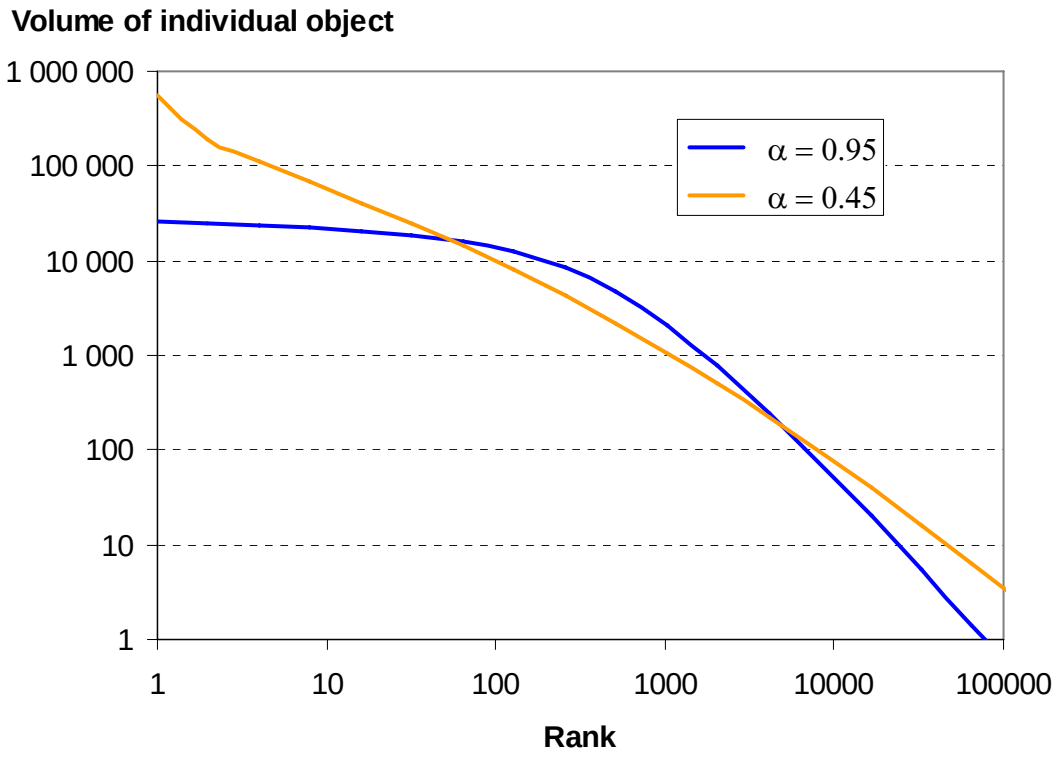

Figure A2. The volume of individual objects as a function of the rank of the object.

Finally, Figure A3 shows how the volume is distributed among set of ranks, in which the boundaries between consecutive sets are powers of two. This figure makes the symmetry of the function (on a logarithmic scale) very clear. When α is 0.45, the middle of the tail (six middle columns) covers 44 percent of the total volume, whereas when α is 0.95, the middle of the tail covers 76 percent of the total volume. The difference is even more distinct for the popularity of the most popular object, that is, 5.7 percent versus 0.27 percent. In principle, the symmetry means that if the long tail model is valid, the base of the tail and the end of the tail are equally strong. In reality, many external issues can spoil the symmetry.

