Predicting the long tail of book sales: Unearthing the power-law exponent

Trevor Fenner, Mark Levene, and George Loizou
Department of Computer Science and Information Systems
Birkbeck, University of London
London WC1E 7HX, U.K.
{trevor,mark,george}@dcs.bbk.ac.uk

Abstract

The concept of the long tail has recently been used to explain the phenomenon in e-commerce where the total volume of sales of the items in the tail is comparable to that of the most popular items. In the case of online book sales, the proportion of tail sales has been estimated using regression techniques on the assumption that the data obeys a power-law distribution. Here we propose a different technique for estimation based on a generative model of book sales that results in an asymptotic power-law distribution of sales, but which does not suffer from the problems related to power-law regression techniques. We show that the proportion of tail sales predicted is very sensitive to the estimated power-law exponent. In particular, if we assume that the power-law exponent of the cumulative distribution is closer to 1.1 rather than to 1.2 (estimates published in 2003, calculated using regression by two groups of researchers), then our computations suggest that the tail sales of Amazon.com, rather than being 40% as estimated by Brynjolfsson, Hu and Smith in 2003, are actually closer to 20%, the proportion estimated by its CEO.

Keywords: long tail, power-law distribution, stochastic model

1 Introduction

The long tail is the phenomenon that allows an e-commerce business to make significant profit from small sales volumes of a large number of less popular items. It is well known that in aggregate the tail sales of many online retailers, offering products such as books, music and films, are comparable to the sales of the most popular items, i.e. the "blockbusters". The proportion of tail sales may be estimated using regression techniques on the assumption that the data obeys a power-law distribution. Newman, in [New05], provides evidence that books sales do indeed follow a power law. In this paper we present a different method for estimating the tail sales based on a generative model that simulates a simplified sales process, and results in an asymptotic power-law distribution. Our generative model also has the advantage that the proportion of tail sales can be estimated from the number of available products and the total volume of sales, when these are known. The methodology we present may be useful in providing sales analytics for an e-commerce business in relation to prediction and validation of sales volumes from the tail.

A power-law distribution taking the mathematical form

$$g(i) = \frac{C}{i^\tau}, \tag{1}$$

where C and τ are positive constants, represents the proportion of observations having the value i. The constant τ is called the *exponent* of the distribution [New05]. There are many well-known examples of power-law distributions [Sch91]; for example, *Lotka's law* states that the number of authors publishing a prescribed number of papers is inversely proportional to the square of the number of publications. *Pareto's law*, which is a cumulative version of (1), states that the number of people whose personal income is above a certain level follows a power law with an exponent between 1.5 and 2. *Zipf's law*, which states that the relative frequency of a word in a text is inversely proportional to its rank, is the inverse of the cumulative power-law distribution for the proportion of words whose frequency is above a certain level. (We note that for $\tau > 1$ the cumulative distribution corresponding to the distribution (1), i.e. the proportion of observations greater than i, also follows a power law, but with exponent $\tau - 1$. Its inverse is a *Zipfian* distribution of frequency against rank, which also follows a power law, now with exponent $1/(\tau - 1)$.)

The tail of a power-law distribution decays polynomially, in contrast to the exponential decay characteristic of distributions such as the Normal and geometric. Power-law distributions are notoriously hard to fit [GMY04], and often there is an exponential cutoff present in the power-law scaling, although this cutoff may only be observable in the tail of the distribution for extremely large data sets [FLL06]. A power-law distribution with exponential cutoff [FLL05] is of the mathematical form

$$g(i) = \frac{C\, q^i}{i^\tau}, \tag{2}$$

where $0 < q < 1$, and frequently $q \approx 1$.

The concept of the *long tail* has been recently popularised by Anderson [And06] (see also `www.longtail.com`) and is currently used to explain the phenomenon in e-commerce where the total volume of sales of the items in the tail of a Zipfian distribution of sales volume against sales rank is comparable to that of the most popular items. One category of sales to which long tail analysis has been applied is online book sales. In [BHS03, BHS06] it was argued that the considerable increase of product variety in online book stores has a significant positive impact on consumer welfare. Their analysis also applies to other products such as CDs and DVDs. Table 1, taken from [BHS03], shows the numbers of products available from Online and Brick-and-mortar stores.

