NET Institute*

www.NETinst.org

Working Paper #08-15

September 2008

Consumer Search on the Internet

Babur I. De los Santos
Indiana University

* The Networks, Electronic Commerce, and Telecommunications ("NET") Institute, http://www.NETinst.org, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, "virtual networks" comprised of computers that share the same technical standard or operating system, and on network issues in general.

Consumer Search on the Internet

Babur I. De los Santos*
Kelley School of Business
Indiana University

September 15, 2008

Abstract

This paper uses consumer search data to explain search frictions in online markets, within the context of an equilibrium search model. I use a novel dataset of consumer online browsing and purchasing behavior, which tracks all consumer search prior to each transaction. Using observed search intensities from the online book industry, I estimate search cost distributions that allow for asymmetric consumer sampling. Research on consumer search often assumes a symmetric sampling rule for analytical convenience despite its lack of realism. Search behavior in the online book industry is quite limited: in only 25 percent of the transactions did consumers visit more than one bookstore's website. The industry is characterized by a strong consumer preference for certain retailers. Accounting for unequal consumer sampling halves the search cost estimates from $1.8 to $0.9 per search in the online book industry. Analysis of time spent online suggests substitution between the time consumers spend searching and the relative opportunity cost of their time. Retired people, those with lower education levels, and minorities (with the exception of Hispanics) spent significantly more time searching for a book online. There is a negative relationship between income levels and time spent searching.

*I benefited from comments from Jean-Pierre Dubé, Jeremy Fox, Matt Gentzkow, Austan Goolsbee, Günter Hitsch, Steven Levitt, Jesse Shapiro, Chad Syverson, and seminar participants at the University of Chicago, Indiana University, and the Federal Reserve Board - Kansas City. I am especially grateful to Ali Hortaçsu for his support and guidence. I thank Janet McCabe at ComScore Networks for her help with the acquisition of the data. I gratefully acknowledge financial support from the NET Institute (www.netinst.org), the Kauffman Foundation, and CONACYT. E-mail: babur@uchicago.edu.

1 Introduction

This paper uses consumer search data to explain search frictions in online markets, within the context of an equilibrium search model. I use a novel dataset of consumer online browsing and purchasing behavior. This dataset is unique in that it allows tracking of all consumer search prior to each transaction. Using observed search intensities, I estimate search cost distributions that allow for asymmetric sampling by consumers. These estimates can help explain price dispersion in online markets. Search data is also useful in identifying sources of search cost heterogeneity, and the resulting substitution patterns between time spent searching and online expenditures.

The expansion of e-commerce has motivated a large body of research that analyzes search mainly through measures of price dispersion in online markets. This research relies predominantly upon prices from price comparison websites, and there is a general notion in these studies that substantial price dispersion persists in online markets.[1] However, estimates of price dispersion appear to be highly sensitive to implied market structure. In the absence of quantity data, most studies weight prices from different firms equally and assume that sales occur at each observed price. These are questionable assumptions in most online markets given the presence of large dominant firms and retailers that sell very small quantities. Using prices from the eight bookstores with the largest number of online visitors, Brynjolfsson and Smith (2000) estimate significantly lower price dispersion when controlling for a firm's market share. In addition, the high concentration of some online markets suggests that the distribution of price offerings from comparison sites is likely to differ greatly from equilibrium price distribution. This could have a significant effect on measured price dispersion and therefore on search cost estimates.

In the model, I relax the assumption that consumers randomly sample from the price distribution. Research on consumer search often assumes a symmetric sampling rule for analytical convenience. However, this is not a realistic assumption: search patterns in this dataset indicate a strong consumer preference for certain retailers. For example, in only 25 percent of transactions do consumers visit more than one bookstore. Amazon was the first bookstore visited by a consumer in 65 percent of the transactions. In about 17 percent of transactions, consumers visited another bookstore before completing their transaction at Amazon. In contrast, about 39 percent of Barnes and Noble's customers visited another bookstore, mainly Amazon. In fact

[1] See e.g. Clay et al. (2001, 2002), Baye et al. (2004), Brynjolfsson and Smith (2000), and Ellison and Ellison (2004). Baye and Morgan (2005) provide a good summary of this research.

most of the online bookstores seem to be ignored by consumers in their search. I found that only 15 online bookstores from the dataset had book sales, with Amazon and Barnes and Noble capturing 84 percent of the market. Most of the online bookstores had no visits by consumers in the dataset: of the more than 230 bookstores, 15 bookstores in the sample capture 98 percent of all consumer visits. These search patterns in the online book industry further support the assumption that consumers have prior beliefs about the market distribution of prices.

These consumer search patterns indicate asymmetry within the online book market, which must be accounted for in analyzing search frictions. Empirical price distributions that incorporate these asymmetries exhibit smaller gains from search, thus implying lower search costs. Accounting for unequal consumer sampling halves the search cost estimates from \$1.8 to \$0.9 per search in the online book industry.

Analysis of time spent online suggests substitution between the time consumers spend searching and the relative opportunity cost of their time. Retired people, those with lower education levels, and minorities (with the exception of Hispanics) spent significantly more time searching for a book online. There is a negative relationship between income levels and time spent searching. As indicated by the smaller number of bookstores they visit, individuals with income greater than \$100,000 have significantly higher search cost.

The rest of the paper proceeds as follows. Section 2 outlines the basic framework of a nonsequential search model. Section 3 discusses the relevant literature on the online book industry. Section 4 describes the data and discusses consumer search patterns. Section 5 compares estimates of search cost models under symmetric and asymmetric search, and analyzes the sources of search cost heterogeneity. Section 6 presents concluding remarks.

