A Longer Tail?: Estimating The Shape of Amazon's Sales Distribution Curve in 2008

Erik Brynjolfsson, Yu (Jeffrey) Hu, Michael D. Smith

1. Introduction

The term "The Long Tail" was coined by *Wired*'s Chris Anderson (Anderson 2004) to describe a phenomenon where niche products account for a much larger proportion of sales in Internet markets than they do in brick-and-mortar markets. This phenomenon has captured much attention and debate in the popular press (e.g., Gomez 2006, Orlowski 2008) and in the information systems, marketing, and operations management literatures. In an earlier study of the Internet's Long Tail phenomenon, Brynjolfsson, Hu, and Smith (2003) found that a log-linear relationship (a power law) can be used to describe the relationship between Amazon sales and Amazon sales rank. Assuming such a log-linear relationship holds for all the books sold by Amazon, we estimated that niche books that are not typically stocked in brick-and-mortar bookstores accounted for 39% of Amazon's total sales in 2000, and that the sales of these niche titles enhanced consumer surplus by $731 million to $1.03 billion in 2000.

Following Brynjolfsson et al. (2006), we hypothesize that there are several factors that might increase the proportion of "niche" sales over time. First, exposure to niche products could drive consumers to develop a taste for more niche products. Second, producers could have an increased incentive to create more new niche products over time. Finally, search tools, product reviews, product popularity information, and recommendation engines could increase the sales of niche products disproportionately.

On the other hand, some have argued that the Long Tail may be a short-lived phenomenon. For instance, over time, consumers who buy from Amazon could become less dominated by "early adopters" of e-commerce who may have a strong taste for niche products. A larger proportion of mainstream consumers could lead to proportionately more sales of popular products, reducing the size of Amazon's Long Tail. In addition, Amazon's search and recommendation tools could be tuned (intentionally or unintentionally) to promote popular products (Hosanagar 2008). Finally, producers of popular products could employ online marketing strategies to promote their products and counteract the effect of Amazon's search and recommendation tools promoting niche products.

To analyze whether the Long Tail phenomenon represents a temporary of permanent shift, we collected Amazon sales and sales rank data in 2008 on a larger and broader sample of books than was available in our 2000 data. We then match this sample to our 2000 sample to compare changes in the profile of sales over time. Our analyses suggest that the "long tail" of Internet book sales has gotten longer from 2000 to 2008. We also develop a new methodology to fit the relationship between sales and sales rank and apply it to our 2008 data. Our analyses suggest that niche books account of 36.7% of Amazon's sales in 2008 and the consumer surplus generated by niche books has increased at least five fold from 2000 to 2008.

2. Literature

Economic explanations for the existence of superstars and popular products can be traced to Rosen (1981) and Frank and Cook (1995). Brynjolfsson et al. (2006) point out several demand-side and supply-side factors that could drive sales to niche products on the Internet. These demand-side and supply-side factors can even reinforce each other. For instance, Cachon et al. (2006) show that low consumer search costs can enhance a retailer's incentive to provide a large product selection. This could lead to even more sales of niche products. There is also a growing body of literature that empirically examines sales distributions in various product markets. Brynjolfsson et al. (2007) find the sales distribution of an Internet channel is less concentrated than that of a catalog channel, using data from a clothing retailer. Elberse and Oberholzer-Gee (2008) find evidence that Internet retailing has shifted

demand toward niche video products over time, although they also find that a substantial proportion of niche products have almost zero sales. Chellappa et al. (2007) have similar findings for music sales.

However, none of these papers addresses whether the Long Tail phenomenon is a temporary or permanent shift, or how it might change over time. To answer these questions, one needs to compare the sales distribution of a similar profile of products over a sufficiently long period of time . In this paper, we address this question by collecting data on Amazon sales and sales rank in 2008. We then use the "sample matching" statistical technique to construct a 2008 sample that is comparable to the 2000 sample used in Brynjolfsson et al. (2003).

3. Data

The data for this paper come from a major publisher with annual sales of more than $1 billion. The publisher provided us with their Amazon sales and sales rank data on a sample of 1,598 titles over 10 weeks from June to August 2008. Overall, we have 15,980 observations of Amazon sales and sales ranks. Table 1 compares the summary statistics for our 2000 and 2008 samples. It is clear that our 2008 sample has more observations (15,980 vs. 901) and covers a much broader spectrum of books (sales ranks of 71 to 5,350,140 versus sales ranks of 238 to 961,367 million) than our 2000 data does.

