Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales

Goodbye Pareto Principle, Hello Long Tail:
The Effect of Search Costs on the Concentration of Product Sales

This version: January 2011

Erik Brynjolfsson*, Yu (Jeffrey) Hu **, Duncan Simester***

ABSTRACT

Many markets have historically been dominated by a small number of best-selling products. The Pareto Principle, also known as the 80/20 rule, describes this common pattern of sales concentration. However, information technology in general and Internet markets in particular have the potential to substantially increase the collective share of niche products, thereby creating a longer tail in the distribution of sales.

This paper investigates the Internet's "Long Tail" phenomenon. By analyzing data collected from a multi-channel retailer, it provides empirical evidence that the Internet channel exhibits a significantly less concentrated sales distribution when compared with traditional channels. Previous explanations for this result have focused on differences in product availability between channels. However, we demonstrate that the result survives even when the Internet and traditional channels share exactly the same product availability and prices. Instead, we find consumers' usage of Internet search and discovery tools, such as recommendation engines, are associated with an increase the share of niche products. We conclude that the Internet's Long Tail is not solely due to the increase in product selection but may also partly reflect lower search costs on the Internet. If the relationships we uncover persist, the underlying trends in technology portend an ongoing shift in the distribution of product sales.

We thank Chris Anderson, Jerry Hausman, Lorin Hitt, Hank Lucas, Jiwoong Shin, Hal Varian, Pai-Ling Yin, seminar participants at Carnegie Mellon University, New York University, Purdue University, University of Minnesota, University of Pennsylvania, Yale University, the Workshop on Information Systems and Economics (WISE), and the American Economic Association Annual Meeting (AEA) as well as the review team at Management Science *for valuable comments on this paper. Generous funding was provided by the MIT Center for Digital Business and NSF Grant IIS-0085725.*

* MIT Sloan School of Management and the National Bureau of Economic Research; email: erikb@mit.edu; web: http://digital.mit.edu/erik

** Purdue University, Krannert School of Management; email: yuhu@purdue.edu

*** MIT Sloan School of Management; email: simester@mit.edu

1. Introduction

Many markets have traditionally been dominated by a few best-selling products. Book sales are concentrated in a relatively small number of titles by established best-selling authors (Greco 1997); Billboard "top 40" hits account for the lion's share of radio playlists and music sales; and movie rentals are dominated by a few "new releases". Economists and business managers often use the Pareto Principle to describe this phenomenon of sales concentration. The Pareto Principle, which is sometimes called the 80/20 rule, states that a small proportion (e.g., 20 percent) of products in a market often generate a large proportion (e.g., 80 percent) of sales.[1] However, the Internet has the potential to shift this balance. Anderson (2004) coined a term—"The Long Tail"—to describe the phenomenon that niche products can grow to become a large share of total sales. On the Internet, the Pareto Principle may be giving way to the "Long Tail".

Anecdotal evidence suggests that Internet markets have helped shift the balance from a few best-selling products to niche products that were previously obscure. For example, Frank Urbanowski, Director of MIT Press, observes that the increased accessibility to backlist titles through the Internet has resulted in a 12% increase in sales of these titles (Professional Publishing Report 1999). This increase happened despite flat growth in overall book sales. Similar observations have been made in electronic markets for music, DVDs, and electronics. Rhapsody, an online music provider, streams more songs each month beyond its top 10,000 than it does its top 10,000. While "new release" movies account for a dominant share of revenue in a video rental shop, DVDStation, a company that allows consumers to search and reserve movies online and pick them up in a DVD kiosk, reported that more than 50% of their rental revenue came from titles that are not new releases (DVDStation 2005).

Two basic explanations have been offered for the Internet's Long Tail phenomenon (Brynjolfsson, Hu, and Smith 2006). The first explanation focuses on the supply side. The Internet channel can carry a much larger product selection than traditional retail channels. For example, Brynjolfsson, Hu and Smith (2003)

[1] The 80/20 rule was first suggested by Vilfredo Pareto in his study of wealth distribution (Pareto 1896), and has since been applied to the analysis of city population, product sales, and sales force management.

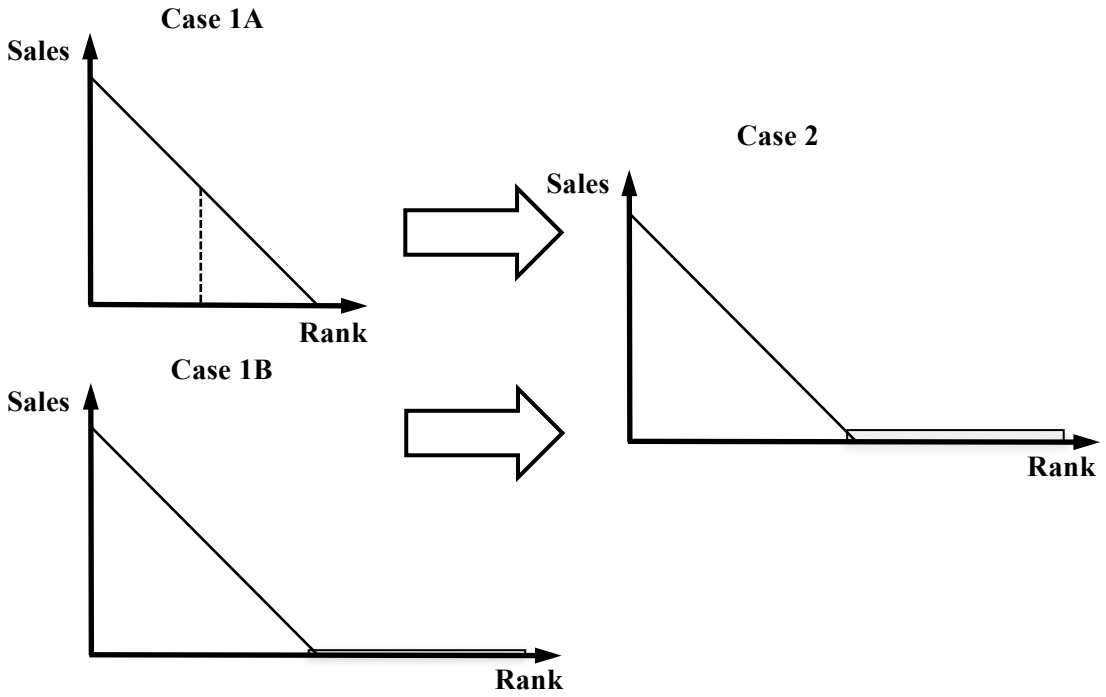
document how centralized inventory and drop-shipping agreements allow online book retailers to offer over 2 million book titles (and millions more used and out-of-print titles). In contrast, physical space restrictions, logistics, and holding costs limit the product selection in a typical brick-and-mortar book store to between 40,000 and 100,000 titles. By increasing the supply of niche products that are unavailable through traditional channels, Internet commerce may boost the share of sales generated from these niche products, leading to a Long Tail.

The second explanation centers on the demand side. The Internet channel's ability to allow consumers to acquire product information with greater convenience and at lower costs leads to increased *demand* for niche products. Many offline book shoppers do not search deeply, simply because of the inconvenience of locating a niche product in a big-box store with thousands of products on its shelves. Many catalog shoppers do not venture beyond a few popular products, even though more obscure products are available by going through the catalogs carefully or by calling sales representatives over the phone. In contrast, a retail website provides consumers with IT-enabled search, discovery tools and recommendation systems, lowering consumers' costs of acquiring product information.

