Evaluating Pricing Strategy Using eCommerce Data: Evidence and Estimation Challenges*

Anindya Ghose† Arun Sundararajan‡
Leonard Stern School of Business, New York University

October 15, 2005

Abstract

As Internet-based commerce becomes increasingly widespread, large data sets about the demand for and pricing of a wide variety of products become available. These present exciting new opportunities for empirical economic and business research, but also raise new statistical issues and challenges. In this article, we summarize a program of research that aims to assess the optimality of price discrimination in the software industry using a large ecommerce panel data set gathered from Amazon.com. We describe the the key parameters relating to demand and cost that must be reliably estimated in order to successfully accomplish this research, and outline our approach to estimating these parameters. This includes a method for "reverse engineering" actual demand levels from the sales ranks reported by Amazon, and approaches to estimating demand elasticity and variable costs directly from publicly available ecommerce data. Our analysis raises many new challenges to the reliable statistical analysis of ecommerce data, and we conclude with a brief summary of some salient ones.

1. Introduction. The adoption of Internet-based commerce has provided academic researchers with a wealth of new data on demand and pricing across a number of industries, although the availability of these data and their growing use in empirical studies of electronic commerce raises a number of new statistical and econometric issues. In this article, we

*We thank seminar participants at New York University, participants at *First Annual Symposium on Statistical Challenges in E-Commerce* for their comments, Wolfgang Jank and Galit Shmueli for their detailed feedback on an earlier draft of this paper, and Rong Zheng for outstanding research assistance in data collection. This research was partially supported by a grant from the NET Institute (www.NETinst.org).

†Tel: (212) 998-0807, E-mail: aghose@stern.nyu.edu

‡Tel: (212) 998-0833, E-mail: asundara@stern.nyu.edu

summarize a program of research, which consists of multiple related studies, each of which aims to empirically analyze and evaluate pricing strategy in the consumer software industry using a large-scale ecommerce data set from Amazon.com. We describe some of the statistical and econometric methods we have applied to our analysis of these data, how we have adapted them to address issues unique to ecommerce data, and summarize open challenges whose resolution will help facilitate more robust empirical research in electronic commerce.

Pricing strategy in the consumer software industry (and in many other industries) often involves the use of *price discrimination*, which, broadly, aims to identify (directly or otherwise) customers who are willing to pay more for a product and to charge them a higher price. Beyond the idealized notion of "first-degree" price discrimination (charging different consumers different prices for an identical good), there are a variety of ways by which firms price discriminate. For example, a seller may price differently depending on whether a consumer has purchased from the firm before (these are typically called *introductory offers*). A seller may vary the price of a product depending on how many units of the product are purchased by an individual consumer; this is commonly referred to as *nonlinear pricing* (Sundararajan, 2004). A seller may base the price of a product on whether other related products are also purchased from the same firm: this is called *bundling*, and a seller may choose to implement either *pure* bundling, under which a set of products are sold only as a bundle, or *mixed* bundling, under which both the bundle and individual products are sold. As an example of the latter, Microsoft sells its Office suite of software as a bundle of Word, Excel and PowerPoint, while selling each of these products individually as well. A seller may create different but related versions of a product (typically one of higher quality, or with

more features), and price them differently. This is referred to as *versioning*, and aims to price discriminate by exploiting differences in how much different customer value product quality. There are multiple versions of a large number of popular desktop software titles that differ only in their quality or number of features (rather than in their development or release date), and which are sold at different prices. Current examples include Adobe Acrobat, TurboTax, Microsoft Money and Norton AntiVirus. These are examples of software titles for which a firm has developed a flagship version, disabled a subset of the features or modules of this version, and released both the higher quality version and one or more lower quality versions simultaneously. A related form of price discrimination is based on releasing *successive generations* of the same product in multiple periods, with period of time where the old and new generation overlap; since each new generation represents an improvement in the overall performance of the product, the simultaneous presence of two or more successive generations is analogous to the presence of two or more related products of varying quality.