In [CG03] an analysis of online book sales data was carried out to compare the demand and price competition between Amazon.com and BarnesandNoble.com. In order to analyse the long tail, the exponent of the assumed cumulative power-law distribution relating the sales rank of a book to the number of copies sold needs to be estimated. In [CG03] the estimate of $\tau - 1$ used was 1.2, while in [BHS03] the slightly lower value of 1.1481 was used. (The sales rank of a book is one greater than the number of books that have sold more copies.) Based on the latter estimate of the power-law exponent and assuming, as shown in Table 1, that the most popular 100,000 titles are stocked in Brick-and-mortar stores, Brynjolfsson et al. [BHS03] concluded that about 40% of Amazon.com's sales are represented by titles that would not normally be found in these stores. It is interesting to note that Jeff Bezos, the

Product Category	Online	Brick-and-mortar
Books	2,300,000	40,000 – 100,000
CDs	250,000	5,000 – 15,000
DVDs	18,000	500 – 1,500
Digital cameras	213	36
Portable MP3 players	128	16
Flatbed scanners	171	13

Table 1: Product variety comparison for large Online and Brick-and-mortar retailers.

CEO of Amazon.com, thought that the 40% figure was too high and the real figure was closer to 20% [And05]. So, assuming that Bezos is correct, how can we explain this discrepancy?

There is some inconsistency in the estimation of the power-law exponent for Amazon.com's sales data, and different researchers have reported values for $\tau-1$ in the range from just below 1.0 to approximately 1.3 [BHS03, CG03]. This is not surprising, since there are inherent difficulties in fitting power-law distributions [GMY04, FLL06] and it is often unclear whether or not the distribution is indeed a pure power law. We will show that the generative model we describe in the next section supports the exponent $\tau - 1$ being in the region of 1.1 for Amazon.com's sales data, assuming that Jeff Bezos's estimate of 20% tail sales is closer to reality than 40%.

A recent approach, which to a certain extent circumvents the above problems, is to assume a generative model that results in a distribution that is asymptotically either a pure power-law distribution or a power-law distribution with exponential cutoff, where q in (2) is close to 1.0 [New05]. (The latter covers a wider range of real-world scenarios than a pure power law.) The details of such a model are given in Section 2. We use this model to investigate the possible range of power-law exponents that are consistent with the book data given in Table 1 corresponding to 20%, 30% or 40% of the sales being in the tail of the distribution. The methodology we use and our results are presented in Section 3, and analysis of a sparse data set from [And06] is presented in Section 4. Finally, in Section 5 we give our concluding remarks.

2 A stochastic model exhibiting power-law behaviour with an exponential cutoff

The stochastic model presented in [FLL05] can be described, in the context of the sales of products, and in particular books, as follows. We have at our disposal a countable number of urns, say $urn(i), i = 1, 2, \ldots$, where each urn contains a number of products, for example, books or CDs. A product is in $urn(i)$ if i copies of it have been sold since it entered the system. Initially all the urns are empty except $urn(1)$, which has one product in it (of which one copy has been sold). At time $t + 1$, one of the following two things occurs:

(a) with probability p, $0 < p < 1$, a new product is inserted into $urn(1)$ (this represents the first sale of this product), or

(b) with probability $1 - p$, a product is chosen with probability proportional to the number

of sales of that product up to time t, i.e. proportional to i for a product in $urn(i)$ (thus the selection follows the rule of preferential attachment [AB02] originally suggested by [Sim55] and [Pri76]); then,

(i) with probability q, $0 < q \leq 1$, the chosen product is transferred from $urn(i)$ to $urn(i+1)$ − this represents an additional copy of the chosen product being sold, or

(ii) with probability $1-q$, the chosen product is discarded from $urn(i)$ − this represents the chosen product being discontinued, for example, if a book has gone out of print; when $q = 1$, "old" products are always available.

We have assumed the above initial conditions of the model for the sake of simplicity. However, it can be shown that any other initial conditions will lead to the same asymptotic distribution described below.

For the following analysis we will assume from now on that $q = 1$, since, for the Amazon.com sales data considered here, the researchers [BHS03, CG03] have assumed a pure power-law distribution (i.e. that $q = 1$). From a web perspective such an assumption is not unreasonable, since second-hand out-of-print books are often available online, and the trend towards a print-on-demand model is increasing the availability of less popular books. Nevertheless, if online sales data sets become available, it could be tested whether in practice there exists a noticeable cutoff in the power-law distribution.