Typically these two explanations exist concurrently, making it difficult to disentangle them. For instance, a firm can respond to the Internet's lower information costs by increasing its product selection on the Internet. To make matters worse, although concentration is a popular measure of the importance of the Long Tail, the effect of product availability on the concentration of product sales may be non-monotonic. A moderate increase in production selection may lead to a less concentrated distribution of product sales; but if the market is flooded by a large number of products that have minimal sales, product sales can actually appear to be more concentrated even if the sales don't change for any of the previously existing products. The comparison of Case 1A and Case 2 in Figure 1 provides a simple illustration of this.[2] As a result, the concentration of product sales can be a misleading indicator of changes in the importance of niche products, especially when the number of available products changes.

[2] See Brynjolfsson, Hu and Smith (2010) for a related discussion of this issue.

Figure 1: Concentration Can Be a Misleading Indicator of the Importance of the Long Tail



Case 1A: 100 products are available and the top 50% of products account for 75% of total sales.

Case 2: Add a "tail" of 100 niche products with small sales, while leaving the sales of existing products unchanged. Now 200 products are available, and the top 50% of products account for 95% of total sales.

Case 1B: Sales of the top 100 products are exactly the same as in Case 1A. The only change from Case 1A is we now consider 100 niche products that have zero sales. In this case, 200 products are available, and the top 50% of products account for 100% of total sales.

Fortunately, the concentration of product sales will still be a good indicator of changes in the importance of niche products if the number of available products does not change. The comparison of Case 1B and Case 2 in Figure 1 provides a simple illustration of this.

A unique feature of our study is the ability to control for differences in product availability. This not only allows us to distinguish the "supply-side" and "demand-side" stories proposed by Brynjolfsson, Hu and Smith (2006), but it also allows for an unambiguous interpretation of relative product concentration as a metric for the Long Tail.

The data we analyze was provided by a retailer that aims to supply consumers with an identical selection of products (and at an identical set of prices) through both the Internet and catalog channels. These two

channels use the same prices and the same order fulfillment facilities. This controls for differences in prices, sales tax policies, shipping costs, product selection, and stock outs. In addition, we pick one catalog and the products that are printed in that catalog; we then study how those same products are being purchased through the catalog and Internet channels within an identical time window. Thus, the products we study are not only available but also "visible" in the catalog channel as well as in Internet channel. Finally, a third way to control for product availability and visibility is to focus on just one channel. Later in this paper we focus on purchases from just the Internet channel and investigate how consumers' use of IT-enabled search and recommendation tools can have an effect on their tendency to purchase niche products through this channel.

We provide empirical evidence that the Internet channel exhibits a significantly less concentrated sales distribution than the catalog channel, even when supply-side factors such as product availability and visibility are held constant. Moreover, on average, niche products make up a larger percentage of products sold in an Internet order than in a catalog order. This finding persists even after we introduce controls for differences in the characteristics of customers that use each channel.

Our data allows us to directly measure how consumers use the search and recommendation tools provided by the company's website. We explore how these demand-side (or consumer-side) factors can lead to changes in consumers' purchasing patterns. We find that as consumers' use of IT-enabled search and discovery tools increases, the percentage of sales generated by niche products becomes larger. These findings are robust to using alternative definitions of niche products and using different sets of variables as controls for consumer heterogeneity.

Related Literature

Brynjolfsson, Hu and Smith (2003) first analyzed the Long Tail phenomenon on the Internet when they showed that niche products that were unavailable through conventional channels accounted for a large share of sales, and consumer welfare, on the Internet. They noted that product selection was much greater on the Internet than offline in categories ranging from books to DVDs to cameras, and introduced Pareto

Curve as a way to quantify the longer tail on the Internet. In a subsequent article (Brynjolfsson, Hu and Smith 2006), they identified supply-side explanations (i.e., product availability), along with demand-side explanations, such as search tools and recommendation systems, for the Long Tail phenomenon.

There are two recent papers that use aggregate sales data to study how the distribution of aggregate sales has changed over time in the market for video and music products. Both of these papers find some evidence of a shift toward niche products. Elberse and Oberholzer-Gee (2007) find evidence that a larger share of video sales have shifted toward niche products from 2000 to 2005. Bestselling videos as a category generate fewer sales in 2005 than in 2000, and studios sell fewer copies of a larger number of titles. Studying music sales data, Chellappa et al. (2007) find that the share of total sales generated from platinum albums has dropped from 33% in 2002 to 23% in 2006. They also show that the number of albums released doubled in this period and so overall sales became more concentrated at the top when using a relative concentration metric— the top 0.5% titles accounted for 56% sales in 2002 but 68% sales in 2006, although it should be noted that the absolute number of albums counted as part of the top 0.5% doubled in those four years. As noted above, a key feature of our paper is we study the concentration of product sales across channels and hold the number of available products fixed, which makes it easier to interpret the relative concentration metric.

Two recent papers study the effect of recommendation systems on the concentration of sales.[3] Oestreicher-Singer and Sundararajan (2006) measure how products are hyperlinked together on Amazon via its recommendation network, and show that such a hyperlinked content network can cause product sales to be more evenly distributed. In contrast, Fleder and Hosanagar (2009) use a theoretical model and simulations to demonstrate that recommendation systems that recommend products with high sales can lead to an increase in the concentration of sales.

[3] In addition, De, Hu, and Rahman (2010) study how consumers' use of search and recommendation tools can have an effect on overall Internet sales.

Our paper is also related to a growing body of empirical research that investigates how a reduction in consumer search costs on the Internet can impact competition (Brynjolfsson, Hu, and Rahman 2009), prices and price dispersion (see Brynjolfsson and Smith 2000, Morton, Zettelmeyer, and Silva-Risso 2001, Brown and Goolsbee 2002, Hann, Clemons and Hitt 2003, Clay, Krishnan, Wolff, and Fernandes 2003, Anderson, Fong, Simester and Tucker 2010). There is also a significant theoretical literature on how search costs and other types of information costs can affect price, price dispersion, entry, and product variety (see Diamond 1971, Wolinsky 1986, Anderson and Renault 1999, Bakos 1997, Cachon, Terwiesch, and Xu 2008). None of these models consider the concentration of product sales.

Structure of the Paper

In Section 2 we describe the design of our empirical analyses, before presenting the findings in Sections 3 and 4. The paper concludes in Section 5 with a review of the findings and broader implications.

2. Design of Empirical Study

The company we study is a medium-sized retailer selling women's clothing at moderate price levels.[4] All of the products carry the company's private label brand and are sold exclusively through the company's catalog channel (mail and telephone) and the Internet channel (website), with the Internet channel contributing roughly 60% of the company's sales. The retailer also has a physical store that accounts for a negligible percentage of overall sales. Because the company is unable to adequately identify the relatively small number of consumers purchasing in its physical store, we do not have a record of purchases made by consumers in the physical store.[5] This limits our study to the Internet and catalog channels.

2.1. Product Selection

A key feature of the company is that it offers the same product selection (and prices) through its Internet and catalog channels. This policy simplifies the company's logistics and ordering processes. In addition, it avoids potential consumer dissatisfaction if consumers observe that they have paid higher prices for an

[4] The company asked to remain anonymous.
[5] Our results are robust to including or excluding consumers who live near the physical store.

item than other consumers (see for example Anderson and Simester 2010). There are differences in the ease with which consumers can acquire information across the two channels, due in large part to specific technologies like search and recommendation tools available only on the Internet. However, the company does not respond by varying the number of available products across the channels. In addition to offering the same set of products (and prices) through both channels, the company uses the same photo and product description in both channels. It also uses the same order fulfillment methods and facilities for the two channels. This controls for differences in sales tax policies, shipping costs, and the possibility of stock outs, eliminating several alternative explanations for potential differences in the sales distribution across the two channels.