The objective of a software company that price discriminates is to maximize the profits it generates from the sale of its products. However, price discrimination can often have countervailing effects on a firm's profits. For instance, two consequences of introducing a lower quality version of an existing product in order to price discriminate are the loss of profits from customers who switch from purchasing the higher quality version to purchasing the lower quality version (commonly termed "cannibalization") and a gain in profits from new customers, for the lower quality version, who either did not purchase the product earlier, or who purchased a competing product. The interplay between these consequences eventually determines the optimality of versioning. A similar pair of consequences, with opposing

effects, characterizes the eventual profitability of bundling. Similarly, nonlinear pricing that discounts high usage levels too extensively can reduce a seller's profits.

Thus, to profit from price discrimination, a software company must make an appropriate choice of the form of price discrimination; it must choose its prices optimally, and sometimes, it must determine optimal quality levels for an inferior (related) set of products, or the size of a bundle. There is no published research with evidence that software companies in fact make these price discrimination choices optimally; however, the availability of detailed price and demand data from ecommerce sites like Amazon now makes it feasible to empirically assess the optimality of these. A first goal of our research program is therefore to use this data to empirically evaluate the optimality of such price discrimination strategies in the software industry. This is a problem of significant economic importance.

In order to do so, we require a method for converting "sales ranks" (to be described in more detail in what follows) reported by Amazon.com into actual demand levels. We then need to estimate the demand system associated with our products (that is, associate variation in prices with variation in demand). Amazon.com does not provide any data about the variable cost of the products it sells, and we therefore also need to infer these costs from our data (since the profit to a seller is determined not just by price charged and quantity sold, but also by its cost per unit). We describe our approach towards accomplishing each of these in some detail. We briefly summarize other estimates that contribute to our research program, and conclude with some of the key statistical challenges that emerge from our analysis.

2. Summary of Data. Our data are compiled from publicly available information on

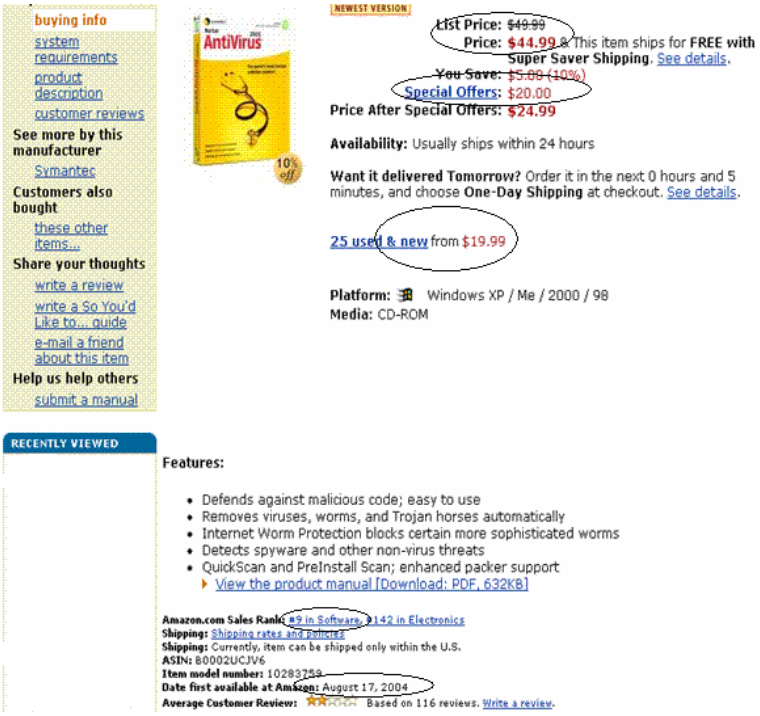


Figure 1: Illustrates how the data we gather from Amazon.com is displayed on its web site.

new software prices and sales rankings at Amazon.com, the largest online retailer of consumer software. Our data are gathered using automated Java programs to access/parse HTML and XML pages downloaded from its web site, three times each day, at equally spaced intervals. Our data includes 330 individual software titles. We collect all relevant data on list prices (the manufacturer's suggested price), new prices (the price charged by Amazon.com), sales ranks (to be discussed soon), product release date, average customer review and number of reviewers. To facilitate an understanding of how each of these is reported to a consumer on Amazon.com's web site, we illustrate a screenshot of an Amazon page in Figure 1.