2 Model

In this section, I present a search model based on Burdett and Judd's (1983) framework, generalized by Hong and Shum (2006), but with two important deviations.[2] First, I assume that consumers are knowledgeable about the market's equilibrium price distribution, but do not know which firm charges each price. Stahl (1996) exemplifies the main difference between this approach and the Nash equilibrium approach. According to Stahl, in the case of N firms, whose symmetric mixed strategy

[2]See Moraga-Gonzalez and Wildenbeest (2007) for an application of Hong and Shum's (2006) model in an oligopolistic setting.

is to draw a price from a equilibrium price distribution, $F(p)$, these N draws generate a discrete distribution of actual prices, $M(p)$, or market distribution. The main distinction between these two approaches is the information available to consumers. Under the Nash paradigm, consumers have no information regarding actual prices and their search process is optimal according to firms' mixed strategies, thus consumers randomly sample prices from $F(p)$. In contrast with this approach, I assume that consumers have some information about the market distribution $M(p)$.[3]

This assumption more accurately reflects markets where consumers have a great deal of information. For example, in the case of a finite number of multiproduct firms, consumers learn about the relative price distribution through repeated transactions with the firms. This is particularly important in cases where a firm's relative prices for a range of products are stable over time. These are features of some online markets, in particular the book industry analyzed here.

Second, I relax the assumption that consumers randomly sample from the distribution of prices. Observed search patterns in the data indicate a strong consumer preference for certain retailers, derived from brand, trust, or overall consumer awareness. This pattern further supports the assumption that consumers have prior beliefs about the market distribution of prices. Asymmetric search determines market shares. One advantage of search data and transaction prices is that we can approximate the equilibrium price distribution in the presence of firm heterogeneity.

I derive a nonsequential search model for two main reasons. First, as shown by Morgan and Manning (1985) there is not a clear advantage to sequential search over nonsequential search. In fact, their analysis shows that an optimal search rule combines the elements of nonsequential search with the flexibility of sequential search. In general, nonsequential search is preferred when there are fixed costs for search. This might be the case when online consumers budget time for their Internet shopping and have to stop when time runs out (e.g. consumer visits to online bookstores last 11 minutes on average). Second, a nonsequential model makes the best use of the available data. While I observe consumer search behavior prior to a transaction, prices are observed only when consumers complete a transaction. In addition, a nonsequential model allows the use of consumer search data to explore the sources of search cost heterogeneity using an ordered response model. A nonsequential search rule is reasonable where there are consumers who are informed about the past pricing

[3] A similar assumption can be found at Salop and Stiglitz (1977) and Rob (1985). The drawback of this approach is that it gives consumers a discrete distribution of actual prices, limiting the use of firms' mixed strategies.

strategies of a small number of firms.

2.1 Nonsequential Search

Consumers inelastically demand one unit of a homogenous good. Under a nonsequential search rule, consumers decide the number of price quotations, n, to sample prior to observing prices. The first price quote is obtained for free and consumers incur a cost c for each price quotation thereafter.[4] Consumers optimally decide n, which minimizes the total expected cost of search

$$n^* = \arg\min_{n>1} c\,(n-1) + Ep^n_{(1)} \tag{1}$$

where $Ep^n_{(1)}$ is the expected minimum price for a sample of size n. Let the equilibrium price distribution of the market, described by a probability mass function, be given by

$$f_p\,(p) = \pi_j \quad \text{for } p = p_j,\ j = 1, ..., N$$

where $\pi_j > 0$ for $j = 1, .., N$ and $\sum_{j=1}^{N} \pi_j = 1$. Let prices$\{p_i\}_{i=1}^n$ be an i.i.d. random sample rearranged in ascending order of magnitude, $p_1 \leq p_2 \leq ... \leq p_n$. The expected minimum price from a sample of size n is given by

$$Ep^n_{(1)} = E\left[\min\left\{p_1, ..., p_n\right\}; n\right] = \sum_{j=1}^{n} p_j f^n_{p_{(1)}}\left(p_j\right) \tag{2}$$

where $f^n_{p_{(1)}}\left(p\right)$ denotes the p.m.f. of the minimum order statistic when consumers sample n prices without replacement. In Appendix A, I describe in detail the methodology to compute $f^n_{p_{(1)}}\left(p_j\right)$ from consumer search data.

The optimal sample size, n^*, is a decreasing function of c and has a unique solution for a positive integer value of n. Denote the expected savings from increasing the sample size by one as

$$\Delta_n = Ep^n_{(1)} - Ep^{n+1}_{(1)}. \tag{3}$$

Given that $\Delta_i \geq 0$ for $i = 1, .., N$, and the sequence of expected savings $\{\Delta_i\}_{i=1}^N$ is nonincreasing, the optimal sample size, n^*, satisfies

$$\Delta_{n^*} < c_i \leq \Delta_{n^*-1}. \tag{4}$$

[4]This is a common assumption in the literature (e.g. Stahl, 1989). See Janseen and Moraga-Gonzalez (2004) for a oligopolistic model that analyzes the implications of deviations from this assumption.

Notice we can use reinterpret Δ_n as the largest search cost of a consumer who is indifferent between searching n^* and n^*-1 firms.[5] Hence, Δ_{n^*} can be used as cutoff values that generate partitions of search cost distribution $G(c)$. The proportion of consumers who sample $n=1,...,N$ prices is given by

$$\begin{aligned} q_1 &\equiv 1-G(\Delta_1) \\ q_n &\equiv G(\Delta_{n-1})-G(\Delta_n) \quad n=2,...,N-1 \\ q_N &\equiv G(\Delta_{N-1}). \end{aligned} \tag{5}$$

In order to recover the parameters $q_1,..,q_N$ using solely price data, Hong and Shum (2006) impose firms' pricing equilibrium conditions, and estimate the model with maximum empirical likelihood. This approach imposes conditions on the empirical price distribution that do not necessarily provide a minimum variance estimator (see Moraga-Gonzalez and Wildenbeest, 2007).