2.2. Catalog Channel

Consumers ordering through the catalog channel place their orders either by calling the company's toll-free number and speaking to a service representative, or by completing the physical order-form bound into the middle of a catalog and mailing it to the company. The vast majority of orders through the catalog channel are made over the telephone.

The company sends out catalogs to its consumers every four or five weeks, although they do not send out any catalogs in December. A typical catalog contains a few hundred products. The majority of the purchases made through the catalog channel are on products printed in the current catalog. Consumers occasionally purchase products that are not printed in the current catalog, by calling the company's sales representatives and by using past catalogs. From the company's own analyses, the impact of a current catalog lasts for about four weeks, and it diminishes greatly after the next catalog is mailed.

It seems likely that the products printed in a catalog are more visible than products not printed, and such variations in visibility could alter sales and sales patterns. Our approach to controlling for such variations in visibility is straightforward—we pick a catalog and study only the products printed in that catalog during the time period when that catalog is the current catalog. To do so, we first obtain the mailing schedule of the company's catalogs in Fall 2006, as shown in Table 1.

Table 1: Catalog Mailing Schedule

Catalog Name	Date
August Catalog	August 16, 2006
September Catalog	September 13, 2006
October Catalog	October 18, 2006
November Catalog	November 15, 2006

We then obtain the exact list of products in the August catalog, and focus on how consumers purchase those products printed in that catalog during the period between August 16, 2006 and September 12, 2006.

2.3. Internet Channel

Consumers can also visit the company's website and place their orders through the company's Internet channel. In 2006 Internet sales accounted for approximately 60% of the company's sales. To ensure there are no differences in product availability, we only study how consumers purchase via the Internet channel those products printed in the August catalog between August 16, 2006 and September 12, 2006. This approach allows us to tightly control for variations in supply-side factors such as product availability and visibility.

The company offers one version of their website to all visitors. A visitor to the company's website has several options. A website visitor may simply browse through the products that are available under each product category, moving from one product to the next one. Each product page shows a picture of a model wearing the product, as well as the price, available sizes, and colors. The process of browsing through available products on the website is similar to the process of browsing through available products in a catalog.

However, a visitor may use the website's more advanced features, such as a search function and a recommendation system. When a visitor use the search function to search for a specific product, with either its SKU or its product name, the website takes the visitor directly to the product page of that specific product. On the other hand, if a visitor searches with a non-specific keyword, the website presents a long list of relevant products that match the search keyword. In addition, when a visitor views a

product page, the website always recommends five other products that the retailer feels the visitor "may also like". These five products are picked by the company's experts based on their similarity and relevance to the focal product, and then recommended to all visitors. We note that this type of semi-personalized recommendation system is widely adopted by many Internet retailers on their product pages. If a consumer responds to the recommendation system by clicking on one of the recommended products, she will be taken to the page of the clicked product.

2.4. Hypotheses

Next, we will formulate our hypotheses on how the Internet channel's unique search and recommendation features can lead to changes in the purchasing patterns of consumers. Standard models of search behavior in the marketing and economic literatures predict that consumers search for information to improve their purchasing decisions (Stigler 1961, Engel et al. 1996, Kotler 2002). Consumers first conduct internal searches by scanning their memory and retrieving products for which they have ex ante awareness; when internal searches prove inadequate, they then decide to acquire additional information from external sources (Engel et al. 1996).

More importantly, rational consumers continuously weigh expected benefits against search costs and will stop searching whenever expected benefits are lowers than search costs (Stigler 1961, Rothschild 1974). Therefore, in an extreme case where consumers can search for product information at zero costs, consumers will exhaustively search for all available products and the distribution of sales across products will fit closely to consumers' "true" preferences. In the other extreme case where search costs are prohibitively high, consumers will conduct no external searches; as a result, consumers' consideration sets will be limited to products for which they already have ex ante awareness and the sales distribution will be extremely concentrated on such products. Between these extreme cases, consumers will search for a subset of all available products; as search costs decrease, consumers will conduct more searches (Rothschild 1974) and the sales distribution will fit closer to consumers' true preferences, being less concentrated on products for which consumers have ex ante awareness.

In the clothing industry, consumers are more likely to be aware of popular products than niche products. This is because consumers could obtain information of popular products from external sources, including fashion magazines and other media, which play a key role in setting the current trend for popular designs (Agins 1999). In addition, since each clothing retailer would offer its own version of those popular products that are consistent with the fashion trend (Rantisi 2002), consumers could gain ex ante awareness of the focal retailer's popular products from previous shopping experiences at competing retailers. Finally, ex ante awareness could come from firm advertising and word-of-mouth referrals, which tend to favor popular products (Frank and Cook 1995).

Compared with the catalog channel, the Internet channel provides unique search and recommendation tools. These tools can lower consumers' search costs (Alba et al. 1997). As discussed above, lower search costs lead to a sales distribution that fits closer to consumers' true preferences and is less concentrated on products for which consumers have ex ante awareness. Because consumers are more likely to be aware of popular products than niche products, we hypothesize that the Internet channel should have a sales pattern that places less weight on popular products and more weight on niche products, compared with the catalog channel.

Each of the three search and recommendation tools may affect the sales pattern of the Internet channel in different ways. On one hand, Internet consumers can actively perform specific searches by searching for product SKUs or exact product names. This type of specific searches, or "directed searches" as called by Moe (2003), take consumers directly to the product page that displays the product being searched for, helping consumers quickly locate a product which they have already been aware of. As a result, the use of directed searches may not lead to a sales pattern with more weight on niche products for the Internet channel.

On the other hand, consumers may use the search function to perform non-specific searches by typing in a keyword that is not a product SKU or an exact product name. Such searches are called "non-directed searches", and they typically lead to a list of products. Furthermore, while viewing a product page, the

website will display five products that are related to the focal product. If a consumer clicks on one of the products being recommended, the recommendation system enables the consumer to explore and discover products that she otherwise may not have found and, hence, provides new information (Mobasher et al. 2001). The recommendation system, along with the feature of non-directed searches, lowers the costs incurred by consumers when searching for additional information (Brynjolfsson et al. 2006). We hypothesize that the use of the recommendation system and non-directed searches should lead to a sales pattern with more weight on niche products for the Internet channel.

These are not the only possible outcomes. For example, sellers may wish to push certain products,via the Internet channel, the catalog channel, or both. In particular if a seller prefers to steer consumers toward popular products, the recommendation engine could be tuned to disproportionately favor such products. Of course, the opposite is also possible, with niche products being favored, thereby leading to a longer tail. Because of these possibilities, theory alone cannot predict whether improvements in search or recommendation technologies will lead to a longer tail. Ultimately, this is an empirical question.

2.5. Cross Channel Transactions

Consumers can find products they like from the catalog channel, and then visit the company's website and order the same items from the Internet channel. The industry wisdom is that this certainly occurs. It is also true that the reverse may occur: consumers may identify products on the company's website and then place an order over the telephone. We will present evidence that there is a longer tail on the Internet, compared to the catalog channel. Notice that this outcome cannot be explained by consumers obtaining information in one channel and then ordering through another channel. Indeed, to the extent this occurs it will *reduce* the differences between the channels and undermine our ability to detect any difference in the sales concentration across the two channels.

3. Empirical Results

Our empirical analyses focus on two questions:

1. Does the Internet channel exhibit a less concentrated distribution of product sales?

2. Do demand-side factors, such as consumers' usage of recommendation and search tools, have an effect on the sales of niche products on the Internet?