We also collect data on secondary market activity, including used prices (prices charged by sellers who have posted second-hand copies of the product for sale) and new prices from

non-Amazon sellers (these are sellers who are not affiliated with Amazon but are allowed to sell goods on Amazon in exchange for a 15% commission on their transaction price). We provide the details of the data in a summary table.

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
Sales Rank	1649.61	1971.26	1	11622
List Price	69.16	226.17	19.95	1799.99
Amazon Price	65.53	208.57	14.95	1699.99
New Non-Amazon Price	17.74	23.08	10.01	209.99
Customer Rating	3.14	0.99	1	5
Number of Reviewers	25.72	66.3	1	606
DaysRelease	717.7	1336.22	0	1750

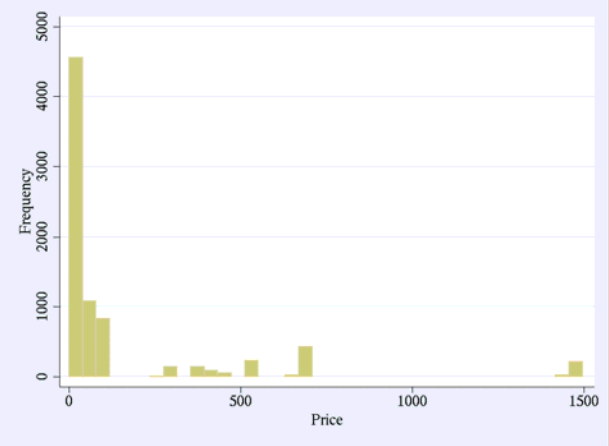
Table 1: Summary Statistics

We have categorized our software titles in three ways: (i) based on those titles that have just two versions, and those which have more than two versions, (ii) based on whether the title is sold as part of a bundle or as a stand alone product, and (iii) based on whether the title is from the most recent generation or from a previous generation. This categorization is summarized in Table 2.

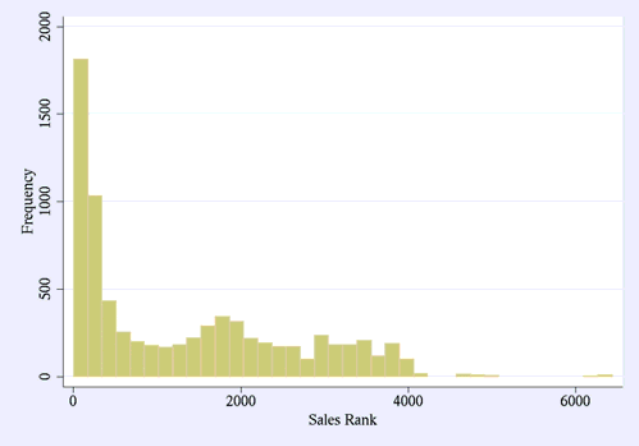
Product Category	*Number of unique titles*	*Total number of products*
Bundles	68	136
Versions (2)	32	64
Versions (>2)	19	57
Successive Generation	56	112

Table 2: Various Product Categories in Sample

Our data was collected between January 2005 and September 2005. During the entire duration of our study there were very few days over which the JAVA program was unable to collect data all three times during the day. In most cases this happened if the Amazon server



(a) Distribution of prices across our data



(b) Distribution of salesranks across our data

Figure 2: Summarizes the distribution of prices and salesranks across observations in our data set.

was not functioning properly during the time the data was being gathered. The distributions of price and sales rank across our products are summarized in Figure 2.

3 Demand Estimation. Like many other ecommerce firms, Amazon.com does not report its periodic demand levels. Instead, it reports a sales rank for each product sold on its site, and these sales ranks are updated every hour. The sales rank of a product indicates how sales of the product are relative to other products in its category. Thus, the lower the sales rank, the higher the sales for that particular item. The calculation is based on Amazon.com sales and is updated each hour to reflect recent and historical sales of every item in that category sold on Amazon.com. Prior research (for instance, Chevalier and Goolsbee, 2003 Brynjolfsson, Hu and Smith, 2003) has associated these sales ranks with demand levels. To do so, they needed to know the probability distribution of book sales. A common distributional assumption for this type of rank data is a Pareto distribution (i.e.,

a power law). Hence, they convert sales ranks into quantities by conjecturing the Pareto relationship $\log[Q] = \beta_1 + \beta_2 \log[Rank]$, where Q is the (unobserved) demand for a product, $Rank$ is the (observed) sales rank of the product, and β_1, β_2 are industry or category specific parameters[1].