Table 1: Summary Statistics for Our 2008 Sample and Our 2000 Sample

Variable	Obs.	Mean	S.D.	Min	Max
2008 Sample					
Weekly Sales	15,980	3.04	17.79	0	950
Weekly Sales Rank	15,980	338,238	330,780	71	5,350,140
2000 Sample					
Weekly Sales	901	18.32	30.20	0	480
Weekly Sales Rank	901	34,054	61,001	238	961,367

4.1 Sample Matching

We use a statistic technique called "sample matching" (Rassler 2002) to construct a sub-sample from our 2008 sample that matches our 2000 sample on the basis of weekly sales rank. Summary statistics for the 2008 matched sample (reported in Table 2) are then very comparable to those for our 2000 sample (shown at the bottom of Table 1).

Table 2: Summary Statistics for 2008 Matched Sample

Variable	Obs.	Mean	S.D.	Min	Max
Weekly Sales	901	21.32	41.30	0	354
Weekly Sales Rank	901	34056.1	60999.1	226	961,146

4.2 Re-estimating Amazon's Long Tail

We then repeat the estimation of the log-linear relationship between Amazon sales and sales rank, using the 2008 matched sample. The linear regression model we use is:

$$y_i = \beta_0 + \beta_1 x_i , \quad (1)$$

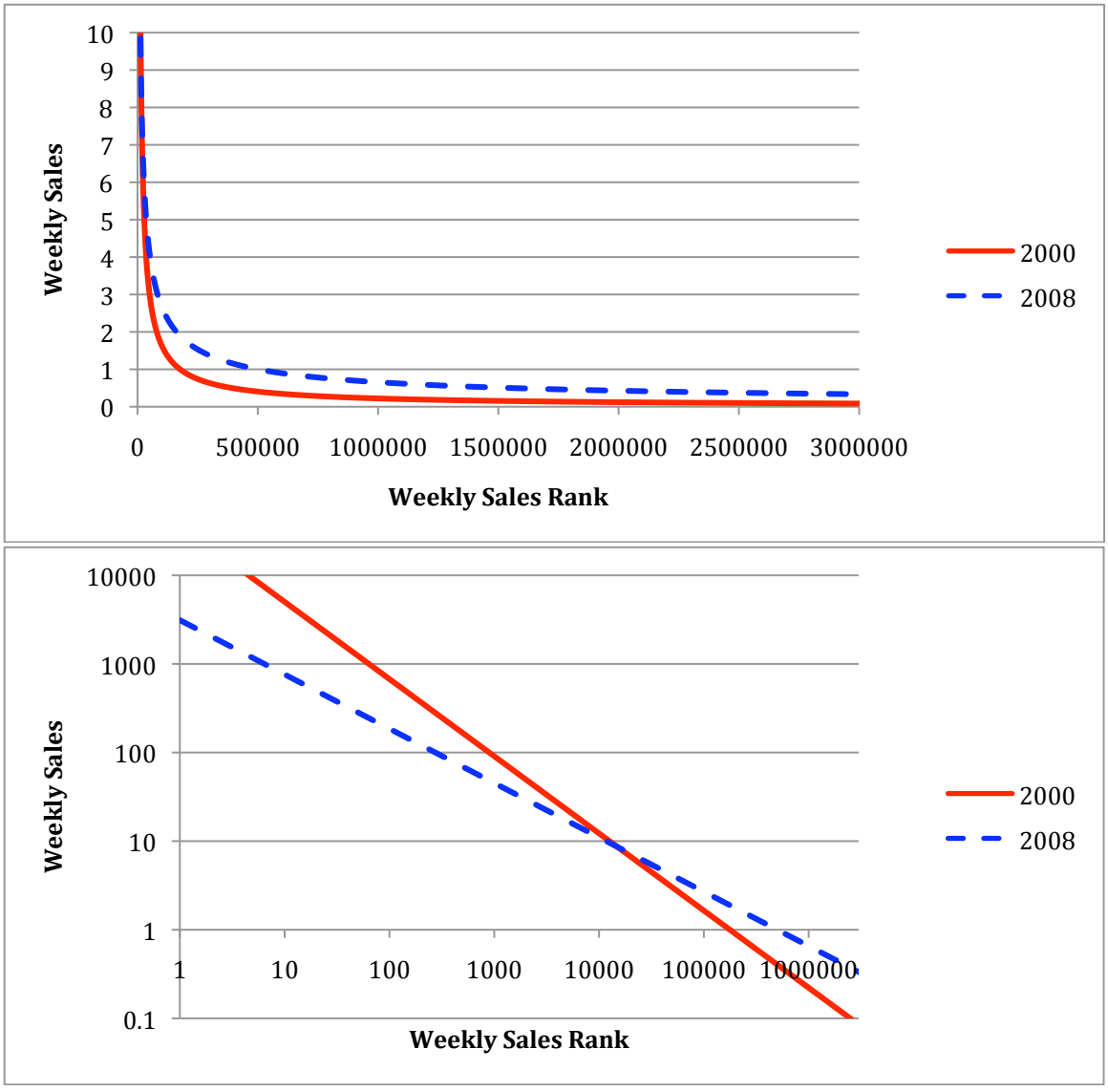
where y_i is the natural log of *Weekly Sales*, and x_i is the natural log of *Weekly Sales Rank*. The estimation results using the 2008 matched sample are reported in Column (1) of Table 3, with the analogous results from our 2000 data in Column (2). Note that 41 observations are dropped after taking the natural log of *Weekly Sales* in the 2008 matched sample, and 40 observations are dropped in the 2000 sample. The

