The Long Tail is Longer than You Think:
The Surprisingly Large Extent of Online Sales by Small Volume Sellers

Joe Bailey, Gordon Gao, Wolfgang Jank,
Mingfeng Lin, Hank Lucas, and Siva Viswanathan

The Robert H. Smith School of Business
University of Maryland

May 12, 2008

DRAFT

Abstract

Sales by small volume sellers are systematically undercounted in public and private surveys of ecommerce. The twin results are that the contribution of small sellers to the ecommerce marketplace is considerably larger than generally assumed and the overall market is larger by this difference.

As the costs of selling things online have fallen with cheaper equipment and communications fees, and with the availability of retail platform services provided by eBay, Amazon.com, Google, and many other firms, Internet retailing has grown to include many small businesses and individual occasional sellers, particularly in the United States. But how much do these "small sellers" sell each year?

U.S. Government statistics give some insight into the type and sales volume of online sellers, but the Government's current methods of data collection and analysis are better suited to tracking larger, traditionally organized businesses, rather than "small sellers," whether operating as small businesses or as individuals. Traditionally, small sellers simply were ignored. In traditional retail markets, the number of businesses with low annual revenues may not be significant because the contribution of such small sellers to the overall size of the market is relatively small. However, in Internet retailing, there are millions of small sellers that, in the aggregate, make a large contribution to the overall market. Yet these small sellers are systematically overlooked in government and private data collection and analysis.

In this paper, we estimate the size of Internet retailing in 2004 to have been over 20% above U.S. Government estimates – and the difference is explained by a more accurate accounting of sales by small sellers. We do this through a variety of methods and the development of confidence intervals in our data. We hope that the techniques outlined in this paper will give greater insight into the magnitude of Internet retailing, particularly in the "long tail" of the ecommerce market occupied by small volume sellers.

1. Introduction

Traditionally, estimates of non-B2B ecommerce are derived from investigating sales revenue from some of the largest electronic commerce retailers and sampling revenue from some of the smaller ecommerce companies. For example, the quarterly reports from U.S. Census Bureau's "eStats" that measure the size of electronic commerce is partially derived from the quarterly reporting of publicly-traded firms with quarterly revenue that is often measured in the millions of dollars. Every five years, Census is able to supplement these statistics with a more in-depth analysis of smaller companies that may only have $1 million in revenue per year. Since this more in-depth data gathering and analysis process takes longer, the 2002 data was just released in 2007. Therefore, the timeliness as well as the ability to capture sales information from ecommerce activities from firms of less than $1 million in revenue is a concern.

The concern about ability to capture sales data from small firms is heightened when one considers the increased ability for small Internet retailers to enter the market. A small retail location or individual may use a site such as eBay or Amazon.com to sell goods and services and make a profit with little or no fixed costs. The result of large entry by small firms and individuals into the seller marketplace increases the problem of underestimating ecommerce sales by the methods used by the U.S. Census Bureau.

Private surveys of ecommerce have similar flaws. The 2008 *Internet Retailer* list of the Top 500 ecommerce firms shows Amazon.com as the leading retailer, with $14.8B in 2007 revenue, but this figure is a combination of sales by Amazon's retail subsidiary and service fees from sales by other, smaller sellers, through Amazon's retail platform. That is, all of Amazon's own sales are counted, but only the service fees on sales by Amazon's seller customers, which number over one million. The actual sales by these small sellers are not counted.

Moreover, as Amazon pointed out in recent congressional testimony, pure platform and search service providers like eBay and Google don't even make the *Internet Retailer* roster, in spite of the fact that, in the same year (2007) that Amazon had $14.8 billion in revenues, eBay transacted more than $59 billion through its site (gross merchandise volume). In other words, although the top listed retailer (Amazon) had less than $15 billion in revenue shown, little if any of the online sales through eBay – four times that amount – are captured by the *Internet Retailer* list simply because the sales through eBay are mostly by small volume sellers. Also missing are other billions of dollars transacted by small sellers online with the help of service providers such as Google and Microsoft.

In this paper, we introduce an alternative methodology that uses consumer-based ecommerce transactions. By examining the activities of the consumers and, hence, the demand-side of the market, we are able to understand how consumers purchase online

from different retailers and individuals without being tied to sales as measured from the supply side of the market.