Using data on consumer search patterns and transaction prices greatly simplifies the estimation of search cost distribution. From consumer search, I calculate $q_1,..,q_N$ directly as the proportion of consumers that search $n=1,...,N$ without imposing firms' equilibrium conditions. From these values, using equation (5) I recover the search cost distribution $G(c)$ evaluated at cutoff points Δ_i, for $i=1,..,N-1$. I estimate Δ_i from the empirical distribution of transaction prices $f_p(p)=\pi_j$ for $p=p_j,\ j=1,...,N$ using data on consumer search to estimate the weights π_j. I compare the resulting search cost distributions to those that result from a random sampling rule, $f_p(p)=\pi=1/N$ for every p.

2.2 Consumer Search Cost Heterogeneity

In this section, I use consumer search data to explore the sources of search cost heterogeneity. The nonsequential search model is suitable to fit an ordered response model, given that I observe the number of firms a consumer samples before a purchase,

[5] In the case of a continuous equilibrium price distribution $F(p)$ with support $\left[\underline{p},\overline{p}\right]$, the minimum price is $m_n=\int_{\underline{p}}^{\overline{p}} pn\left[1-F(p)\right]^{n-1}dF(p)$. It is straightforward to show that it can be rewritten as $m_n=\underline{p}+\int_{\underline{p}}^{\overline{p}}\left[1-F(p)\right]^n dp$ which is a monotone decreasing sequence of n, bounded below by $\underline{p}$. The expected gain for searching one more firm is

$$\Delta_n=\int_{\underline{p}}^{\overline{p}}\left[1-F(p)\right]^{n-1}F(p)\,dp$$

which is in turn a nonincreasing and convex function of $n=1,...,N$. See the work of Burdett and Judd (1983), Hong and Shum (2006) and MacMinn (1980) for a derivation of these models.

but do not directly observe search cost for each consumer. Define Y_i as the number of firms that consumer i samples, which takes values $n = 1, ..., N$. Consumer search costs are

$$c_i = x_i\beta + \varepsilon_i. \tag{6}$$

where x_i is a vector of explanatory variables, β is a vector of parameters, and ε is an i.i.d. error with distribution H. Search costs are not directly observed in the data, but I observe the number of firms consumers sampled according to $\Delta_{n+1} < c_i \leq \Delta_n$. For $N = 4$ we have

$$Y_i = \begin{cases} 1 & \text{if } c_i > \Delta_1 \\ 2 & \text{if } \Delta_2 < c_i \leq \Delta_1 \\ 3 & \text{if } \Delta_3 < c_i \leq \Delta_2 \\ 4 & \text{if } c_i \leq \Delta_3 \end{cases} \tag{7}$$

The probabilities of a these outcomes are given by

$$\begin{aligned} \Pr(Y_i = 1) &= \Pr(c_i > \Delta_1) = \Pr(x_i\beta + \varepsilon_i > \Delta_1) \\ &= 1 - H(\Delta_1 - x_i\beta) \\ \Pr(Y_i = 2) &= \Pr(\Delta_2 < c_i \leq \Delta_1) = \Pr(\Delta_2 < x_i\beta + \varepsilon_i < \Delta_1) \\ &= H(\Delta_1 - x_i\beta) - H(\Delta_2 - x_i\beta) \\ \Pr(Y_i = 3) &= H(\Delta_2 - x_i\beta) - H(\Delta_3 - x_i\beta) \\ \Pr(Y_i = 4) &= H(\Delta_3 - x_i\beta) \end{aligned} \tag{8}$$

The likelihood function is given by

$$\ln \mathcal{L} = \sum_{i=1}^{M} \sum_{n=1}^{N} \mathbf{1}\{Y_i = n\} \ln [H(\Delta_{n-1} - x_i\beta) - H(\Delta_n - x_i\beta)] \tag{9}$$

where $\mathbf{1}\{Y_i = n\}$ is an indicator function with $\mathbf{1}\{Y_i = n\} = 1$ if $Y_i = n$ and $\mathbf{1}\{Y_i = n\} = 0$ otherwise.

Ordered response models require the distribution H to be fully specified. In the case $\varepsilon_i \sim N(0, 1)$ this is the standard ordered probit setup. However, if ε_i is not normally distributed, maximum likelihood estimates are inconsistent. Klein and Spady (1993) provide a semiparametric methodology to approximate the distribution for binary response models. Klein and Sherman (2002) extend the methodology to ordered response models. Gallant and Nychka (1987) provide a semi-nonparametric approximation of the distribution using an Hermite form, which is the product of a squared polynomial and a normal density, but could be used with any distribution

with a moment generating function (see Stewart, 2005 for an application). Both of these approaches approximate the distribution up to a location and a scale.

3 Background and Literature Review

The book industry has been the focus of studies of online markets given the maturity and predominance of the industry.[6] Since Amazon's launch in 1995, the online industry has grown to represent 17 percent of the total sales of the $24.2 billion book industry.[7] With the exception of travel services, the book industry has the highest penetration among Internet users. More than 30 percent of Internet users that responded the Forrester Technographics Survey of 2003 declared to have bought a book online. This is a highly concentrated industry, with the two dominant firms capturing 83 percent of the market: Amazon (66 percent of book sales) and Barnes and Noble (17 percent).[8]

The expansion of e-commerce has motivated a large body of research that analyses search frictions mainly through measures of price dispersion in online markets. Using predominantly prices from price comparison websites, there is a general notion in these studies that substantial price dispersion persists in a large number of online markets.[9] Brynjolfsson and Smith (2000) report price dispersion of 33 percent for 20 books sold at the eight online bookstores with the largest number of visitors. Clay et al. (2001), using prices for 32 online bookstores, estimate that the price dispersion for 399 books is between 32 and 65 percent. These studies show that online price dispersion is higher than dispersion among traditional brick and mortar retailers (e.g. Clay et al. 2001, 2002; Scholten and Smith, 2002; Pan et al. 2003).