As Figure 1 illustrates, it is important to control for product availability when we examine how the concentration of product sales varies across different channels. Next we describe how we have controlled for product availability.

3.1. Controlling for Product Availability

In principle, customers ordering through the catalog channel are not limited to the items included in the most recent catalog; they may also order other items that they have seen in prior catalogs or that are suggested by the customer service representative. The same product selection is available in the Internet and catalog channels. However, the majority of items ordered through the catalog channel are for items included in the most recent catalog. Therefore, we will adopt a conservative definition of which products are "available" by restricting attention to only those products that were included in the most recent catalog. This ensures that any evidence that there is a longer tail of products on the Internet than in the catalog channel cannot be attributed to some items not appearing in the catalog.

The retailer's August 2006 catalog included a total of 734 products. This catalog was sent to consumers on August 16, while the next catalog was mailed on September 13 and so we will focus on purchases made between August 16 and September 12. This yields a sample of 26,686 units purchased over 12,081 Internet orders and 18,663 units purchased over 6,905 catalog orders.[6]

3.2. Aggregate-level Analyses

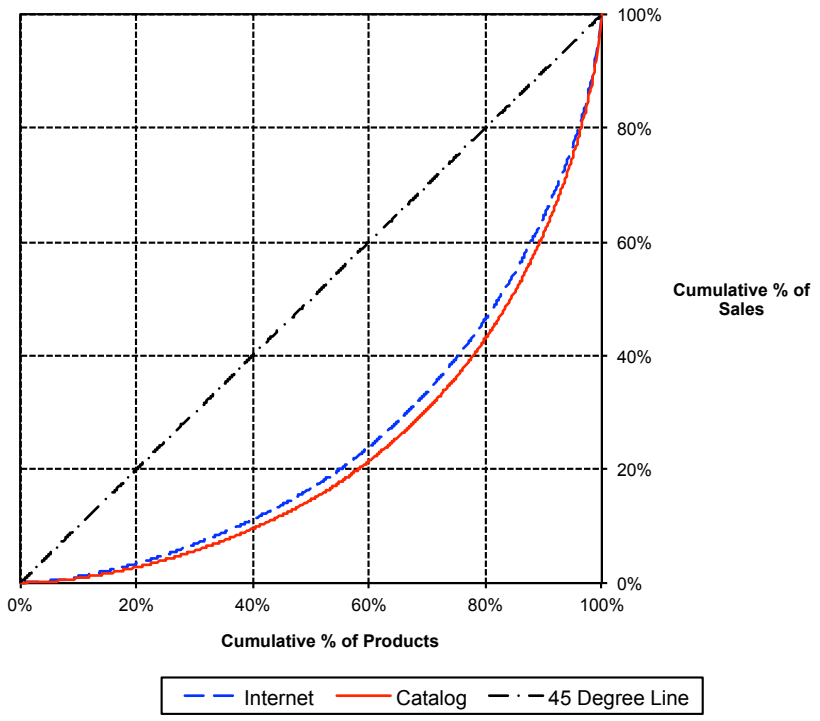
We first calculate the aggregate sales for each of the 734 products in each channel between August 16 and September 12. We then use the Lorenz Curve and Gini Coefficient to study the concentration of product sales in each channel.[7] Figure 2 presents two Lorenz Curves, the blue one for the Internet channel and the

[6] As we will discuss, the results are replicated when we use the products in the September catalog and focus on the time period between September 13 and October 17.

[7] Economists have long used the Lorenz Curve and Gini Coefficient to describe the inequality in income and wealth distribution (Lorenz 1905, Gini 1912). This paper is among the first to apply these two concepts to measure the

red one for the catalog channel. The Internet channel's Lorenz Curve lies above the catalog channel's Lorenz Curve, implying that the Internet channel exhibits a less concentrated distribution of product sales than the catalog channel. Correspondingly, the Gini Coefficient for the Internet channel (0.49) is lower than that for the catalog channel (0.53). From the Lorenz Curve, one can easily obtain the percentage of total sales generated by the bottom 80% products. For the catalog channel, the bottom 80% products generate 43% of sales; for the Internet channel, the bottom 80% products generate 47% of sales.

Figure 2: Lorenz Curve and Gini Coefficient for the Internet and Catalog Channels



Using the Lorenz Curve and Gini Coefficient, we have shown that these exists a difference between the sales distribution in the Internet channel and that in the catalog channel. However, these two tools do not allow us to conclude whether such a difference is statistically significant. In order to do so, we then fit the

---

concentration of product sales. The Lorenz Curve is drawn inside a square box with the x-axis being cumulative percentage of products and the y-axis being the cumulative percentage of sales. The Gini Coefficient is the ratio of the area between a Lorenz Curve and a 45 degree line to the total area under a 45 degree line. When sales are perfectly evenly distributed among products, the Lorenz Curve coincides with a 45 degree line and the Gini Coefficient equals zero. As the distribution becomes more concentrated, the Lorenz Curve curves away from a 45 degree line and the Gini Coefficient increases.

sales and sales rank data to the following log-linear relationship and compare the *Sales Rank* coefficient obtained when using data from the two channels:

$$\ln(Sales_j) = \beta_0 + \beta_1 \ln(SalesRank_j) + \varepsilon_j . \quad (1)$$

The *Sales Rank* is an ordinal ranking of the frequency with which each item was purchased, and the log-linear curve described by Equation (1) is known as a Pareto Curve. Previous research has shown that it fits the relationship between product sales and sales rank very well across the full distribution of products (Brynjolfsson, Hu and Smith 2003). This curve has also been used to successfully describe the distribution of income, wealth, and city size (Pareto 1896, Zipf 1949). Given this specification, β_1 measures how quickly product *j*'s demand in a channel falls as the sales rank increases. If the Internet channel has a longer tail, then β_1 would be less negative (i.e. lower in absolute value) in the Internet channel than in the catalog channel, indicating that products with large sales ranks retain a larger share of demand in this channel.

We first estimate Equation (1) separately for the Internet and catalog data and report both sets of findings in Models 1 and 2 of Table 2. Both coefficients are highly significant, while the high R^2 values suggest that the log-linear relationship fits the data well. The β_1 coefficient is -0.925 for the Internet data, and -0.877 for the catalog data. Next, we test whether the β_1 coefficient in Equation (1) is significantly less negative for the Internet channel than for the catalog channel. To do that, we pool Internet and catalog data into one data set and run a single regression. We create an "Internet" dummy indicating whether an observation is for the Internet channel, and interact the "Internet" dummy with $\ln(SalesRank_j)$:

$$\ln(Sales_j) = \beta_0 + \beta_1 \ln(SalesRank_j) + \beta_2 Internet_j + \beta_3 Internet_j * \ln(SalesRank_j) + \varepsilon_j . \quad (2)$$

Estimates for Equation (2) are reported in Table 2 (Model 3). The β_3 coefficient on the interaction term is positive and highly significant, indicating that the original β_1 coefficient in the model of Equation (1) is significantly less negative for the Internet channel than for the catalog channel. To check the robustness

of these results, we re-estimate Equation (2) using Quantile Regression and report the results in Table 2 (Model 4). Quantile regression relates the conditional median of the dependent variable to independent variables and is more robust to outliers than linear regression. The β_3 coefficient on the interaction term remains positive and highly significant.[8] We conclude that the Internet channel has a significantly less concentrated distribution of sales (a longer tail) than the catalog channel.