2. Research Methodology

Our approach to understanding the size of Internet retailing comes from understanding the perspective of an Internet retailer. This perspective helps answer the question, what would an Internet retailer with this level of scale (as measured by annual revenue) do to participate in ecommerce? Our research has divided these firms into three categories:

I. Firms that have significant scale and are likely to invest heavily in technology.
II. Firms that can invest in technology and operate an independent web site but information technology is not core to their operations.
III. Firms that rely on information technology services provided by other firms.

As we identify firms in each of the three categories and rank them from largest to smallest, we would expect an exponential distribution. In other words, the firm with the highest sales rank will contribute the most to Internet retailing. As the sales rank increases, the sales contributed by each firm exponentially decays and approaches the x-axis. This exponential distribution is often referred to as a "long-tailed" distribution. In our analysis, we have decided to examine the long tail based on cumulative sales instead of incremental sales on the y-axis. We have also log transformed the x-axis in order to shrink the long-tailed distribution into a linear representation of this distribution. One important benefit of this transformation of the x-axis is that firms in the three categories identified above have approximately the same area or region size on long-tail graph. It is our hope that this relationship can then be plotted as shown in Figure 1.

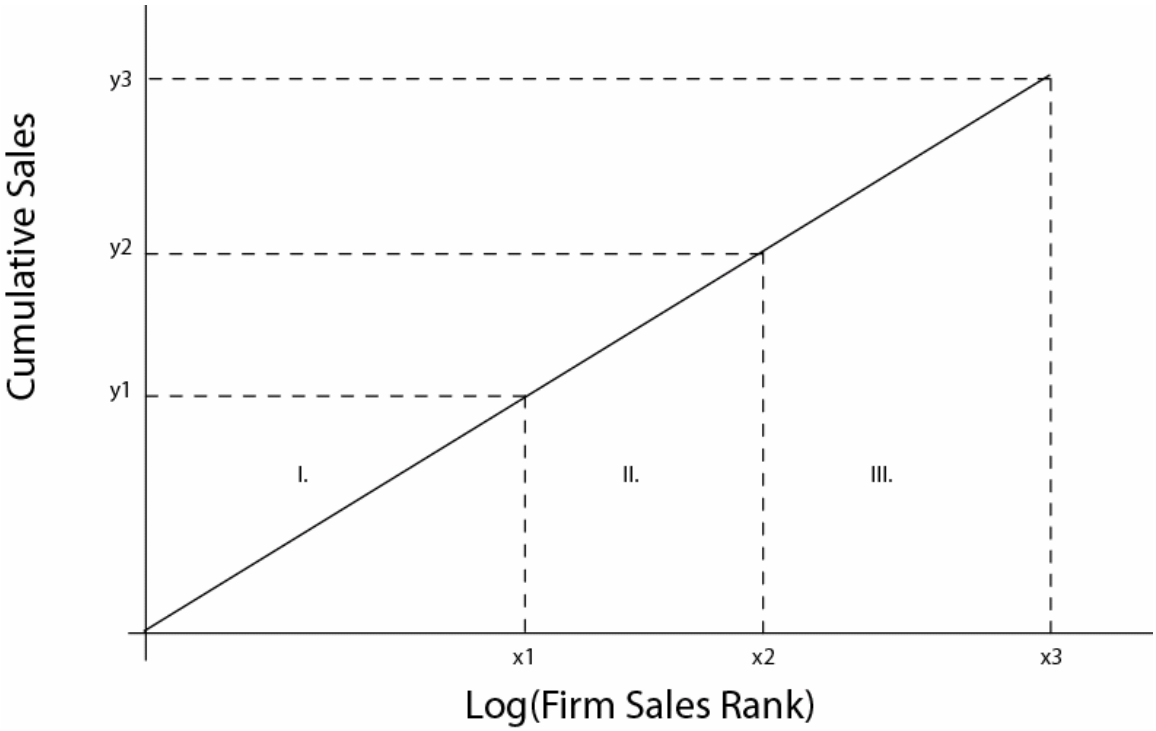

Figure 1. Model of Cumulative Sales vs. Log(Firm Sales Rank)

The goal of our research is to better estimate six variables shown in the graph in Figure 1: x_1, x_2, x_3, y_1, y_2, and y_3. Because our estimates are generated from data sources, sampling, and estimation, it is important for us to provide appropriate confidence intervals with each of these variables. Furthermore, since x_2 depends upon x_1, and x_3 depends upon both x_1 and x_2, it is important for us to understand how variance in Region I affects the variance in subsequent Regions for x_2, x_3, y_2, and y_3. Our analysis begins with some general comments about the nature of firms and data availability for Regions I, II, and III.