A number of recent studies (including Ghose et al. 2004) pertaining to the book industry have used estimates of β_1 and β_2 from this prior literature. However, these are industry specific parameters, and to our knowledge, there are no corresponding estimates available for software. Furthermore, in summer 2004, Amazon altered its sales rank system in the following way: they eliminated their three-tier system, updating ranks each hour for most products (rather than merely for the top products), and they moved to a system that uses exponential decays to give more weight in the sales rank to newer purchases. Towards a more current and accurate reverse engineering of the ranking system to infer actual periodic demand, we have designed and conducted an independent analysis to convert measured sales ranks into demand data. Retaining the assumption of a Pareto relationship between demand and sales rank, we combine a series of purchase experiments with the analysis of a time series of sales ranks of all the 330 products in our sample to estimate both α and β.

Over a two-week period in mid-June 2005, we collected hourly sales rank data for each of the products in our panel, yielding a time series of 336 observations for each product. For products not ranked too high, a general trend in these time series is an extended downward drift in the rank value over many hours (that is, the rank becomes a progressively larger

[1] Chevalier and Goolsbee 2003 report that evidence that the Pareto distribution fits well can be found using the weekly Wall Street Journal book sales index which, unlike other bestseller lists, gives an index of the actual quantity sold. This index is constructed by surveying Amazon.com, BN.com, and several large brick and mortar book chains. For discussions on the use of power law distributions to describe rank data, see Pareto (1897) and Quandt (1964).

number), followed by intermittent spikes which result in a large upward shift in rank (that is, the rank became a smaller number suddenly). This is illustrated for two candidate products in Figure 1. We interpret these spikes as reflecting time periods in which one or more purchases have occurred.

This procedure yielded a data set of a certain number of observations, which associated a weekly demand level with each average sales rank, for two successive weeks. Weekly unit sales ranged from 0 to 16. Using the implied pairs of average weekly demand and average sales rank, we them estimated the OLS equation:

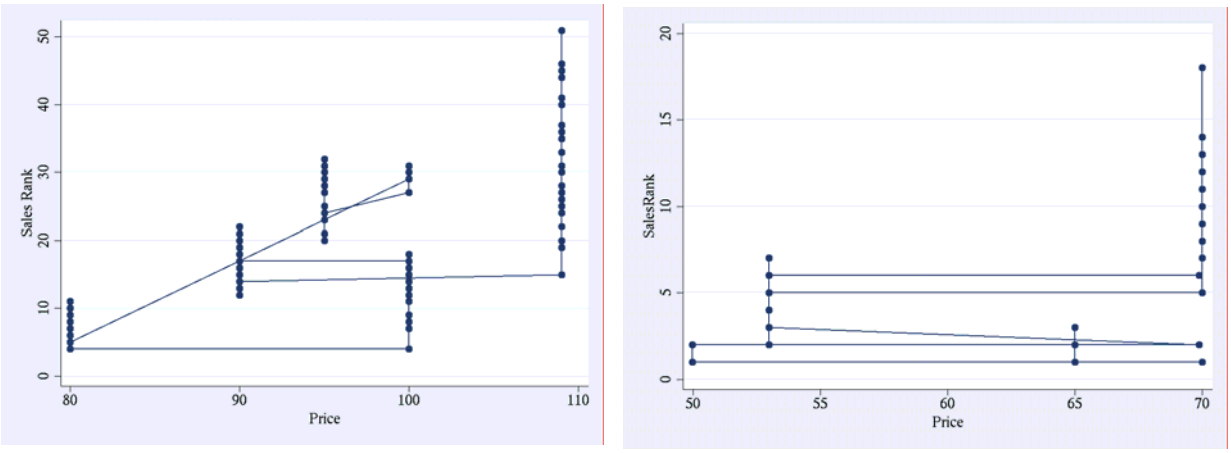
$$\log[q+1] = \log[\alpha] + \beta \log[rank], \tag{1}$$

where q is average weekly demand, and $rank$ is the corresponding average sales rank[2]. The results of these experiments yielded $\alpha = 8.352$ and $\beta = -0.828$. To provide a sense for what this estimate implies: weekly sales of two units correspond to an average sales rank of about 3100, weekly sales of 10 units correspond to an average sales rank of about 440, and weekly sales of 25 units correspond to an average sales rank of about 150.[3]