coefficient on *Log(Weekly Sales Rank)* is -0.613 when the 2008 matched sample is used, significantly smaller in size than the same coefficient when the 2000 sample is used (-0.871).

Table 3: Results of The Log-linear Regression

	2008 Matched Sample (1)	2000 Sample (2)
Constant	8.046	10.526
	(0.432)	(0.156)
Log(Weekly Sales Rank)	-0.613	-0.871
	(0.042)	(0.017)
Obs.	860	861
R^2	0.311	0.801

Figure 1: Amazon's Long Tail in 2008 vs. in 2000



The above results provide empirical evidence that Amazon's Long Tail has become longer and fatter in 2008 than in 2000. As sales ranks increase, book sales decline. Such a decline is at a slower pace in 2008 than in 2000, as shown by the relative smaller size in the coefficient on *Log(Weekly Sales Rank)* in 2008. Figure 1 shows the estimated log-linear relationship between Amazon sales and sales rank, with the 2008 results in blue and 2000 results in red. We plot these two curves on both a normal scale and a

logarithmic scale. These two curves cross when sales rank is 14,949. This means popular books (with sales rank below 14,949) tend to sell fewer copies in 2008 than in 2000, while niche titles (with sales rank below 14,949) tend to generate more sales in 2008 than in 2000.

4.3 Developing A More Accurate Method of Estimating Amazon's Long Tail

The log-linear regression method assumes that the coefficient on *Log(Weekly Sales Rank)* does not vary as a book's sales rank increases. It is possible that this assumption may not hold. In this paper, we fit the relationship between *Log(Weekly Sales)* and *Log(Weekly Sales Rank)* to a series of splines, rather than just a single line. Such a spline fitting technique allows the slope coefficient to vary as a book's sales rank increases, leading to a more accurate estimate of the size of Amazon's Long Tail.

Our 2000 sample does not contain any observation with *Weekly Sales Rank* above 1million. In our 2008 sample, we have 569 observations with *Weekly Sales Rank* above 1 million, allowing us to more accurately estimate the shape of Amazon's Long Tail for books with sales ranks above 1 million.

Finally, books with *Weekly Sales Rank* above 1 million frequently have zero *Weekly Sales* as well. Our original method took the natural log of *Weekly Sales*, dropping observations with 0 sales. To utilize these observations, we now use a negative binomial regression model, rather than a linear regression:

$$f(y_i \mid x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, y_i = 0,1,2,3,... \quad (2)$$

where y_i is *Log(Weekly Sales)*, x_i is *Log(Weekly Sales Rank)*, $E(y_i \mid x_i) = \mu_i$ is the conditional mean, and ε_i is unobserved heterogeneity following a log-gamma distribution with $\varepsilon_i \sim \Gamma(\theta,\theta)$ (Cameron and Trivedi 1998). We model the natural log of conditional mean as a series of splines of x_i that are broken down at the 25th, 50th, and 75th percentile of x_i:

$$\ln(\mu_i) = \beta_0 + \beta_1 x_i + \beta_2 (x_i - p_2) I(x_i > p_2) + \beta_3 (x_i - p_3) I(x_i > p_3) + \beta_4 (x_i - p_4) I(x_i > p_4) \quad (3)$$