These estimates of price dispersion appear to be highly sensitive to the implied market structure. The evidence suggests that price dispersion found in the online book industry is between large branded retailers and unbranded retailers. Clay et al. (2001) find that Amazon, Barnes and Noble, and Borders had the lowest standard deviation of price, in contrast to a large dispersion found for fringe retailers. Brynjolfsson and Smith's (2000) estimates of price dispersion are significantly lower when controlling for a firm's market share, as measured by its website's popularity. The main cause of these results is the high concentration of the industry and the similar

[6] See e.g. Clay et al. (2001) for a review of the industry.

[7] In 2006 Amazon's sales of media (books, music, and DVDs) for North America totaled $3.6 billion, and Barnes and Noble reported sales of $433 million for its online site.

[8] Books sales in dollars for 2004 from the ComScore data sample.

[9] See Pan et al. (2004) for an excellent review of the research that studies online price dispersion.

pricing strategies of large bookstores. Brynjolfsson and Smith's results indicate that the prices for Barnes and Noble and Borders average -$0.19 and $0.09 difference, respectively, from Amazon's price. Clay et al. (2001) calculate for a sample of 399 books that 77 percent of Barnes and Noble's and 75 percent of Borders' prices are within 1 percent of difference when compared to Amazon's prices.[10] This evidence suggests that controlling for market share would lead to lower estimates of price dispersion.

Limitations to the use of price data could account for some of the unexplained puzzles in the literature.[11] For example Clay et al. (2001), using prices from online comparison websites, show that small online bookstores have varied price strategies. Most of these stores set prices slightly below Amazon's prices, about $0.10 below Amazon's price. In some cases, small bookstores set prices above Amazon's prices. Clay et al. find an Amazon price premium of between 10 and 25 percent.[12] These results have the same limitations as price dispersion estimates in the absence of quantity data, since most of the price difference is between Amazon and smaller retailers. Chevalier and Goolsbee (2003) exemplify this limitation using prices and sales rank data from Amazon and Barnes and Noble. They find that prices weighted by sales differ significantly from prices estimated with sales weighted equally. Although indicative of firm heterogeneity in terms of brand, service quality, or consumer awareness, there is no conclusive evidence that higher-quality firms command higher prices (see e.g. Baylis and Perloff, 2002; Pan et al., 2003).

Structural estimates that use only price data could lead to biased estimates of search cost. Hong and Shum (2006) show that price data can also be used to estimate search cost distributions consistent with theoretical models by using information on supply and demand equilibrium conditions. However, this assumes that consumers randomly sample prices from an infinite number of firms. This assumption increases the benefit of search and could lead to higher search cost estimates.

The data on consumer search presented in the next section help to explain some of the patterns found in online markets. In particular, search data is crucial to understanding search costs in the online book industry. Search patterns indicate that consumers visit only a small number of online bookstores. Consumers might have never observed the full set of prices posted in online comparison websites. As

[10] For the market of consumer electronics, Baye et al. (2004) report that the levels of price dispersion are sensitive to variations in the number of firms that post price quotes in price comparison sites.

[11] Additionally users of price comparison sites may not represent the typical Internet user. ComScore Media Matrix found that only 4 percent of Internet users visited these sites in 2000.

[12] Average savings from buying from the lowest price among 32 bookstores listed in two major price comparison websites instead of Amazon, measured as a percentage of Amazon's price.

a result, the distribution of price offerings is likely to differ greatly from transaction prices.

4 Data

The dataset was constructed from the ComScore Web-Behavior Panel which includes detailed online browsing and transaction data from 100,000 Internet users in 2002 and 52,028 users in 2004 chosen at random from a universe of 1.5 million global users. ComScore is a leading provider of information on consumers' online behavior and supplies Fortune 500 companies and large news organizations with market research on e-commerce sales trends, website traffic, and online advertising campaigns. Each user's online activity is channeled through ComScore proxy servers that record all Internet traffic, including information on visits to a website or domain (browsing), as well as secure online transactions. The data include date, time, and duration of visit, as well as price, quantity, and description of each product purchased during the session.

The dataset contains users' transactions for products and services from June 2002 to December 2002 and for the full year of 2004. Approximately 38 percent of the users realized a product transaction in 2002 (48 percent of users in 2004), and 7 percent of users bought at least one book online in 2002 (10 percent in 2004). The book transactions exclude observations from websites that could not be identified as online bookstores, such as unidentified domains and auction sites (Appendix B describes the sample construction in detail). This results in transactions from 15 online bookstores with 17,956 observations in 2002 and 17,631 observations in 2004. Each observation represents a single book purchased during one transaction; if multiple copies of the book are purchased, it is recorded as one observation.

The browsing activity of all users consists of 112,361 visits to the websites of online bookstores in 2002 and 214,713 visits in 2004.[13] In order to identify a user's visit to a website as search behavior related to a particular transaction, I link the browsing history up to 7 days before that transaction, which I label the cutoff period. There is no evidence to guide the definition of a search time span in relation to a transaction. One week is long enough to capture all search behavior related to a transaction; any longer intervals are likely to also capture unrelated website visits. A search history could be less than 7 days if another transaction has occurred within the cutoff period

[13] This large increase was the result of a more than twofold increase in the number of visits to Amazon, which is the largest online bookstore and had 80 percent of website visits in 2004.

(in these cases, the average time span is 2.9 days between transactions). Limiting browsing to search occurring 7 days prior to a purchase reduces the sample to 18,349 observations in 2002 and 25,513 in 2004. Although some user search may not be linked to the next transaction, but to a subsequent one, there is no clear way to link this intervening search to a later transaction. For example, if a user searches prices for book A but buys book B first, the search for book A is linked to book B. In the case where multiple books are acquired in the same transaction, browsing is linked to all books purchased.