Our sample encompasses multiple products with observations collected over time, and our data set therefore has elements of both cross-sectional and time-series data. Consistent with existing published and current research, we analyze our observations as panel data (for a detailed treatment of the econometric analysis of panel data, see Wooldridge, 2002). Figure 3 below shows a scatterplot of the changes in prices and sales ranks for the two versions (of high and low quality) of a specific product for a specific period of time.

[2]Similar to Brynjolfsson, Hu and Smith (2003), we used White's heteroskedasticity-consistent estimator (see Greene 2000, page 463) to estimate both parameters.

[3]Further details of this experiment and its results are presented in Ghose and Sundararajan (2005a).



(a) Higher quality version

(b) Lower quality version

Figure 3: Illustrates the variation of salesrank with price for two versions of a specific software title in our data set. A line between two points indicates that the (price, salesrank) had changed from one of the points to the other in successive periods.

3.2 Estimates of Price Elasticity

In order to compute own-price and cross-price elasticities, we estimate OLS regressions with product and category level fixed effects. Ideally demand estimation involves regressing quantities and prices. Since we do not have direct measures of quantities in our data, we regress prices on sales ranks. Based on these estimates, we subsequently compute the own and cross-price elasticities by weighing them with the appropriate Pareto mapping parameter. This is similar to the approach used in prior literature (Chevalier and Goolsbee 2003, Ghose et al. 2004). These regressions have the following general form:

$$Log(Rank_i) = a + \phi Log(P_i) + \sum_j \gamma_j Log(P_j) + \lambda \sum_k Log(P_k) + \boldsymbol{\omega}\mathbf{x} \qquad (2)$$

where i indexes the product in question (for instance, the high quality version of a specific title), j indexes the related products whose prices P_j affect the demand for product i (for

example, the price of a lower quality version corresponding to a high-quality version, or the price of a bundle which contains product i as one component), k indexes the lowest price posted for a related non-Amazon marketplace product (the best price which is the best price across all conditions by competing sellers on Amazon's secondary market, and $\mathbf{x}$ is a vector of control variables. Our control variables include the time since the product was released ($DaysRelease$), the average customer rating ($CustomerRating$) and the number of reviewers ($NumberofReviewers$) who have reviewed the product. Note that we also use the fixed effects transformation (Wooldridge, Chapter 10.5) to control for unobserved heterogeneity across product categories.

We can use the results of this regression to calculate the relevant own- and cross-price elasticities, η_{ii} and η_{ij}, respectively. One can easily show that having estimated the parameters ϕ and each of the γ_j's, the own-price elasticity of demand for product i is

$$\eta_{ii} = \beta\phi, \tag{3}$$

where β was estimated in Section 3.1, and the cross-price elasticity of demand for product i with respect to product j is

$$\eta_{ij} = \beta\gamma_j. \tag{4}$$

These elasticity estimates describe how demand varies with price, and form the basis for analyzing the optimality of a firm's chosen price-discrimination strategy, since they enable us, for example, to assess how demand would vary if the firm altered its price discrimination by removing a version, or discontinuing a bundle. They also are inputs into the estimation process for variable costs, as described in the following section.

3.3 Cost Estimation. Contrary to what is commonly assumed in the IT economics

literature, packaged consumer software is not an "information good". Many products in IT industries have an unusual cost structure: high fixed costs of production, but near-zero or zero variable costs of production. This cost structure characterizes a class of technology products which are collectively termed information goods. Put differently, the cost of producing the first unit of an information good is very high, and yet the cost of producing each additional unit is virtually nothing. For instance, Microsoft spends hundreds of millions of dollars on developing each version of its Windows operating system. Once this first copy of the OS has been developed, however, it can be replicated costlessly. Early examples of information goods were computer-based information services and software; currently, a wide variety of diverse products – video, music, textbooks, digital art, to name a few – share this unique cost structure.