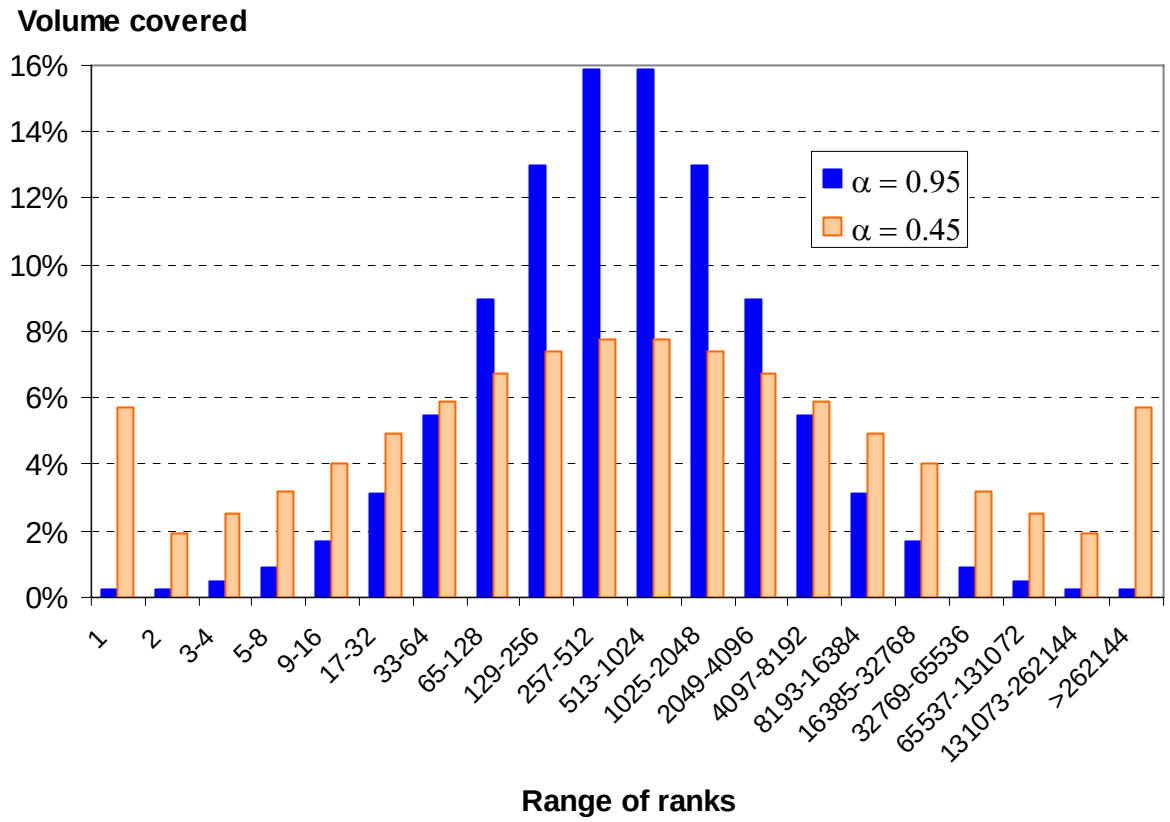

Figure A3. The volume covered by a range of ranks.

The difference in the form of long tails also has business consequences. If α is small (< 0.6), the overall business could be more easily divided into two separate areas: The first consists of a limited number of high popularity objects and the second consists of a very large number of low popularity objects. In contrast, if α is large (> 0.8) the clear majority of the overall business is in the middle of the tail, while the ends of the tail do not provide clear separate business potential. In addition, the business reasoning depends on the length of tail defined by parameter N_{50}: With very long tails, the middle part is relatively strong even when α is small.

There are lots of long tails that seem to behave similarly. Is there something more fundamental than an arbitrary function behind the popularity distribution of diverse objects? The special property of the selected function is the symmetry as regards point (N_{50}, $\beta/2$), which is evident particularly in Figure A3. The symmetry and smooth form of the distribution in Figure A3 also evokes the possibility to use normal

distribution. In this case what is normally distributed is the volume covered by a set of ranks, with equal ratio (e.g. from 33 to 64, and from 2049 to 4096).

Indeed, we could obtain quite similar results based on normal distribution as with the long tail model used in this article. The main difference is that the end of the tail with normal distribution is thinner than with the long the model. Moreover, the behavior of normal distribution is quite problematic with the smallest ranks. Still, the normal distribution assumption may give an opportunity for finding a general process that generates the long tails presented in this article. This idea is left for further studies.

Brief instructions for long tail analysis

1. Gather all relevant statistics about the phenomena. Particularly useful is information about the cumulative volume over the whole tail.
2. Try to ascertain that the data is consistent and reliable. Preferably utilize data from several sources. For instance, if you use search engine hits to define the popularity of objects, carry out some tests to identify the possible anomalies.
3. In case of limited statistics, it is necessary to know at least the total volume and two points of the cumulative curve. Optimal points for cumulative volumes are near the boundaries between the base and the middle of the tail (x_{bm}), and between the middle and the end of the tail (x_{me}).
4. If you know only the volume for some individual objects, the minimum useful information is three points: one point in the base, another one in the middle, and a third one in the end of the tail. Without better data, you have to rely more on general knowledge about the subject. For instance, the knowledge that books typically form a tail with large N_{50} and α around 0.50 may essentially improve the reliability of your analysis. Be particularly careful with data that does not reveal the turning point of the S-curve.
5. There is no evident criterion for selecting the optimal values for the model parameters. One possibility is to minimize the square error between the real data and data predicted by the long tail model. A possibility is to use data points that locate evenly on logarithmic scale (e.g., ranks 1, 2, 4, 8, 16, 32 …). If all data points are included, the individual points in the base of the tail should

obtain a higher weight than the data points in the end of the tail. It might also be reasonable to omit some of the first data points, because we may assume that the values of those points possess more randomness. Always check the result on a graph.

6. If there is a significant difference between the real data and the long tail model, try to find credible explanations. If the real tail seems to be shorter than what the model predicts, possible explanations are latent demand for low popularity objects and the finite size of objects (e.g., one book, or one inhabitant). If the tail is longer than what is expected, a possible explanation is the demand among minorities (e.g., related to language) that does have an effect in the middle of the tail, but may significantly lengthen the end of the tail.
7. Remember the fundamental limitations of the model when you make conclusions. Although the long tail model cannot ever be as exact as the laws in physics, it can still offer a useful estimate of complex behavior of real phenomenon.