Let $g(i)$ be the asymptotic proportion of products in $urn(i)$. It was shown in [Sim55] (cf. [FLL05]) that, for $i > 1$,

$$g(i) = \frac{i-1}{i+\tau-1}\ g(i-1) = \frac{(\tau-1)(i-1)!}{\tau(1+\tau)\cdots(i+\tau-1)} = \frac{(\tau-1)\Gamma(i)\Gamma(\tau)}{\Gamma(i+\tau)}, \tag{3}$$

and, for $i = 1$,

$$g(1) = \frac{\tau-1}{\tau},$$

where

$$\tau = \frac{2-p}{1-p} \tag{4}$$

and Γ is the gamma function [GKP94]. It was also shown that

$$\sum_{i=1}^{\infty} g(i) = 1 \quad \text{and} \quad \sum_{i=1}^{\infty} i g(i) = \frac{1}{p}, \tag{5}$$

which is consistent with the fact that p is the proportion of sales that are the first sale of that product, i.e. the ratio of the number of products sold to the total number of sales.

On using Stirling's approximation [GKP94], it can be shown from (3) that for large i, corresponding to (1), we obtain:

$$g(i) \sim \frac{C}{i^{\tau}},$$

where $\sim$ means *is asymptotic to*, and $C = (\tau-1)\Gamma(\tau)$. (We have shown in [FLL05] that $g(i)$ is asymptotic to the more general form given in (2) when $q < 1$.) We note that, for any given exponent τ, the value of $g(i)$ will be less than the above asymptotic approximation. Moreover, the difference will be greater for small values of i, which correspond to the tail sales, as we will see in the next section.

3 How long is the tail?

From now on we will assume that we are dealing with books. We define the long tail to comprise those titles available online but not in Brick-and-mortar book stores. To find the proportion of sales in the long tail, taking the figures from Table 1, we assume that $2,300,000$ is an estimate of α, the total number of titles in print, and thus the number of titles potentially available online; we also assume that β, the number of titles that are stocked in a very large Brick-and-mortar book store, is approximately $100,000$.

Let N be the minimum number of copies sold by an online store of any title stocked in a Brick-and-mortar store. In order to find the proportion of tail sales, we first need to find N, i.e. the smallest integer such that

$$\alpha \sum_{i=1}^{N} g(i) > \alpha - \beta, \tag{6}$$

where the right-hand side represents the number of titles in the long tail of the distribution. We note that, for the stochastic model of the preceding section, when $j > i$ the titles in $urn(j)$ are more popular than those in $urn(i)$, since they have sold more copies. Thus in (6) we sum $g(i)$, the proportion of titles in $urn(i)$, from 1 to N in order to find a bound on the number of titles in the tail.

Now let δ be the number of titles in $urn(N)$ that are in the tail, so

$$\delta \;=\; \alpha \;-\; \beta \;-\; \alpha \sum_{i=1}^{N-1} g(i). \tag{7}$$

The proportion λ of sales in the long tail is given by

$$\lambda \;=\; p \sum_{i=1}^{N-1} i g(i) \;+\; \frac{pN\delta}{\alpha}, \tag{8}$$

since the proportion of sales corresponding to the titles in $urn(i)$ is $ig(i)p$ by (5).

Given α and β, we obtain an estimate for the power-law exponent τ that corresponds to a given value of λ in the following way. For a suitable range of values of τ we compute p from (4) and the $g(i)$ from (3). Inequality (6) is then used to obtain N, the threshold volume for the tail. Finally, λ is calculated from (7) and (8). We can then choose the value of τ that corresponds most closely to the given value of λ. Moreover, if we know the average number of copies sold per title, which by (5) is $1/p$ in our model, τ can be calculated from (4) and then the corresponding value of λ can be computed as described above.

The figure of 40% sales in the tail was estimated by Brynjolfsson et al. [BHS03] using a power-law exponent for the cumulative distribution of $\tau - 1 = 1.1481$, the reciprocal of 0.871, the exponent obtained by fitting a Zipfian distribution [New05] relating sales volume to sales rank. (A previous estimate from 2002, relying on only 2 points rather than 800 points, gave an exponent of 1.0917, the reciprocal of 0.916; see [BHS03].) The estimated exponent used in [CG03], however, was the higher value of 1.2.