2.1. Region I

Region I consists of the largest Internet retailers. Fortunately, there is a significant amount of information about these retailers. In this region we use the information from the *Internet Retailer* Top 500 list. As is shown in Figure 2, the actual relationship between cumulative sales and the log of the firm sales rank (shown in blue) fits very well to a linear relationship (shown in red) with an adjusted r^2 value of 94.9%. Given this information, we are able to estimate $x_1 = 500$ and $y_1 = \$55.7$ billion.

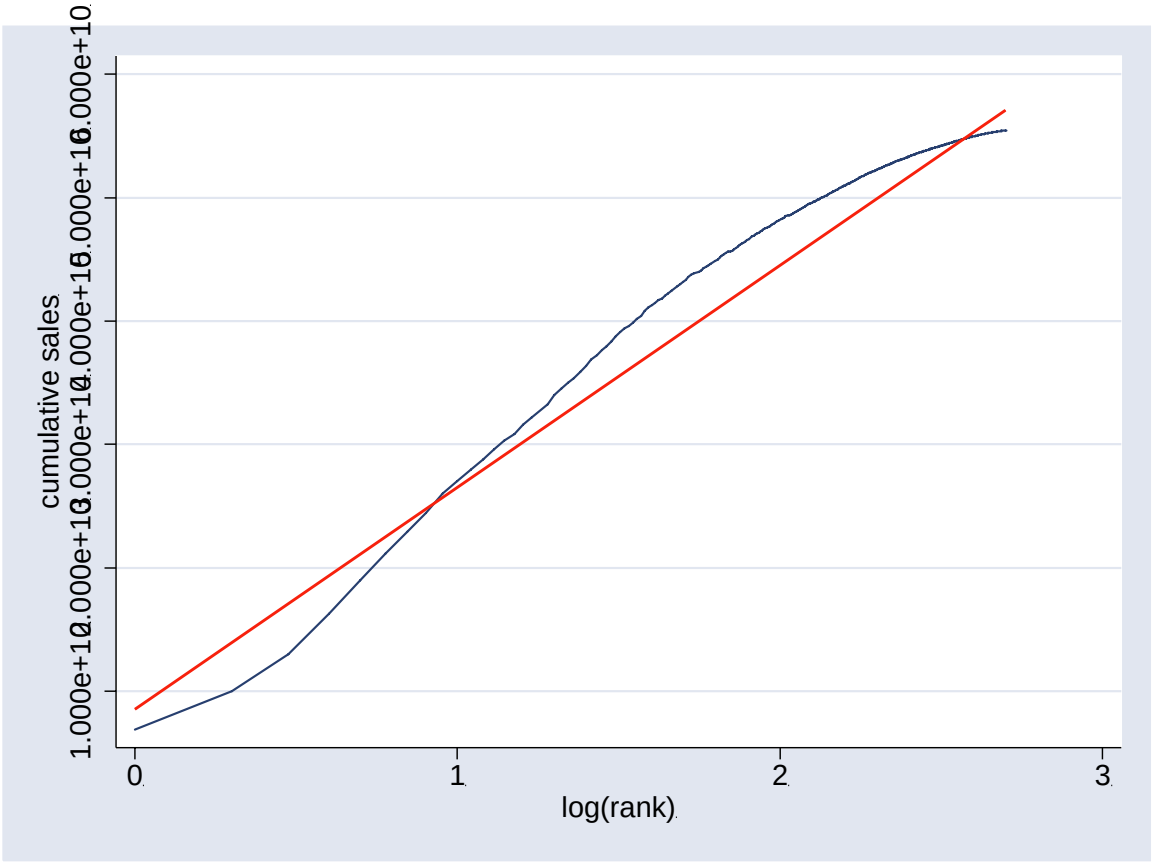

Figure 2. Region I Firm Cumulative Sales vs. Log(Rank)

2.2. Region II

Region II consists of medium-sized Internet retailers that have their own web site but are not captured by the *Internet Retailer* Top 500 directory. These companies have their own web sites and have annual revenue between $1 million and $10 million dollars. Since we were able to determine a representation of the relationship in Region I, we can use this information to extrapolate to Region II.