As a final note, a fundamental property of long tail leads to a phenomenon that could be called the long tail paradox. Let us take an example. There are perhaps 50 million authors writing articles that could be considered scientific. The same persons who write the articles read the articles of other authors. Now let us assume that an average author writes two articles and reads 200 articles per year. A criterion for an article is that someone else reads the whole article.

Now it is obvious that, on average, an article will be read by 100 persons belonging to the scientific community. So if you pick a random article, you may assume that there are 10,000 other persons that have or will read the same article. Because the popularity of articles apparently forms a long tail, the variations in popularity are certainly huge. Based on our general understanding about long tails, we may assume that the parameters of the distribution are something like: $N_{50} = 175000$ and $\alpha = 0.65$. According to this model, the most popular article will then be read about 4 million times and 100 million articles will be read at least once.

But now if we take one of those real occurrences of reading an article, we may ask, what is the expected number of readers for that particular article? The correct answer is not 99, but about 42,000. Why?

Because there are many more reading occurrences for the most popular articles than for the less popular articles. This reasoning should be valid for this article as well. So thank you for selecting and reading this article — now I, the author of this article, can be proud of the large number of expected readers.

About the author

Kalevi Kilkki is a Principal Scientist at Nokia Research Center, Helsinki, Finland. He received Master of Science and Doctor of Technology degrees from Helsinki University of Technology in 1983 and 1995, respectively. He had made analytical models in various areas including quality of service in communications networks, social networks, and the business of service providers.

Notes and references

[1] C. Anderson, 2004. “The Long Tail,” *Wired*, Issue 12.10, October 2004, at http://www.wired.com/wired/archive/12.10/tail.html, accessed 9 October 2006.

[2] C. Anderson, 2006. *The Long Tail*. London, UK: Random House.

[3] Long tail is not a new term. It has been used in statistics to describe distributions that have essentially thicker tail than exponential distribution, see e.g., http://en.wikipedia.org/wiki/Long_tail. Still, without a doubt, the long tail concept has become much more popular after the article by C. Anderson.

[4] As Stephen Hawkings has expressed it in "A Brief History of Time": “Someone told me that each equation I included in the book would halve the sales.” Anderson’s book has only one equation and it was 47th in the list of Amazon.com’s most popular books in August 2006 (http://www.amazon.com/gp/bestsellers/).

[5] The data is taken from the long tail book p. 121, and is originally from Book Industry Data Group and obviously refers to book sales in the United States.

[6] See e.g. the analysis by Morris Rosenthal at http://www.fonerbooks.com/surfing.htm, accessed 9 October 2006.

[7] The data was obtained from http://www.accessabc.com/reader/top150.htm, accessed September 2006.

[8] Statistics are available at http://www.ses.fi/fi/tilastot.asp, under topic: “Vuonna 2003 ensi-illan saaneet elokuvat. (pdf)”, accessed 9 October 2006.

[9] The statistics are available at http://www.movieweb.com/movies/boxoffice/alltime.php?q=&page=1. The presented results were accessed in August 2006.

[10] Last.fm statistics are available at http://www.last.fm/charts/.

[11] The data was obtained from http://www.last.fm/charts/music/artist/, accessed September 2006.

[12] The data was from http://www.last.fm/charts/music/tag/, accessed 7 September 2006.

[13] The data was obtained from http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists, most common words (TV and movie scripts), accessed in September 2006.

[14] The data was obtained from Wigblog - Things Internet and Otherwise by Richard Wiggins, at http://wigblog.blogspot.com/2005/08/long-tail-and-short-head-of-zipf-curve.html, accessed in 9 October 2006

[15] The data about Google hits (or results) was gathered in August and September 2006 using the Finnish language version of Google at http://www.google.fi/. The number of hits somewhat depends on the selected language. There also are variations between different dates that seem to be quite random.

[16] The data was obtained using Google at http://www.google.fi/, accessed in September 2006.

[17] The data was obtained using Yahoo at http://www.yahoo.com/, accessed in September 2006.

[18] The data was obtained using MSN Search at http://www.live.com/?searchonly=true, accessed in September 2006.

[19] The data was obtained from http://www.census.gov/genealogy/names/names_files.html, accessed in September 2006.

[20] The data was obtained from http://www.census.gov/genealogy/names/names_files.html. File: dist.female.first, 29-Sep-94, accessed in September 2006.

[21] The data was obtained from http://money.cnn.com/magazines/fortune/global500/2006/full_list/, accessed 9 October 2006.

[22] The data was published by Talouselämä magazine. Data is available at http://www.talouselama.fi/te500list.te, accessed in September 2006.

[23] The data has been provided by Statistics Finland, http://www.stat.fi/index_en.html.

[24] This choice of including the volumes with $k < 1$ in the volume of the volume of most popular object raises an interesting philosophical question of what are those objects with a rank of below 1 in the theoretical model. We

may, for instance, say they are those objects that the audience is still waiting for. Because that need does not yet have any concrete target, the most popular object gets extra interest. Regardless of the speculativeness of the idea, the model often works amazingly well.