Table 1 presents descriptive statistics for the consumer browsing and transaction data. Website visits that are not linked to any transaction are significantly shorter than visits occurring within 7 days of a transaction, even when lengthy transaction visits are not included. Although the average duration of website visits has diminished from 2002 to 2004, the total duration of search has increased in this period. The dominance of Amazon and Barnes and Noble in the market might explain the low levels of consumer search: users on average searched 1.2 bookstores in 2002 and 1.3 in 2004. The average number of books bought (2.2 to 2.4 books) and average expenditure per transaction can be explained by consumers taking advantage of some bookstores' offers for free shipping for purchases above $25.

4.1 Consumer Search Patterns

Search behavior provides insight into the nature of consumer awareness, brand recognition, and preference for some firms. Amazon and Barnes and Noble capture 83 percent of book transactions and thus it is expected that most consumer search is directed at those stores. This work uncovers two important consumer search patterns in the online book industry. First, search is limited. In only 25 percent of transactions did consumers search more than one bookstore. The fraction of consumers that price shop is small: 27 percent of consumers searched more than one firm in any of their transactions in 2002, and 33 percent of consumers in 2004. Second, consumers do not visit the majority of bookstores available, they show a strong retailer preference in their search patterns, visiting 1.29 online bookstores on average.

In order to analyze consumer search of online bookstores, I grouped small bookstores into two categories to create four firms: Amazon (63 percent of transactions), Barnes and Noble (21 percent), Book clubs (12 percent), and Other bookstores (4 percent). "Book clubs" include the following sites (.com): Christianbook, Doubledaybookclub, Eharlequin, Literaryguild, and Mysteryguild. Other bookstores include

(.com): 1bookstreet, Allbooks4less, Alldirect, Booksamillion, Ecampus, Powells, Varsitybooks, and Walmart. In order to determine whether restricting consumer search to the 4 firms adequately captures consumer behavior in this market, I estimate the amount of consumer browsing directed at all 234 online bookstores listed on the Yahoo directory. As expected, consumer browsing of the four firms captures most consumer search; only about 1.6 percent of all consumer visits were directed to excluded bookstores.

One important consideration is Amazon Marketplace, first launched in November 2000, which allows third-party sellers to offer items through Amazon's website. When available, third-party offerings appear below Amazon's price on a book's webpage. Since purchases of third-party books are processed through Amazon's payment system, these transactions are indistinguishable from Amazon's direct transactions. According to Amazon's financial reports for the third quarter of 2002, third-party seller transactions represented 23 percent of North American sales units. However, this figure includes new, used, and refurbished items in several product categories in addition to books.

Table 2 displays consumer visits to any of the four firms for each book transaction. The first part of the table shows the proportion of times a particular bookstore was visited first by a consumer within the search history of each transaction. In the first column, the proportions for all transactions correspond closely the firm's market shares: Amazon was visited first in 65 percent of the sample; Barnes and Noble, 17 percent; Book clubs, 11 percent; and Other bookstores, 7 percent. The rest of the columns are conditioned on the bookstore where the consumer purchased the book. This allows me to analyze consumer retailer preferences. For shoppers who bought a book from Amazon, 91 percent visited Amazon first, compared with 68 percent of Barnes and Noble buyers who visited Barnes and Noble first. A significant share of consumers of Barnes and Noble, Book clubs and Other bookstores visit Amazon first in their search proccess (in 19 to 29 percent of transactions of these bookstores).

The second part of Table 2 shows consumer visits to bookstores at any point in the search process. Amazon was visited in 74 percent book transactions, and in only 17 percent of transactions did Amazon buyers browse any other bookstore. In contrast, Barnes and Noble buyers searched other bookstores (mainly Amazon) in 39 percent of cases; Book club shoppers, 31 percent of cases; and Other bookstore shoppers, 46 percent of cases. The limited search process is reflected in the number of stores that consumers search for each transaction. On average Amazon buyers search 1.2 bookstores, compared to Barnes and Noble, 1.5; Book clubs, 1.4; and Other

bookstores, 1.6. These patterns show the asymmetric nature of the search process in this industry.

4.2 Patterns of the Search Stopping Decision

I use observed patterns of consumer search to shed light on differences in features of common search rules found in the literature. In particular, I examine the importance of recall, which is consumers' ability to buy an item at a previously observed price. In a sequential search model with perfect recall, a consumer must decide after observing a price to stop the search and buy at that price or continue the search. Under perfect recall it is optimal to continue searching if the lowest observed price is higher than a reservation price and stop if the lowest price is less than the reservation price. As consumers do not want to incur costly search, they stop at the first price at or below the reservation price. As a consequence, if there are an infinite number of firms, consumers will always buy from the last visited firm since it is the first price below the reservation price, and they will never recall a previously observed price (see e.g. Stahl, 1996). In the case of a finite number of stores, the only reason a consumer will recall is if they visit all stores without observing a price below the reservation price. In contrast, in a nonsequential search rule, consumers choose the minimum price after observing all the prices in their optimal sample. One important note is that in cases where consumers visit only one bookstore, we cannot distinguish between these two search rules.

Table 3 presents a more detailed picture of the search process; in particular consumers' decision to halt search either by buying from the last firm or by recalling a previously searched firm. Every consumer visits at least one firm, which is the firm where they complete their transaction. The top panel of the table shows the proportion of transaction sessions where consumers visited only one store for a variety of lengths of search period. All search behavior is linked to the next transaction, and since there is no research to identify a correct search span, I have limited the lengths to 7, 5, 3 days and 1 days prior to each book purchase, or to the same day of the transaction.[14]

The table reiterates the previous findings,that consumer search in this industry is very limited. The first column shows that in 76 percent of all transactions consumers visited one bookstore when we examine search performed 7 days prior to a transaction. The proportion of people that search one store increases as we shorten the length of

[14]Note that the table refers to transaction sessions, in which consumers can purchase more than one book. In data from 2002 and 2004, consumers bought on average 2.29 books per transaction.

the search period. When we consider search on the transaction day, in 90 percent of cases consumers visited only one store. This is an expected result. For example, consider a consumer who visits firm X one week prior to purchasing a book at firm Y. If I establish a search period of 7 days, the visit to firm X will be counted as search for that transaction, but it will not be included if the time span is 6 days or fewer. As a result, the proportion of sessions where consumers visit only one firm will be larger as we consider shorter search periods and omit visits to other firms. The breakdown of transactions by firm shows the same pattern presented in the previous section: Amazon's buyers are less likely to visit other stores.