However, packaged consumer software has positive variable costs associated with its production, packaging and distribution, and these may represent a substantial fraction of the price of such software, especially since a number of titles are priced under fifty dollars. Therefore, in order to assess the optimality of a seller's choice of price discrimination, we need estimates of the variable costs of the software titles in our data set. We estimate the variable costs by inferring the Lerner index for each product version i, defined as the ratio of the markup to the price, or as $((p_i - c_i)/p_i)$, where p_i is the price and c_i is the variable cost of product i. Towards doing this reliably using ecommerce data, we have developed an extension of the model of Hausman (1994) that provides a way to estimate markups using just sales rank data and prices. We begin with the approach of Hausman (1994), who provides the following equation to estimate the markups for products sold by multi-product

oligopolists, weighted by their market share. Consider a set of related products indexed by i. His first-order conditions for oligopoly profit maximization yield the following system of equations:

$$s_j + \sum_i \left[\frac{p_i - c_i}{p_i} \, s_i \right] \eta_{ij} = 0, j = 1, 2, ..., n \tag{5}$$

Here, s_i is the demand share of product i [demand share is the ratio of revenues from product i to the total revenues from all related products], η_{ii} is product i's elasticity of demand with respect to its own price, and η_{ij} is the cross-price elasticity of demand with respect to the price of product j. He therefore has a system of linear equations

$$s + \mathbf{N}'\mathbf{m} = 0 \tag{6}$$

, where s is the vector of revenue shares, $\mathbf{N}$ is the matrix of cross price elasticities $[[\eta_{ij}]]$, and $\mathbf{m} = [m_0, m_1, ..., m_n]$,where

$$m_i = ((p_i - c_i)/p_i)s_i \tag{7}$$

is the Lerner index of product i multiplied by its product share. The marginal costs c_i of each individual product can then be estimated by inverting $\mathbf{N}$ to solve the system of equations (6).

Our extension of this approach allows the estimation of variable costs from just sales ranks, the parameter β which we estimate in section 3.1, and observed prices. Our system of equations for a set of related products $0, 1, ..., n$ with prices p_i and sales ranks R_i is:

$$s_j = \beta \sum_i \left[(p_i - c_i) \frac{s_i}{R_j} \frac{dR_j}{dp_i} \right], \tag{8}$$

where

$$\frac{1}{s_i} = 1 + \frac{p_i}{p_j} \left(\frac{R_j}{R_i} \right)^{\beta} \tag{9}$$

Our application of this model to our data set has indicated, for instance, that the inferred variable costs of security software are significantly higher than the average variable costs of other consumer software, which is consistent with the additional lifetime costs of maintenance/updates associated with this product category (Ghose and Sundararajan, 2005b).

4. Conclusion

Our objective in this paper is to outline analyzing the optimality of pricing strategy using ecommerce panel data. While we have shown how the widespread availability of ecommerce data presents a number of novel empirical research opportunities, it is important to point out that there are significant new challenges faced by researchers who aim to analyze these data in a statistically valid and economically meaningful way. Although our context is of price discrimination in software, the methods we use apply equally well to ecommerce data about any consumer product category.

A key statistical challenge in demand estimation of this kind is that the time structure of ecommerce data is not well understood. Granted, one can control for systematic seasonal effects (such as time of day that the data was collected, or month of year), for major event effects (such as the release of a new version of Windows), and check one's data for autocorrelation. However, ecommerce is still at a relatively early stage of its evolution, and the fraction of retail demand fulfilled by ecommerce sites continues to grow over time. This is driven by both an increase in the number of consumers who shop online, and an increase in the fraction of their purchases made online. Each of these factors may affect the relationship of observed ecommerce demand and price, which in turn suggests that ecommerce data may have a complex underlying time structure.