Table 2 shows the values of λ obtained for several values of $\tau - 1$, calculated as described above. It can be seen that the estimated proportion of sales in the tail is very sensitive to the

power-law exponent. In particular, if the power-law exponent of the cumulative distribution is close to 1.1, the earlier estimate in [BHS03], rather than to 1.2 [CG03], then the results obtained using our model suggest that the tail sales for Amazon.com would, in fact, be closer to 20%. This figure is consistent with the estimate reported by its CEO, rather than 40% as estimated in [BHS03]; if the exponent is close to 1.15, the more recent estimate in [BHS03], our results suggest that the proportion of sales in the tail would be close to 30%.

Tail sales (λ)	Exponent ($\tau - 1$)
20.0%	1.0896
20.4 %	1.0917
29.4 %	**1.1481**
30.0%	1.1522
36.1 %	**1.2000**
40.0%	1.235

Table 2: Proportion of tail sales for various cumulative power-law exponents.

The interpretation of our results is *not* that the method used in [BHS03] to estimate the power-law exponent was more accurate than that used in [CG03], or vice versa, but rather a general critique on fitting power-law distributions. A generative model, such as the stochastic urn model presented in Section 2, can be useful for validating power-law statistics, especially if additional information is available, such as Jeff Bezos's estimate in this case.

4 Further analysis of the proportion of tail sales

We now attempt to verify the above results using the book sales data presented by Anderson [And06, p.121], which he used to support his argument that sales data follows a power-law distribution or, in his terminology, is "long-tailed". (This was the only book sales data set we were able to obtain – book sellers are rather reluctant to provide their sales data, presumably for commercial reasons.) In Table 3 we reproduce this sparse data set. *Range* refers to the range of the number of copies sold for each book, *Books* refers to the number of different book titles that sold within the range, and *Units* refers to the total number of copies of books sold within the range. We assume as before that the data comes from an asymptotic power-law distribution following the model presented in Section 2. Unfortunately, due to the sparseness of the data in Table 3, we cannot reliably determine the exponent of the distribution using regression techniques. Thus we will resort to measuring the distance between the empirical distribution, as given in Table 3, and the distribution according to the model, as given by (3).

From the raw data in the table we can calculate that the tail sales for this data set is approximately 12.16%, and from our model we can then deduce that $\tau - 1 \approx 1.065$. We note that it is reasonable to expect that the tail sales for this data set will be less than that for Amazon's, since the total number of books available here is approximately $1,240,000$, as opposed to Amazon's $2,300,000$. Moreover, because of the self-similarity property of power laws [DKM$^+$02], it is also reasonable to assume that the distributions will be similar.

We now make use of two well-known non-parametric measures of distance between two distributions, namely the *Hellinger distance* and the *relative entropy* [GS02], in order to estimate the tail sales corresponding to the data in Table 3 using our model. Let $(a_1, a_2, \ldots, a_n)$ be

Range	Books	Units
$\geq$ 1,000,000	10	17,396,510
500,000 to 999,999	22	13,798,299
250,000 to 499,999	64	22,252,491
100,000 to 249,999	324	46,932,031
50,000 to 100,000	767	51,858,835
5,000 to 49,999	23047	280,000,591
1,000 to 4,999	67008	149,093,614
100 to 999	202938	69,548,499
$\leq$ 99	948005	14,346,417
Total:	1,242,185	665,227,287

Table 3: Book sales data from 2004 [And06].

the empirical probabilities of the distribution obtained from the data, and let $(b_1, b_2, \ldots, b_n)$ be the corresponding probabilities of the distribution as predicted by the model. Then, the Hellinger distance (He) is defined by

$$\mathrm{He} = \sum_{i=1}^{n} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2,$$

and the relative entropy (Re) is defined by

$$\mathrm{Re} = \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}.$$

We note that the Hellinger distance is bounded between 0 and 2, and its value cannot be greater than that of the relative entropy [GS02].

Tail sales (λ)	Exponent ($\tau-1$)	He Units	Re Units
2.10%	1.0100	0.2361	0.8008
10.00%	1.0522	0.1654	0.5541
12.16%	1.0650	0.1471	0.4934
16.30%	1.0917	0.1146	0.3888
20.00%	1.1180	0.0907	0.3160
23.80%	**1.1481**	0.0739	**0.2699**
29.60%	**1.2000**	**0.0716**	0.2818
30.00%	1.2043	0.0729	0.2878
40.00%	1.3410	0.1959	0.8098

Table 4: Data analysis for sales data from Table 3.