Table 2: Pareto Curve Estimates

	Model 1: Internet Data	**Model 2: Catalog Data**	**Model 3: Pooled Data, Linear Regression**	**Model 4: Pooled Data, Quantile Regression**
Constant	8.126*** (0.084)	8.002*** (0.081)	8.002*** (0.083)	8.519*** (0.077)
Sales Rank	-0.877*** (0.015)	-0.925*** (0.014)	-0.925*** (0.015)	-0.991*** (0.014)
Internet			0.124 (0.117)	0.140 (0.109)
Internet * Sales Rank			0.048** (0.021)	0.045** (0.019)
Adjusted R^2	0.830	0.852	0.848	0.672
Sample Size	733	728	1,461	1,461

Models 1 and 2 present the coefficients from Equation 2 estimated using OLS. Model 1 uses sales in the Internet channel and Model 2 uses sales in the catalog channel. Model 3 presents the OLS coefficients when estimating Equation 3 using the pooled data from Models 1 and 2. Model 4 re-estimates Model 3 using Quantile Regression rather than OLS. Standard errors are in parentheses; ***Significantly different from zero, $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

We further check robustness by repeating the analysis using products and sales from the September catalog. In particular, we use the 441 products printed in the September catalog and focus on the time period between September 13 and October 17. Detailed findings are reported in the Online Appendix, Tables A4-A7.

[8] We can also compare the concentration of sales in the two channels using quantile-quantile plots. We report these results in the Online Appendix, Figures A1. The plot starts as an almost straight line and then curves downward. This is consistent with the findings reported for the Pareto Curve. We thank an anonymous reviewer for suggesting the use of quantile regression and quantile-quantile plots.

3.3. Order-level Analyses and Controlling for Consumer Selection via Sample Matching

The analyses in Section 3.2 demonstrate that the Internet has a longer tail than the catalog channel. One could argue that consumers who purchase through the Internet channel could be systematically different from consumers who purchase through the catalog channel. This consumer selection effect could have confounded the results shown in Section 3.2. We use the propensity score matching method suggested by Rosenbaum and Rubin (1983) to control for this consumer selection effect.

Before matching the two samples we first identify "niche" products as products that are purchased infrequently. Specifically, for each channel, we rank products by their aggregate sales and define the bottom 50% of products as "niche products". Table 3 compares the unit sales and price across these two types of products. On the Internet, an average niche product sells just 12.1 units, while the other products average 60.6 units sold; through the catalog channel, an average niche product sells 7.4 units, while the other products average 43.4 units sold. We note that, in the Internet channel, the top 50% products has an average price of $30.80, while the average price for the bottom 50% products is $32.59; the difference has a t-statistic of 1.48 and is not statistically significant. Similarly, in the catalog channel, the average price is $31.78 for the top 50% products and $29.60 for the bottom 50% products; the difference has a t-statistic of 1.75 and is not significant.

Table 3: Niche Products

	Average Unit Sales	Average Price
Internet Channel		
Top 50%: 367 most frequently purchased products	60.6	$30.80
	(2.3)	(0.70)
Bottom 50%: 367 least frequently purchased products	12.1	$32.59
	(0.3)	(1.00)
Catalog Channel		
Top 50%: 367 most frequently purchased products	43.4	$31.78
	(1.9)	(0.76)
Bottom 50%: 367 least frequently purchased products	7.4	$29.60
	(0.2)	(1.00)

Standard errors are in parentheses.

The propensity score matching method suggested by Rosenbaum and Rubin (1983) matches samples of data on observable dimensions.[9] In this case we match the Internet observations (orders) with catalog observations (orders) based on the observable characteristics of consumers who made those orders. This matching approach drastically reduces the difference between the two groups of observations.[10]

We begin the analysis by matching observations using demographic and socioeconomic variables collected from the 2000 U.S. Census at the zip code level. A consumer's demographic and socioeconomic variables such as her income, age, education, and gender may influence her demand (Goolsbee 2000), and whether a consumer lives in an urban area may influence that consumer's demand (Glaeser et al. 2001). We focus on five demographic and socioeconomic variables: *Population Density* (population per square mile), *Median Household Income*, *Percentage with Bachelor's Degree*, *Percentage Female*, and *Median Age*. For observations with Canadian zip codes or other zip codes that cannot be matched to the U.S. Census data (8.0% of the sample), we create a dummy variable *No Demographics Information*.

We use the PSMATCH2 propensity score matching module in Stata to match samples (using the nearest 10 neighbors).[11] Before matching the Internet sample (orders made by consumers who purchase from the Internet channel) with the catalog sample (orders made by consumers who purchase from the catalog channel), we find that the Internet sample has a significantly higher *Median Household Income*, a significantly higher *Percentage with Bachelor's Degree*, and a significantly lower *Median Age*. However, after matching, the matched Internet sample is no longer significantly different from the catalog sample on any of the five observable dimensions. These results are shown in Table 4a.

[9] Additional details on propensity score matching can be found in Rassler (2002). We note that after sample matching, the matched samples could still differ on unobservable dimensions, which is a limitation of this approach.
[10] Very few consumers made more than one order within this time period. Thus, our results are robust to conducting either an order-level analysis or a consumer-level analysis.
[11] The results are robust to using different numbers of neighbors and different algorithms.

Table 4a: Matching the Internet and Catalog Samples

Matching Variables	Catalog Sample	Internet Sample	Matched Internet Sample
Population Density (00s)	29.2	27.6 (0.148)	29.0 (0.825)
Median Household Income ($000s)	46.2	47.4*** (0.000)	46.1 (0.896)
Percent with Bachelor's Degree	10.5%	11.1%*** (0.000)	10.5% (0.755)
Percent Female	47.1%	46.7%* (0.093)	47.0% (0.781)
Median Age	34.1	33.7*** (0.005)	34.1 (0.895)
No Demographics Information	7.8%	8.1% (0.350)	7.8% (0.954)
Sample Size	6,905	12,081	6,905

The numbers in parentheses are p-values, measuring the probability that the difference between the Internet and Catalog sample averages will be larger than the observed difference, under the null hypothesis that the true averages are identical. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

For each order we calculate the percentage of both unit and dollar sales generated by niche products. The findings are presented in Table 4b, where we compare the outcomes for the catalog sample, the Internet sample, and the matched Internet sample. Before matching, we find that the percentage of unit sales generated by niche products is significantly higher in the Internet sample than in the catalog sample: 15.2% versus 12.7%, with a t-statistic of 5.63. Similarly, the percentage of dollar sales generated by niche products is on average 15.4% in the Internet sample, versus 12.7% in the catalog sample, with a t-statistic of 6.10. These results are consistent with the aggregate-level results in Section 3.2.

After matching, the percentage of unit sales generated by niche products remains significantly higher in the matched Internet sample than in the catalog sample: 14.8% vs. 12.7%, with a t-statistic of 4.39. The

percentage of dollar sales generated by niche products is 15.0% in the matched Internet sample, versus 12.7% in the catalog sample, with a t-statistic of 4.77.[12]

We conclude that the difference in sales patterns across the Internet sample and the catalog sample persists, even after controlling for the consumer selection effect using sample matching.