Based on the relationship between cumulative sales and the log of the firm sales rank , we can calculated the rank of the company which has $1million in revenue. Our estimates indicate that there are approximately an additional 28,128 firms in Region II. Therefore, x_2 = 28,628 (the sum of the number of Region I and Region II firms). We can then use this value along with the fitted curve shown by the red line in Figure 2 to estimate y_2. Based on our formula, our estimate is that y_2 = 85.6 billion.

2.3. Region III

Finally, in Region III, retailers and individual sellers are too small to measure activities at a unique web site. For example, a small business or individual may sell a few thousand

dollars worth of merchandise on eBay but would not set up a web site to do so. Within the comScore data, customer purchases from Region III players cannot be aggregated to the seller's website because they do not have one. Fortunately, it may not matter when estimating y_3 since that only depends upon an overall size of ecommerce transactions.

We begin by investigating how many dollars were spent by a sample of 52,028 users who made purchases from Region I companies. This involves aggregating the purchases from these users into the firms we identified in Region I from the *Internet Retailer* list. Then, we use this data and compare our sample to the known revenue of those Region I firms. For example, in 2004 Target.com has an online sales of $756.1 million, while the total purchases from Target.com in the comScore sample is $57,254.92 , resulting a sales ratio of 13,205[1]. In Figure 3 we depict the average of the sales ratios, based on the top companies in the *Internet Retailer* list from top 50 up to top 140.

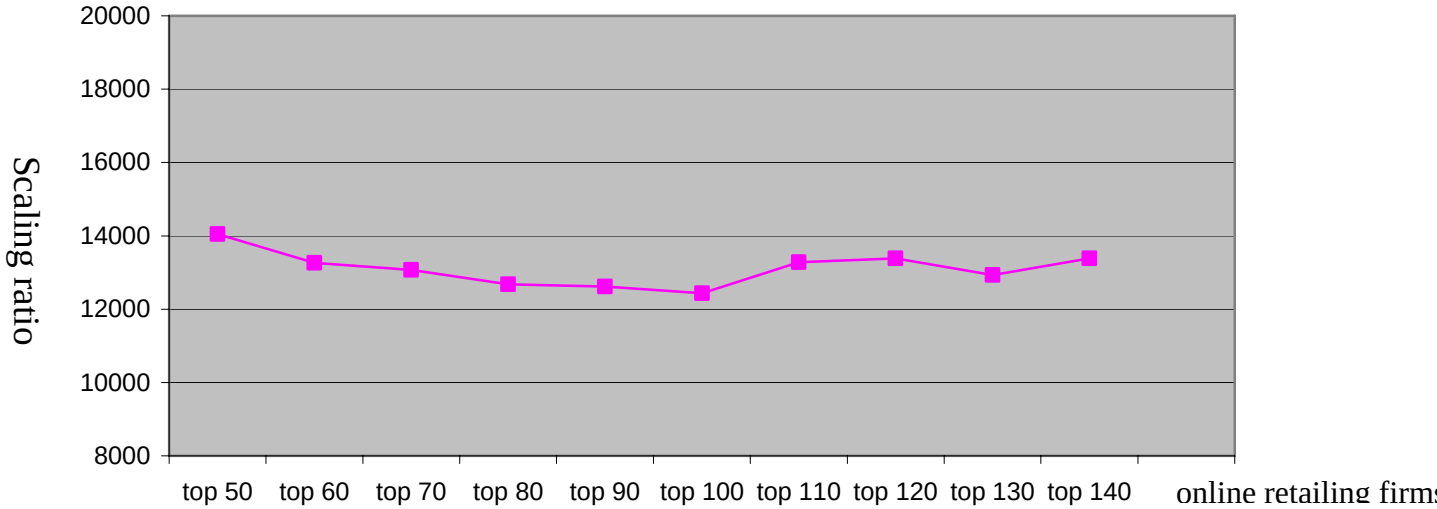

Figure 3. comScore Sampling and Scaling Ratio from Region I

Figure 3 is important because it shows that there appears to be a relatively constant scaling ratio from this sample which can then be used beyond Region I.