Since the sales data in Table 3 are given for ranges rather than point sales, we computed the distance between the cumulative empirical distribution of sales for the data given in Table 3, and the cumulative distribution of sales predicted by our model, both normalised

to sum to one. Using (3), it can be shown that the formula for the cumulative sales in our model is

$$\sum_{i=j}^{\infty} ig(i) = \frac{(\tau-1)^2\ \Gamma(\tau-2)\Gamma(j+1)}{\Gamma(j+\tau-1)},$$

and this can be used to calculate the cumulative sales distribution for our model.

The results are shown in Table 4. It can be seen that according to these two distance measures the most likely percentage of tail sales (not necessarily Amazon.com's) is between 20% and 30%, which is consistent with our findings in Section 3. We stress that for this data set we have assumed that the total number of books α is as given in Table 3, which differs from the value from [BHS03] used in Section 3. This could explain why, for an exponent of 1.1481, the model predicts the tail sales as being 23.8% rather than 29.4% as in Section 3. Although a definitive answer of the percentage of Amazon.com's tail sales can only be verified from their sales data, we have demonstrated that our methodology may be useful for estimation and confirmation purposes.

5 Concluding Remarks

We have investigated the long tail using a generative model of book sales. This attempts to model the process that gives rise to an asymptotic power law, rather than using techniques such as regression for fitting an apparent power-law distribution. The advantage of this new approach is that the parameters of the model, p and q, are related to the rates at which new products are introduced and existing ones discontinued. When data on these is available our methodology can be used to calibrate and validate the model. The generative approach can also be useful in providing sales analytics for an e-commerce business in relation to predicting and validating tail sales volumes. It may also be useful for a manager who wishes to investigate alternative long tail strategies when considering target sales of both popular and niche items.

Our computations show that the estimated proportion of sales in the tail is very sensitive to the estimated power-law exponent. Consequently, given that an exponent estimated using regression may not be very accurate, the error margin of the resulting tail sales estimate may be quite large. When sales data are believed to asymptotically follow a power law, our model could be used to reduce the margin of error of such estimates.

It would be interesting to be able to test the validity of the model we have presented on other e-commerce data, where this is available.

As an epilogue, we believe that as more households world-wide progressively move to conduct more of their daily business through the internet, the long tail phenomenon will have to be periodically re-examined. Thus the close monitoring of long tail sales could potentially become a useful economic instrument, which can contribute to the profitability of companies.

References

[AB02] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[And05] C. Anderson. A methodology for estimating Amazon's long tail sales, August 2005. See `http://longtail.typepad.com/the_long_tail/2005/08/a_methodology_f.html`.

[And06] C. Anderson. *The Long Tail: How Endless Choice is Creating Unlimited Demand*. Random House Business Books, London, 2006.

[BHS03] E. Brynjolfsson, Y.J. Hu, and M.D. Smith. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science*, 49:1580–1596, 2003.

[BHS06] E. Brynjolfsson, Y.J. Hu, and M.D. Smith. From niches to riches: Anatomy of the long tail. *Sloan Management Review*, 47:67–71, 2006.

[CG03] J. Chevalier and A. Goolsbee. Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics*, 1:203–222, 2003.

[DKM+02] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2:205–223, 2002.

[FLL05] T.I. Fenner, M. Levene, and G. Loizou. A stochastic evolutionary model exhibiting power-law behaviour with an exponential cutoff. *Physica A*, 335:641–656, 2005.

[FLL06] T.I. Fenner, M. Levene, and G. Loizou. A stochastic model for the evolution of the web allowing link deletion. *ACM Transactions on Internet Technology*, 6:117–130, 2006.

[GKP94] R.L. Graham, D.E. Knuth, and O. Patachnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, Ma., 2nd edition, 1994.

[GMY04] M.L. Goldstein, S.A. Morris, and G.G. Yen. Problem with fitting to the power-law distribution. *European Physical Journal B*, 41:255–258, 2004.

[GS02] A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–435, 2002.

[New05] M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[Pri76] D.J. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society of Information Science*, 27:292–306, 1976.

[Sch91] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman, New York, NY, 1991.

[Sim55] H.A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.