Table 4b: Results Using the Matched Samples

Percentage of Total Sales Generated by Each Sample of Niche Products	Catalog Sample	Internet Sample	Matched Internet Sample
Unit Sales			
Bottom 40% (294 products)	8.2%	10.0%*** (0.000)	10.0%*** (0.000)
Bottom 50% (367 products)	12.7%	15.2%*** (0.000)	14.8%*** (0.000)
Bottom 60% (440 products)	18.9%	21.7%*** (0.000)	21.2%*** (0.000)
Dollar Sales			
Bottom 40% (294 products)	8.2%	10.3%*** (0.000)	10.3%*** (0.000)
Bottom 50% (367 products)	12.7%	15.4%*** (0.000)	15.0%*** (0.000)
Bottom 60% (440 products)	18.9%	22.0%*** (0.000)	21.3%*** (0.000)
Sample Size	6,905	12,081	6,905

The numbers in parentheses are p-values, measuring the probability that the difference between the Internet and Catalog sample averages will be larger than the observed difference, under the null hypothesis that the true averages are identical. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

3.4. Robustness Checks

Because the cutoff at the bottom 50% (or 367) seems somewhat arbitrary, we try to replicate our results using alternative definitions such as bottom 40% (or 294) products and bottom 60% (or 440) products. We find that our results are robust to using alternative definitions of this cutoff. These findings are also

[12] A shift of 2.3% (the difference between 15.0% and 12.7%) in dollar sales from other products to niche products, across both channels, would be equivalent to a shift of $1.38 million, given that the company's annual sales are roughly $60 million.

presented in Table 4b. The difference in sales patterns between the catalog sample and the matched Internet sample remains significant.

In addition to using demographic and socioeconomic variables to match samples, we also investigated other ways of constructing matched samples. In particular, we used the *Recency*, *Frequency*, and *Monetary Value* of customers' historical purchases. These measures were calculated using customers' transactions made prior to August 16, 2006.[13] These so-called "RFM" measures are widely used in both the catalog industry and the marketing literature to segment consumers. Customers whose previous purchases are more recent, more frequent, and/or include higher priced items are generally considered to be "more valuable". These results using "RFM" measures as the basis of sample matching are reported in Table 5. The pattern of findings is similar to the results in Table 4b, demonstrating that our results are robust to using a different set of observable variables in sample matching.

We also created a measure of each consumer's historical tendency to purchase niche products. To do so, we first calculate the aggregate sales for each product sold in the two years prior to August 16, 2006 and define niche products as those products that cumulatively generate 80% of the company's sales in those two years. For each consumer, we define a variable "*Historical Niche Tendency*" as the ratio of the number of niche products purchased by her to the total number of products purchased by her in those two years. For consumers who made no purchases in those two years, the dummy variable "No Historical Niche Information" is equal to one. These two new variables are added to the "RFM" measures as the basis of sample matching. We find that the pattern of findings is similar to the results in Table 5. For the sake of brevity, these results are reported in the Online Appendix as Table A3.

[13] $Recency_i$ is defined as the number of days prior to August 16, 2006 that consumer *i* made a purchase. $Frequency_i$ is defined as the number of items placed by the consumer prior to August 16, 2006. $Monetary\ Value_i$ is defined as the average price of the items in consumer *i*'s historical orders. The dummy variable *No RFM Information* is equal to one for consumers who made no purchases prior to August 16, 2006.

Table 5: Matching Using Historical Transactions

	Catalog Sample	Internet Sample	Matched Internet Sample
Matching Variables			
Recency	3.147	2.668*** (0.000)	3.119 (0.520)
Frequency	2.198	1.510*** (0.000)	2.192 (0.858)
Monetary Value	2.186	1.948*** (0.000)	2.187 (0.988)
No RFM Information	0.334	0.401*** (0.000)	0.334 (0.983)
Percentage of Total Sales Generated by Each Sample of Niche Products			
Unit Sales			
Bottom 40% (294 products)	8.2%	10.0%*** (0.000)	9.9%*** (0.000)
Bottom 50% (367 products)	12.7%	15.2%*** (0.000)	15.0%*** (0.000)
Bottom 60% (440 products)	18.9%	21.7%*** (0.000)	21.3%*** (0.000)
Dollar Sales			
Bottom 40% (294 products)	8.2%	10.0%*** (0.000)	10.5%*** (0.000)
Bottom 50% (367 products)	12.7%	15.4%*** (0.000)	15.6%*** (0.000)
Bottom 60% (440 products)	18.9%	22.0%*** (0.000)	22.9%*** (0.000)
Sample Size	6,905	12,081	6,905

The numbers in parentheses are p-values, measuring the probability that the difference between the Internet and Catalog sample averages will be larger than the observed difference, under the null hypothesis that the true averages are identical. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

3.5. Summary

Our findings confirm that sales in the Internet channel are more evenly distributed across products than in the catalog channel. This difference cannot be attributed to differences in prices or product availability. Moreover, the result survives when we account for customer differences using a propensity matching algorithm. In the next section we explore different explanations for this result by explicitly measuring customers' use of Internet search mechanisms.

4. The Role of Internet Recommendation and Search Tools

To investigate whether Internet recommendation and search tools may have contributed to the differences in these sales patterns between the two channels, we obtained the server log data recorded by the company's website server. Each month the company's website server records about 25 million lines of logs (the monthly server logs are about 20GB in size). Extracting the server log data allows us to trace each click made by a consumer while making an Internet order, and reveals the extent to which each customer used the retailer's Internet recommendation and search tools.

4.1. Measuring Consumers' Use of Recommendation and Search Tools

To measure consumers' use of search and recommendation tools we first count the total number of page requests linked to each Internet order. We then count the number of times a consumer uses the website's search tool to perform directed searches, the number of times she performs non-directed searches, and the number of times she clicks on one of the recommended products. These three measures are then normalized by the total number of page requests linked to an Internet order to produce three variables measuring the use of each mechanism: *Directed Search*, *Non-directed Search* and *Recommendation System*. Each of these three variables measures the percentage of page requests that are related to using a particular tool. This normalization procedure eliminates concerns stemming from the variation in the total number of page requests.

Table 6 provides the descriptive statistics of consumers' use of recommendation and search tools in our sample. We note that during the period between August 16 and September 12, an average consumer makes about 140 page requests when placing an Internet order. Among the 11,648 observations (orders) in our sample, 9,775 observations contain zero use of directed search; the mean of *Directed Search* for the remaining 1,873 observations is 7.6%; and the mean of *Directed Search* for all observations together is 1.2%. Similarly, 10,593 observations contain zero usage of non-directed search; the mean of *Non-directed Search* for the remaining 1,055 observations is 4.2%; and the mean of *Non-directed Search* for all observations together is 0.4%. In addition, 6,971 orders contain zero usage of the recommendation system;

the mean of *Recommendation System* for the remaining 4,677 observations is 5.1%; and the mean of *Recommendation System* for all observations together is 2.0%.

Table 6: Descriptive Statistics

Variable	Average	Standard Deviation	Minimum	Maximum	Average Non-Zero Observations
Directed Search (%)	1.2%	3.5%	0.0%	36.2%	7.6%
Non-directed Search (%)	0.4%	1.7%	0.0%	33.3%	4.2%
Recommendation System (%)	2.0%	3.9%	0.0%	42.9%	5.1%

4.2. Use of Recommendation and Search Tools and the Sales of Niche Products

We follow the same definition of niche products that is introduced in Section 3.3 (the bottom 50% of products), and for each Internet order calculate the unit sales generated by these products. We estimate a negative binomial regression model to understand how consumers' use of recommendation and search effects the unit sales of niche products on the Internet. We estimate the following model:

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

where: y_i is the unit sales of the bottom 50% of products; X_i is a vector of explanatory variables; $E(y_i \mid X_i) = \mu_i = \exp(X_i\beta + \varepsilon_i)$ is the conditional mean; ε_i is the unobserved heterogeneity that follows a log-gamma distribution. The demographic and socioeconomic variables (in natural logs) are used as controls for consumer heterogeneity, and the natural log of the total unit sales for each Internet order is also added as a control. The findings are presented in Table 7. We note that 433 Internet orders (or 3.6% of our sample) could not be matched to the server log data. Thus, the number of observations in our

sample drops slightly to 11,648.[14] The correlation matrix for all of the variables is reported in the Online Appendix, Table A1.