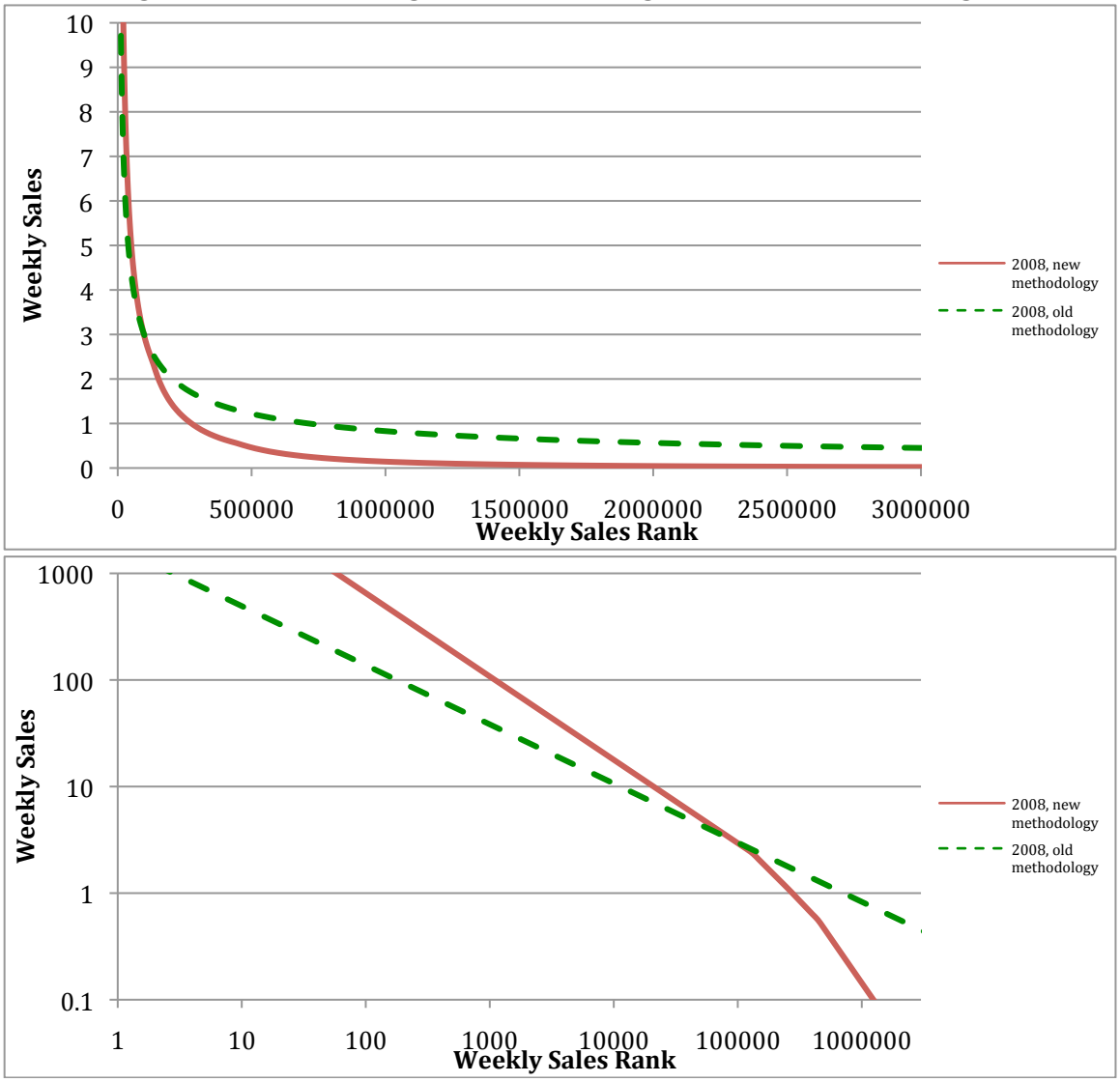
where p_2 is the 25th percentile of x_i (11.78), p_3 is the 50th percentile of x_i (12.46), and p_4 is the 75th percentile of x_i (13.02). The results using the new methodology (negative binomial regression and splines) are reported in Column (1) of Table 4. To show the difference between the new methodology and the old methodology (linear regression and no splines), we estimate the model in equation (1) and report the results in Column (2) of Table 4. Figure 2 shows the estimated curves, with the curve using the new methodology in red and the curve using the old methodology in green.