The bottom panel of the table shows decisions to stop search behavior in cases where two or more bookstores were visited. In these cases, consumers ended their search in one of two ways, by purchasing a book from the last firm they visited or by purchasing a book from a previously visited firm. Given that the proportion of cases where consumers visit two or more bookstores declines as the search period is shortened, I show the proportions for these two cases in relative terms. For example, for a search span of 7 days, in 76 percent of cases one firm was visited. The remaining 24 percent of cases correspond to visits to two or more bookstores. In 65 percent of the latter group of cases (or 16 percent overall) consumers buy from the last firm visited, and in 35 percent of the cases (8 percent overall) consumers recall a previously searched firm.

This table shows consumers exercising their recall option in 35 to 40 percent of cases, and it also shows important differences in search patterns across firms. Amazon consumers who visit other bookstores recall Amazon's price 50 to 54 percent of the time. In contrast, Barnes and Noble's consumers recall its price after visiting other bookstores in 20 to 26 percent of cases. This pattern indicates that consumers start their searches at Amazon. Hence, the majority of consumers who buy a book from Amazon after visiting other bookstores effectively recall Amazon's price (50 to 54 percent). This contrasts with search of Amazon's competition: consumers who search more than one bookstore are likely to have visited Amazon before completing the transaction at a competing bookstore. This behavior explains lower recall proportions at those firms (for example, 20 to 26 percent of consumers return to purchase from Barnes and Noble).

This table provides evidence of an underlying search asymmetry, and further exemplifies the importance of recall in search models. In a sequential model with perfect recall and an infinite number of firms, consumers always buy from the last firm and hence never recall prices. Only in the case where there is a finite number of stores and

consumers visit all stores would consumers recall a previously observed price. There is no indication in this dataset that consumers are searching exhaustively—in fact they are rarely searching more than one firm. Thus a sequential search setting does not account for the large proportion of recall behavior among those who search more than one firm. Studies of search process in a sequential setting have found similar results. This systematic recall supports the nonsequential search process presented here. However, there are other models that might explain this, such as directed search.[15]

4.3 Demographic Characteristics

In addition to browsing and transaction information, the dataset includes a rich set of user demographic characteristics that I use to analyze the components of search costs. In this section, I describe the demographic characteristics of the sample and, using other datasets for comparison, I show that the sample is an appropriate representation of Internet users. I use the Internet and Computer User Supplement of the Current Population Survey (CPS) and the Forrester Technographics Survey (FTS). User characteristics include household income and size, age of the eldest member, education level and racial background of the head of the household, and an indicator if children are present in the household. In addition, there is an indicator for high-speed Internet connection (broadband), region of residence, and zip code information for 2004.[16] Given that the three sources of data have different definitions for some variables, I present the exact methodology in Appendix B.

Table 4 presents demographic characteristics of users from ComScore, the CPS of October 2003, and FTS 2003. I condition the three datasets to those users who made any online transaction. Household composition is similar across samples with an average of about 3 people per household, and 36 to 46 percent of households having a child present. Those who purchased at least one book online (first column) are slightly older, with greater income and more education than those who had any online transaction (second and third columns).

Compared with the CPS data, ComScore Internet users are older, with higher income, but with a lower proportion of users having college and graduate degrees. The

[15] Zwick, Rapoport, and King Chun Lo (2001) find large rates of recall, which violates optimal policy, among participants of an experiment who are able to rank prices and are presented with alternatives in a sequential order.

[16] The online activity recorded cannot be linked to a specific individual in the household. In cases where multiple computers are tracked within a household, each computer is considered a different user.

discrepancy in education level is due to the large proportion of college students (those with "Some college but no degree") in the ComScore sample. The racial composition is similar across samples—online users are predominantly white. But compared with CPS, ComScore oversamples Hispanics and Forrester oversamples whites. The geographic distribution of users is similar to CPS population estimates at the regional and state levels (see Table 9 for state comparison of the samples).

As shown in the Table 4, the demographic characteristics of the users in the sample are representative of online buyers in the United States. In fact, the most-purchased books in the sample reflect purchase patterns of the U.S. population as captured in the *New York Times Best Seller* list. In the next section, I link the demographic characteristics of users in the sample with search behavior.

5 Results

In this section, I present estimates of search cost distribution implied by the non-sequential search model outlined in section 2. The model is estimated using information on consumer search and empirical price distribution. The search cost distribution is characterized by cutoff points, Δ_n, and the quantiles of the distribution q_n for $n = 1, \ldots, N$.

To estimate the model I use search data and transaction prices for a selected number of best sellers. Using the books with the largest number of transactions in the sample has two important advantages. First, observed consumer browsing reflects price search rather than visits to confirm availability of the book, since bookstores keep inventories of best-selling books. Second, using more observations for each book reduces the time difference of transactions, which is the potential bias of using implied prices.

Table 5 displays descriptive statistics for 12 best-selling books in the sample. These books reflect consumer patterns in the United States, as indicated by the fact that all but two were number one on the *New York Times Best Seller* list. For each book, I observe prices for a maximum of 3 bookstores. The mean prices are similar across books, except for *Key of Valor*, with an overall mean price of $15. The proportion of consumers that searched $n = 1, 2$, or 3 bookstores is displayed in the last three columns of the table. The majority of consumers do not search (i.e. they only visit one store), ranging from 52 to 86 percent of consumers (72 percent overall). In about 94 percent of transactions, consumers visit one or two bookstores.