Table 7: Sales of Niche Products and Consumers' Use of Search and Recommendation Tools

	Bottom 40%	Bottom 50%	Bottom 60%
Directed Search	-0.166 (0.504)	-0.606 (0.406)	-0.314 (0.325)
Non-directed Search	2.605** (1.079)	1.976** (0.874)	2.021*** (0.711)
Recommendation System	1.199** (0.493)	1.386*** (0.385)	0.874*** (0.323)
Population Density	0.006 (0.014)	0.010 (0.011)	0.004 (0.009)
Median Household Income	-0.057 (0.086)	-0.002 (0.068)	0.021 (0.056)
Percent with Bachelor's Degree	0.664 (0.573)	0.108 (0.456)	0.017 (0.375)
Percent Female	2.331 (1.564)	1.188 (1.211)	0.700 (0.981)
Median Age	-0.241 (0.181)	-0.070 (0.143)	-0.065 (0.117)
No Demographic Information	-0.513 (1.205)	0.144 (0.954)	0.186 (0.782)
Total Unit Sales	1.206*** (0.031)	1.172*** (0.024)	1.174*** (0.019)
Intercept	-2.055** (1.203)	-2.317** (0.952)	-1.956** (0.780)
Pseudo R^2	0.108	0.121	0.140
Sample Size	11,648	11,648	11,648

The table reports the coefficients when estimating Equation 3 using unit sales of products through the Internet channel. The dependent variable is the units sales of the bottom x% of products, where X varies across the three models. Standard errors are in parentheses; ***Significantly different from zero, $p < 0.01$; ** $p < 0.05$; * $p < 0.10$.

The coefficient on *Directed Search* is not statistically significant, while the coefficients on *Non-directed Search* and *Recommendation System* are positive and statistically significant. This indicates that consumers' use of the company's recommendation system and the non-directed search tool both lead to

[14] Of the 11,648 observations (Internet orders) in our sample, 90 are made by consumers who also purchased from the catalog channel during this period. Our results are robust to including or excluding these 90 observations.

increased sales of niche products. On the other hand, consumers' usage of the directed search tool does not lead to more demand for niche products.

4.3. Robustness checks

The findings in Table 7 survive a wide range of robustness checks. For instance, because the cutoff of bottom 50% seems somewhat arbitrary, we replicate our results using the 40% and 60% cutoffs. These results are also reported in Table 7. The coefficients on *Non-directed Search* and *Recommendation System* remain positive and significant, while the coefficient on *Directed Search* remains insignificant.

We also investigated adding additional controls for consumer heterogeneity. In particular, we included the historical *Recency*, *Frequency*, and *Monetary Value* measures as additional control variables. The results are very similar to those reported in Table 7.

Finally, we also tried using the *percentage* of dollar and unit sales generated by niche products as alternative dependent variables. The pattern of results is again virtually unchanged (these results are reported in the Online Appendix, Table A2).

4.4. Economic Significance

We can easily assess the marginal effects of the coefficients in Table 7 and whether they are economically meaningful. As reported in Table 6, the average of *Non-directed Search* is 4.2% for consumers who engaged in non-directed search. Using the coefficient on *Non-directed Searches* in the second column of Table 7 (1.976), we find that, a change in *Non-directed Searches* from 0% to 4.2% can increase unit sales generated by niche products by 8.3% (calculated as 4.2% of 1.976).[15]

Similarly, the average of *Recommendation System* is 5.1% across consumers who used the recommendation system. Using the coefficient on *Recommendation System* in the second column of Table

[15] We note that this 8.3% increase in the sales of niche products is obtained when *Total Unit Sales* is used as a control variable. Thus, it represents a shift of sales in the amount of $0.47 million toward niche products from other products, given that niche products account for $5.6 million (or 15.2%) of the company's $37 million annual sales on the Internet.

7 (1.386), we estimate that, a change in *Recommendation System* from 0% to 5.1% can increase unit sales generated by niche products by 7.1%.

4.5. Summary

We conclude that the results in this section are consistent with a shift to a longer tail in Internet channels as consumers search more. Use of the recommendation and non-directed search tools are both associated with a significant increase in sales of niche products. In contrast, use of directed search does not lead to an increase in sales of niche products. The difference in the outcomes for directed and non-directed search is consistent with the differences in the extent to which the two mechanisms expose customers to information about products that were not already in their consideration sets. While directed search only provides information about products that customers were specifically considering, non-directed search may present customers with information about a much broader range of products.

5. Conclusions

Most markets have traditionally been dominated by a few best-selling products. However, Internet markets have the potential to increase the share of sales generated by niche products. Previous research on the Internet's Long Tail phenomenon focuses on a "product availability" explanation: the Internet channel has the ability to carry a much larger product selection than traditional retail channels, leading to an increase in the sales of niche products and a longer tail in the sales distribution. While increased product selection is undoubtedly an important driver of the Long Tail, this paper investigates alternative explanations. We control for variation in product availability, and explore whether demand-side factors are associated with increased sales of niche products via the Internet.

First, we present empirical evidence that confirms the existence of the Long Tail on the Internet: the Internet channel exhibits a significantly less concentrated sales distribution when compared with the catalog channel. The Internet channel's Long Tail phenomenon survives even though prices, product descriptions and pictures of the products are identical in both channels. We also ensure that product

"availability" is the same in both channels by restricting attention to only those products that are printed in catalogs.

Consumers who purchase from the Internet channel may differ systematically from consumers who purchase from the catalog channel. To control for this selection effect we construct an Internet sample that matches our catalog sample on multiple observable dimensions. Our finding that niche products generate a larger proportion of sales in the Internet channel than the catalog channel persists when we study these two matched samples.

We believe that lower search costs on the Internet is the most plausible explanation for these empirical results. To investigate this explanation further we directly measure consumers' use of the website's recommendation and search tools. Our analyses indicate that consumers' use of recommendation and non-directed search tools contributes to the increase in demand for niche products on the Internet. Notably, this finding does not extend to directed search, apparently because directed search is less likely to present customers with information about products that they are not already considering. We note that these recommendation and search tools are unique to the Internet channel. Moreover, the changes we detect are not only statistically significant but also economically meaningful. We note that it is possible sellers use search and recommendation tools to bias sales toward niche products. Thus, while we can rule out changes in product selection as a driver of the long tail in our setting, we cannot conclusively attribute the shift purely to lower search costs.

Our results have significant implications for the future evolution of business strategies. As companies invest in ever-more sophisticated information technologies that allow consumers to actively and passively discover products that they otherwise would not have considered, and as consumers gain more experience using these IT-enabled tools, our findings suggest that product sales will become less concentrated. The balance will continue to shift from a few best-selling products to niche products that are previously difficult to be discovered by consumers. This Long Tail phenomenon could have a profound impact on a firm's product development strategy, operations strategy, and marketing strategy.

As consumers use IT-enabled tools to discover niche products that may fit their tastes better than popular products, consumers are likely to obtain higher levels of consumer surplus. Although we do not specifically measure consumer surplus in this paper, this is an important issue for future research. Furthermore, our research shows that more experienced and more loyal consumers tend to have stronger tastes for niche products. One likely explanation is that such consumers may have already purchased popular products. Future research could study how consumers' tastes evolve over time as they make more and more purchases and whether more experienced and more loyal consumers benefits more from the Internet channel and its IT-enabled tools.