We use the comScore data as a sample along with the estimated scaling ratio from Figure 3 to estimate the total sales for ecommerce transactions. Since Figure 3 shows some variability in the data, we use a lower bound scaling ratio of 12,433. We also use bootstrapping to get 95% confidence interval from comScore transaction data, which is [$12.2, $12.9] million dollars. Based on this, the overall magnitude of 2004 sales would

[1] 7561000000/57254.92 = 13205.68

be between $151 billion and $160 billion. We use the conservative estimate of $151 billion as our value for y_3.

Finally, x_3 is estimated at approximately 5 million. This number comes from publicly-available eBay data that discloses the number of sellers as well as the number of transactions it supports. Therefore, a summary of our estimates may be found in Figure 4.

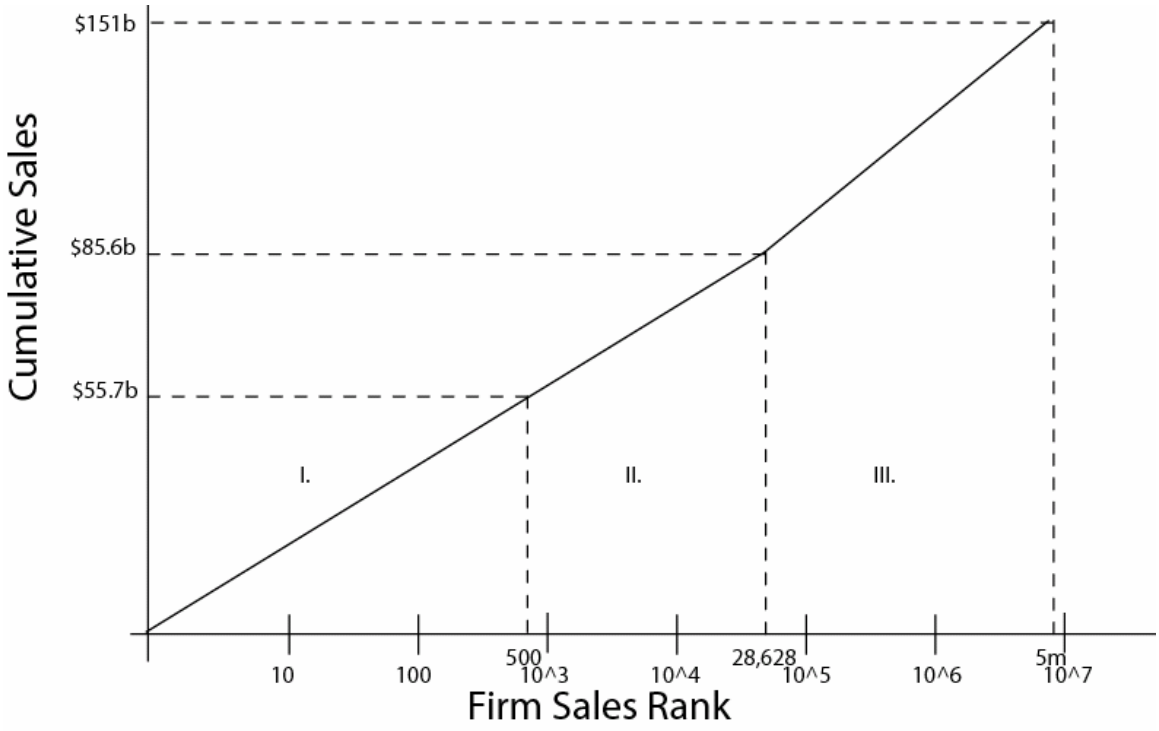

Figure 4. Cumulative Sales vs. Log(Firm Sales Rank)

3. Conclusions

Our estimates of the magnitude of ecommerce indicate that the U.S. Census Bureau and private surveys have underestimated the participation of small volume sellers – and the overall size of the electronic commerce in 2004 by as much as $30 billion. Our approach of using consumer purchasing information that is not tied to sellers with a certain sales volume has the benefit of not overlooking the small retailers and individuals who participate in electronic commerce. In other markets where the number of participants in Region III is small, ignoring these participants may not be a problem since their total sales are too small relative to the whole. However, on the Internet, the number and resulting sales of these participants is too large to ignore. When millions of small ecommerce participant sales are measured, it can lead to significant changes in the estimates.

Given the booming "long tail" phenomenon in recent years, we expect that the omitted part will be even bigger. In the near future, we plan to estimate the size of U.S. ecommerce by utilizing the 2007 data.