Table 6 reports estimates of the empirical search cost distribution. The cutoffs of

the distribution, Δ_n, are estimated from the empirical price distribution in two ways. First, assuming equal sampling probabilities, I calculate the expected minimum price for each sample size by randomly sampling n prices from the empirical distribution and averaging over 100,000 iterations (Appendix A provides a detailed explanation). Since I only observe search at three firms, I can only identify the cutoffs Δ_1 and Δ_2. Recall from section 2 that the we can recover the quantiles of the distribution, $G(\Delta_n) \equiv 1 - \sum_{i=1}^{N} q_n$, using consumer search data to calculate q_n. The results exhibit some variation in the estimates of the search cost. For example, 29 percent of buyers of *The Da Vinci Code* have search cost below $\Delta_1 = \$1.12$.

Second, I take into account the strong preference/awareness for some retailers displayed in consumers' search patterns. Firms' unequal probabilities are calculated assuming a sampling without replacement rule with perfect recall, using the proportion of people that visit each firm as the relevant consumer sampling rule (see Appendix A). I calculate the expected minimum price for each sample size by sampling n prices using this probability.

Columns 4 and 5 of Table 6 report search cutoffs using unequal sampling probabilities. The cutoffs of the search distribution are significantly reduced in all cases. On average, Δ_1 decreases by 45 percent, from \$2.3 to \$1.24. The proportional reduction is much higher for Δ_1. For example, under equal sampling, the 29th quantile of the distribution of *The Da Vinci Code* buyers has a search cost below \$1.12. Under unequal sampling the same quantile has search cost below \$0.55. It follows that the search cost distribution under equal sampling satisfies the criteria for first-order stochastic dominance over the distribution with unequal sampling in all cases, except for the *Lovely Bones*. Thus, expected search cost is consistently overstimated using equal sampling.

The results indicate that the benefits of search are much smaller once I control for the asymmetric nature of search. As incentives for search are small, search is small and reflected in the large proportion of people that do not search within the online book industry. Although using data on consumer search greatly simplifies recovery of the search cost distribution, it has one drawback as stated by Hong and Shum (2006). This methodology cannot identify search cost for non-searchers, that is, for those with search cost above Δ_1.

To address some of these limitations, in the next section I fit an ordered response model that exploits search data to recover search cost distributions from consumer characteristics, and I address the limitations of this methodology with respect to the identification of non-searchers.

5.1 Sources of Search Cost Heterogeneity

In this section, I explore the determinants of consumer search cost heterogeneity. One measure of search cost is the time spent searching for a particular item; however, not all consumer browsing is costly search. For example, some consumers enjoy shopping and spend time browsing the selection of books looking for new acquisitions. The measure of time spent searching comprises both types of consumer browsing. Table 7 presents regression estimates of the total time spent searching for a book based on consumer characteristics. The total duration of search is presented in equation (1), and in (2) excluding those visits where consumers complete the transaction. The same distinction is made in equations (3) and (4), but for the average time spent per book bought.

There are interesting patterns in Table 7 that indicate some consumers enjoy shopping. We would expect consumers to spend less time visiting retailers where they had made transactions in the past if consumers' objective is to minimize time spent online. However, repeated interactions with the same retailer do not decrease the duration of the visits. On the contrary, consumers spend 8 to 11 more minutes per visit to known retailers (equations 1 and 2). Consumers with a larger number of past purchases spent more time visiting bookstores, which clearly indicates that demand effects outweigh any possible learning or time-saving strategies for Internet search. Also, visit duration could derive from consumers' reactions to promotional offers. While consumers spend more time on a transaction that qualifies for free shipping (total book expenses $\geq$ \$25), they spend less time per book, even when I exclude transaction visits.

An important source of search cost heterogeneity is the relative value of the time of different individuals as implied by their socio-demographic characteristics. Aguiar and Hurst (2005) found that at the time of retirement, individuals reduce food expenditures without an equivalent reduction on quantity or quality of food consumption. The discrepancy between expenditure and consumption is explained by retirees spending more time searching for food. Table 7 shows similar evidence of relatively low opportunity cost for retired people. Those with 60 or more years of age spend 5 to 6 minutes more on search than those with 45 to 50 years of age (omitted category). Surprisingly, when lengthy transaction visits are excluded, the discrepancy is greater: 60- to 65-year-olds spend 9 to 11 minutes more on search than younger shoppers.

There is an inverse monotonic relationship between education level and search duration. Those with lower education levels spend more time searching. This could be explained by the higher relative opportunity cost of more educated people, but

also because more educated people might be more efficient in their search strategies. Minorities, with the exception of Hispanics, spent more time searching before a transaction. Income levels exhibit a relationship to search costs that is similar to education levels, with higher-income individuals devoting less time to search. In this framework I cannot distinguish if this difference results from different budget constraints or higher relative value of time.

While time spent searching is a good approximation of search cost heterogeneity as shown above, there are also other important elements that influence the duration of search. Using information on the number of bookstores searched, I can analyze search cost heterogeneity directly. Recall that consumers with higher search cost will optimally visit fewer stores. Table 8 displays ordered probit estimates of the number of bookstores visited by consumers based on household demographic characteristics and transaction variables. The dependent variable is the number of bookstores visited for each transaction ($n = 1, \ldots, 4$). Equations (1) and (2) pool transactions from the years 2002 and 2004, and include the household demographic characteristics summarized in Table 4. Equations (3) and (4) include state of residence indicator variables (this information is only available for 2004). In equation (4), only transactions where a single book title was purchased are considered.

Although broadband users do not spend more time searching, faster speeds make it less costly to visit another bookstore. Relatively lower search cost is reflected in a greater number of bookstores visited by broadband users, for those of 30 to 34 years of age across specifications. Asians and people from 55 to 64 exhibit relatively lower search costs (equations 1 and 2). The price coefficient in equation (4) indicates that higher expected savings induce a larger sampling by consumers. An interesting case is the relatively higher search cost of individuals with income greater than $100,000 per year, which is consistent across specifications.