We might expect that consumers with niche tastes would be attracted to firms that not only offer a larger selection of niche products but also adopt technologies that make it easier for customers to find these products (Brynjolfsson, Hu and Rahman, 2009). In future research it would be interesting to study whether Internet firms strategically compete on product selection and the ease of discovering niche products. If competition does induce Internet firms to compete on these dimensions, we anticipate that it will tend to reinforce the long tail results studied in this paper and lead to even less concentration in product sales.

We use a set of empirical tools that can be readily applied by future researchers to measure and analyze the concentration of product sales in other research settings. Moreover, we study an important phenomenon that is emerging thanks to the unique capabilities of Internet markets. Because the underlying technological drivers are almost certain to continue to progress in advanced economies, the implications of these technologies for firm strategies and economic welfare are likely to become increasingly important.

References

Alba, Joseph, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, Stacy Wood. 1997. Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces. *Journal of Marketing* 61(3) 38-53.

Anderson, Chris. 2004. The Long Tail. *Wired Magazine*, October 2004.

Anderson, Eric T., and Duncan I. Simester. 2010. Price Stickiness and Customer Antagonism. *Quarterly Journal of Economics*, 125(2) 729-765.

Anderson, Eric T., Nathan M. Fong, Duncan I. Simester, and Catherine E. Tucker. 2010. How Sales Taxes Affect Customer and Firm Behavior: The Role of Search on the Internet. *Journal of Marketing Research*, 47(2) 229-239.

Anderson, Simon and Regis Renault. 1999. Pricing, Product Variety, and Search Costs: A Bertrand-Chamberlin-Diamond Model. *RAND Journal of Economics* 30(4) 719-735.

Bakos, Yannis. 1997. Reducing Buyer Search Costs: Implications for Electronic Marketplaces. *Management Science* 43(12) 1676-1692.

Brown, Jeffrey R. and Austan Goolsbee. 2002. Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry. *Journal of Political Economy* 110(3) 481–507.

Brynjolfsson, Erik, Yu J. Hu, and Mohammad S. Rahman. 2009. Battle of the Retail Channels: How Product Selection and Geography Drive Cross-Channel Competition. *Management Science* 55(11) 1755-1765.

Brynjolfsson, Erik, Yu J. Hu, and Michael 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, Erik, Yu J. Hu, and Michael D. Smith. 2006. From Niches to Riches: The Anatomy of the Long Tail. *Sloan Management Review*. Summer 2006, 47(4). 67-71.

Brynjolfsson, Erik, Yu J. Hu and Michael D. Smith. 2010. Long tails vs. Superstars: The Effects of Information Technology on Product Variety and Sales Concentration Patterns, *Information Systems Research* 21(4) 736-747.

Brynjolfsson, Erik, and Michael D. Smith. 2000. Frictionless Commerce? A Comparison of Internet and Conventional Retailers. *Management Science* 46(4) 563-585.

Cachon, Gerard, Christian Terwiesch, and Yi Xu. 2008. On the Effects of Consumer Search and Firm Entry on Multiproduct Competition. *Marketing Science* 27(3) 461-473.

Chellappa, Ramnath, Benn Konsynski, Vallabhajosyula Sambamurthy, and Shivendu Shivendu. 2007. An Empirical Study of the Myths and Facts of Digitization in the Music Industry. Workshop on Information Systems and Economics, Montreal, Canada.

Chevalier, Judith, and Austan Goolsbee. 2003. Measuring Prices and Price Competition Online: Amazon and Barnes and Noble. *Quantitative Marketing and Economics*. 1(2) 203-222.

Clay, Karen, Ramayya Krishnan, Eric Wolff, and Danny Fernnades. 2003. Retail Strategies on the Web: Price and Non-Price Competition in the Online Book Industry. *Journal of Industrial Economics* 49(4) 521-540.

De, Prabuddha, Yu J. Hu, and Mohammad S. Rahman. 2010. Technology Usage and Online Sales: An Empirical Study. *Management Science* 56(11) 1930-1945.

Diamond, Peter. 1971. A Model of Price Adjustment. *Journal of Economic Theory* 3: 156-168.

DVDStation. 2005. Studio Impact: Catalog and Long Tail Product Distribution at Retail. Available at http://www.dvdeverywhere.com/blog/download/DVDStation-CatalogImpact.pdf.

Elberse, Anita and Felix Oberholzer-Gee. 2007. Superstars and Underdogs: An Examination of The Long Tail Phenomenon in Video Sales. Working Paper 07-015, Harvard Business School.

Fleder, Daniel, and Kartik Hosanagar. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* 55(5) 697-712.

Engel, J., R. Blackwell, P. Winiard. 1996. *Consumer Behavior*. Dryden Press, Hinsdale, IL.

Glaeser, E.L., J. Kolko, A. Saiz. 2001. Consumer City. *Journal of Economic Geography* 1(1) 27-50.

Goolsbee, A. 2000. In a World without Borders: The Impact of Taxes on Internet Commerce. *Quarterly Journal of Economics* 115(2) 561-576.

Gini C. 1912. Variabilità e mutabilità. Reprinted in *Memorie di Metodologica Statistica* Editors: Pizetti, E. and T. Salvemini. Libreria Eredi Virgilio Veschi, Rome, Italy, 1955.

Greco, Albert N. 1997. *The Book Publishing Industry*. Allyn and Bacon, Boston.

Hann, Il-horn, Eric Clemons, and Lorin Hitt. 2003. Price Dispersion and Differentiation in Online Travel: An Empirical Investigation. *Management Science* 48(4) 534-549.

Kotler, P. 2002. *Marketing Management*. Prentice-Hall, Upper Saddle River, NJ.

Mobasher, B., H. Dai, T. Luo, M. Nakagawa. 2001. Effective Personalization Based on Association Rule Discovery from Web Usage Data. *Proceedings of the 3rd International Workshop on Web Information and Data Management,* Atlanta, Georgia.

Lorenz, M.O. 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association* 9(70): 209-219.

Moe, W.W. 2003. Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-store Navigational Clickstream. *Journal of Consumer Psychology* 13(1-2) 29-39.

Morton, Fiona S., Florian Zettelmeyer, and Jorge Silva-Risso. 2001. Internet Car Retailing. *Journal of Industrial Economics* 49(4) 501-520.

Nelson, P. 1974. Advertising and Information. *Journal of Political Economy* 82(4) 729-754.

Oestreicher-Singer, Gal, Arun Sundararajan. 2006. Network Structure and the Long Tail of Electronic Commerce. Working Paper, New York University, New York, NY.

Pareto, Vilfredo. 1896. Cours d'economie Politique. Bousquet, G.H., G. Busino, eds. *Oevres Completes de Vilfredo Pareto*, 1. Librairie Droz, Geneva, 1964. Originally published 1896.

Professional Publishing Report. 1999. University Presses Credit Internet for Increased Sales. Volume 3, Number 2, January 29.

Rassler, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications and Alter-native Bayesian Approaches*. Springer, New York, NY.

Rosenbaum, P., D. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70(1) 41-55.

Rothschild, M. 1974. Searching for the Lowest Price when the Distribution of Price is Unknown. *Journal of Political Economy* 82(4) 689-711.

Russo, J.E., E. Johnson. 1980. What Do Consumers Know about Familiar Products? *Advances in Consumer Research* 7(1) 417-423.

Stigler, G.J. 1961. The Economics of Information. *Journal of Political Economy* 69(3) 213-225.

Wolinsky, Asher. 1986. True Monopolistic Competition as A Result of Imperfect Information. *Quarterly Journal of Economics* 101(3) 493-512.

Zipf, George. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.