Further, new theory that models the time structure of such ecommerce data in a more precise way, and techniques that identify and account for time variation may enable future research to assess whether the demand process generating such observations is stationary, and whether the ecommerce market in question is in fact in equilibrium. This is a challenge not just for retail panel data, but for other forms of data generated by consumers interacting with ecommerce sites, such as bidding/reputation data from online auctions. Current research that studies the time structure of bid paths on eBay (Jank and Shmueli, 2005) may be a first step towards understanding similar data generation processes.

A different challenge relates to the extent one can conclude that inferences from data sets such as ours are representative of the characteristics of an industry (in our case, consumer software) in general. Clearly, this is likely to be less of an issue as a larger fraction of commerce is conducted electronically. We have benchmarked our price and demand distributions with a comparable dataset from Buy.com, another large software retailer. However the frequency with which the latter site they updates their sales ranks is different from that of Amazon, and statistical techniques that enable one to assess how representative our intra-day data is based on benchmark data with a different granularity would be helpful.

In addition to the demand and cost estimates we have described in this paper, our research program also involves developing econometric estimates of how consumers perceive the relative quality levels of related products. We have established that these econometric estimates differ significantly from comparable estimates based on self-reported quality assessments from Amazon.com, and subjective assessments by CNET editors, and our results suggest that self-reported ecommerce ratings may provide a good ordinal basis for ranking

products, but their numerical magnitudes may not be appropriate cardinal measures of quality. Since aggregate customer feedback measures from eBay, Amazon.com, and various other review sites are frequently used in ecommerce research as measures of some form of quality, statistical techniques that facilitate assigning appropriate cardinal values to ecommerce ratings data generated by consumers and editors would contribute to the foundations of this line of research. The details of this study are available in Ghose and Sundararajan (2005a).

To summarize, we have described a related set of studies that use ecommerce panel data to evaluate the optimality of different forms of price discrimination in the software industry. By describing our data, our approach to estimating some important parameters, and summarizing some of the issues that researchers face when conducting such statistical analysis on ecommerce data, we have aimed to stimulate thought about statistical challenges that arise when conducting research based on these increasingly widely used data sets. We hope that this summary will encourage future work that identifies and addresses these challenges, thereby strengthening the statistical foundations of this exciting and rapidly evolving new research area.

References

1. Bapna, R., W. Jank and G. Shmueli (2004). Price Formation and its Dynamics in Online Auctions. http://www.smith.umd.edu/faculty/wjank/auctionDynamics.pdf

2. Brynjolfsson, E., Y. Hu, and M. Smith (2003). Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety. *Management Science* 49(11) 1580-1596.

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

4. Ghose, A. and A. Sundararajan (2005a). Software Versioning and Quality Degradation? An Exploratory Study of the Evidence. Working Paper # CeDER-05-20, Stern School of Business, New York University.

5. Ghose, A. and A. Sundararajan (2005b). Pricing and Innovation for Security Software: Theory and Evidence. Working Paper # CeDER-05-24, Stern School of Business, New York University.

6. Ghose, A. K. Huang and A. Sundararajan (2005c). Versions and Successive Generations. An Analysis of Pricing and Product Line Strategies in Software Markets. Mimeo, NYU.

7. Ghose, A., R. Telang and R. Krishnan (2005). Impact of Electronic Secondary Markets on Supply Chain. *Journal of Management Information Systems*, 22(2) 91–120.

8. Ghose, A. M. Smith and R. Telang (2004). Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Social Welfare. Working Paper, Stern School of Business, New York University

9. Hausman, J. (1994). Valuation of New Goods Under Perfect and Imperfect Competition. NBER Working Paper 4970.

10. Jank, W. and G. Shmueli (2005). Functional Data Analysis in Electronic Commerce Research. *Statistical Science*.

11. Pareto, V. (1897). Cours d'Economie Politique, F. Rouge, Lausanne.

12. Quandt, R. E. (1964). Statistical discrimination among alternative hypotheses and some economic regularities. *Journal of Regional Science* 5, 1-23.

13. Sundararajan, A. (2004). "Nonlinear Pricing of Information Goods. *Management Science* 50, 1660-1673.

14. Wooldridge, J. (2002). *Econometric Analysis of Cross-Sectional and Panel Data.* MIT Press.