Table 4: Results Using New Methodology vs. Old Methodology

	2008 Sample and New Methodology (1)	**2008 Sample and Old Methodology (2)**
Constant	10.083	7.480
	(0.210)	(0.092)
x_i	-0.782	-0.555
	(0.019)	(0.008)
$(x_i - p_2)I(x_i > p_2)$	-0.378	
	(0.077)	
$(x_i - p_3)I(x_i > p_3)$	-0.058	
	(0.149)	
$(x_i - p_4)I(x_i > p_4)$	-0.463	
	(0.186)	
Obs.	15,980	7,668

Table 4 shows that the coefficient on *Log(Weekly Sales Rank)* is -0.555 if we use the old methodology. When the new methodology is used, the coefficient on the first spline becomes -0.782, while the coefficients on the other splines are negative (the coefficients on the second and fourth splines are statistically significant). These results indicate that the slope coefficient becomes more negative as a book's sales rank increases. In other words, book sales decrease at a pace that is faster than a regular power law, as the book's sales rank increases.

Figure 2: Amazon's Long Tail in 2008, Using New and Old Methodologies



4.4 Re-estimating The Size of Amazon's Long Tail in 2008

Our new methodology allows us to fit the relationship between *Log(Weekly Sales Rank)* and *Log(Weekly Sales Rank)* more accurately. To obtain an accurate estimate of the total sales and the sales generated by books ranked above 100,000, we simply integrate under the curve as shown in Column (1) of Table 4 and find that books ranked above 100,000 account for 36.7% of Amazon's total sales in 2008. The estimates in Column (2) of Table 4 using the old methodology would have estimated that books ranked above 100,000 account for 82.57% of Amazon's total sales in 2008.

The methodology used by Brynjolfsson et al. (2003) — a linear regression without splines — could have caused an overestimate of the size of Amazon's Long Tail. This is mainly because the assumption that the coefficient on *Log(Weekly Sales Rank)* does not vary as a book's sales rank increases may not hold.

We then use this and our other calculations to estimate the consumer surplus gain from "long tail" books in 2008. Following our prior work, we take 100,000 as the cutoff point for "niche" books, and recalculate the consumer surplus generated from selling these niche books on the Internet. Several changes have happened in the eight years from 2000 to 2008. First, the number of books in print has increase from 2.3 million in 2000 to 3-5 million in 2008. Second, book industry revenue has climbed from $24.6 billion to $37.3 billion. Third, the share of book purchases through the Internet channel has risen from 6% in 2000 to 21-30% in 2008. Combining these changes with the new estimates of the percentage of sales in the Long Tail, we estimate that selling niche books that are unavailable in brick-and-mortar stores leads to a consumer surplus of $3.93 billion to $5.04 billion in the year 2008. These estimates are five times of the estimates in Brynjolfsson et al. (2003), even though the estimates in Brynjolfsson et al. (2003) are likely to have been overestimates.

References

Anderson, C. 2004. The Long Tail. *Wired Magazine* 12(10) 170–177.

Brynjolfsson, E., Y. J. Hu, M. D. Smith. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49(11) 1580–1596.

Brynjolfsson, E., Y. J. Hu, M. D. Smith. 2006. From niches to riches: The anatomy of the long tail. *Sloan Management Review* 47(4) 67–71.

Brynjolfsson, E., Y. J. Hu, D. Simester. 2007. Goodbye Pareto principle, hello long tail: The effect of search costs on the concentration of product sales. Working Paper.

Cachon, G. P., C. Terwiesch, Y. Xu. 2008. On the effects of consumer search and firm entry in a multiproduct competitive market. *Marketing Science* 27(3) 461–473.

Chellappa, R., B. Konsynski, V. Sambamurthy, and S. Shivendu. 2007. An Empirical Study of the Myths and Facts of Digitization in the Music Industry. Workshop on Information Systems and Economics, Montreal, Canada.

Elberse, A., F. Oberholzer-Gee. 2008. Superstars and underdogs: An examination of the long tail phenomenon in video sales. Working Paper.

Fleder, D., K. Hosanagar. 2008. Blockbuster culture's next rise and fall: The impact of recommender systems on sales diversity. *Management Science* (Forthcoming).

Gomes, L. 2006. It may be a long time before the long tail is wagging the web. *Wall Street Journal* July 26th.

Orlowski, A. 2008. Chopping the long tail down to size. The Register Nov 7.

Rassler, S. 2002. *Statistical Matching: A frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer, New York

Rosen, Sherwin. 1981. The economics of superstars. *American Economic Review* 71(5) 845– 858.

Tucker, Catherine, Juanjuan Zhang. 2009. How does popularity information affect choices? A field experiment. MIT Sloan Working Paper.