6 Conclusions

I use novel data on consumer online browsing and purchasing behavior to structurally estimate a nonsequential search model. In contrast to models in which consumers have no information about prices, I assume that consumers are knowledgeable about the equilibrium price distributions. This assumption is supported by consumer search patterns in the online book industry.

Search patterns in the online book industry indicate limited consumer search and a strong preference for particular retailers. In only 25 percent of transactions did con-

sumers visit more than one bookstore website. Amazon's dominance in this industry is reflected in consumer search patterns. I find that symmetry assumptions as to prices in online markets can lead to biased estimates of search cost and possibly measures of price dispersion. Search model estimates that incorporate market asymmetries, as captured by consumer search behavior, imply empirical market distributions that support lower search costs in equilibrium.

In addition, I show that search data is helpful to analyze the sources of consumer search heterogeneity. Estimates of search cost on consumer characteristics show a strong substitution between time and expenditures in online markets.

A Estimation of Minimum Prices under Equal and Unequal Sampling Probabilities

This appendix shows the methodology of unequal probabilities from search data implied by a discrete empirical price distribution. Let the equilibrium price distribution of the market, denoted by probability mass function

$$f_p(p) = \pi_j \quad \text{for } p = p_j,\ j = 1, ..., N$$

where $\pi_j > 0$ for $j = 1, .., N$ and $\sum_{j=1}^{N} \pi_j = 1$. Let prices $\{p_i\}_{i=1}^{n}$ be a sequence of i.i.d. random variables rearranged in ascending order of magnitude, $p_1 \leq p_2 \leq ... \leq p_n$. The expected minimum price from a sample of size n is given by

$$Ep_{(1)}^{n} = E\left[\min\left\{p_1, ..., p_n\right\}; n\right] = \sum_{j=1}^{n} p_j f_{p_{(1)}}^{n}(p_j)$$

where $f_{p_{(1)}}^{n}(p)$ denotes the p.m.f. of the minimum order statistic when a consumer samples n prices without replacement.

In the case equal probability sampling without replacement, $f_p(p) = \pi = 1/N$ for $p = p_1, ..., p_N$. I estimate the p.m.f of the minimum order statistic by combinatorial analysis. In the case that n prices are sampled, the minimum price of the sample is given by

$$f_{p_1}^{n}(x) = \frac{\binom{N-x}{n-1}}{\binom{N}{n}} \qquad x = 1, 2, .., N + 1 - n$$

where $x \in [1, 2, ..., N + 1 - n]$ denotes the reordered support of $p \in [p_1, p_2, ..., p_{N+1-n}]$ (see Evans et al., 2006).

In the case of unequal sampling, there are three cases to consider. First, if only one price is sampled, $n = 1$, the p.m.f of the minimum reduces to

$$f_{p_1}^{1}(x) = f_p(p).$$

Second, if all the stores are sampled, $n = N$, the minimum price is observed, hence $f_{p_1}^{N}(p) = 1$. Finally, $2 < n < N - 2$ is a non-trivial case, with no closed form. It is calculated from combinatorial procedures.

For simplicity, define Ω^n as the set of containing the combination of n prices

sampled from $p_1, ..., p_n$.for the case $N = 4$ and $n = 2$ the set

$$\Omega^2 = \left[\{p_1, p_2\}, \{p_1, p_3\}, ..., \{p_3, p_4\}\right].$$

Let $\omega^n \in \Omega^n$ be a combination of prices sampled by consumers when searching k firms. In order to compute the probability of obtaining ω^n we have to calculate the probability of all the permutations. For example, the probability a combination ω^2, $\{p_1, p_2\}$ is given by the sum of the probability of all permutations $[p_1, p_2]$ and $[p_2, p_1]$

$$f_p(p_1)\frac{f_p(p_2)}{1 - f_p(p_1)} + f_p(p_2)\frac{f_p(p_1)}{1 - f_p(p_2)} = \pi_1\frac{\pi_2}{1 - \pi_1} + \pi_2\frac{\pi_1}{1 - \pi_2}.$$

Given a consumer search process Ψ^n,we can calculate the probability that p_j is observed when n prices are sampled, denoted by λ_j^n=$\Pr\left(p_j \in \omega^2 \mid \Psi^n\right)$. For example, for $n = 2$ the probability of a consumer observing p_1 is

$$\begin{aligned}\lambda_1^2 = \sum_{\substack{\omega^2 \in \Omega^2 \\ j \neq h}} \left[\pi_j\frac{\pi_h}{1 - \pi_j} + \pi_h\frac{\pi_j}{1 - \pi_h}\right] = \quad & \pi_1\frac{\pi_2}{1 - \pi_1} + \pi_2\frac{\pi_1}{1 - \pi_2} + \\ & + \pi_1\frac{\pi_3}{1 - \pi_3} + \pi_3\frac{\pi_1}{1 - \pi_3} + \\ & + \pi_1\frac{\pi_4}{1 - \pi_4} + \pi_4\frac{\pi_1}{1 - \pi_4}\end{aligned}$$

and equivalent probabilities for $p_2, ..., p_n$. Since in this case $p_1$is the minimum price, this corresponds to the probability that p_1 is the minimum order statistic $f_{p_1}^n(p_1) = \lambda_1^n$ for every n. For prices other than the minimum, $f_{p_1}^n(p_1)$ is not trivial, e.g. the probability of p_2 being the minimum of a sample $n = 2$ is equal to the probability that we observe p_2, but not p_1:

$$f_{p_1}^2(p_2) = \pi_2\frac{\pi_3}{1 - \pi_3} + \pi_3\frac{\pi_2}{1 - \pi_3} + \pi_2\frac{\pi_4}{1 - \pi_4} + \pi_4\frac{\pi_2}{1 - \pi_4}$$

The objective is to estimate the probabilities λ_j^n from consumer search data. Define the probability that consumer i samples p_j when optimally sampling n prices as π_{ji}^n such that $\sum_{j=1}^N \pi_{ji}^n = 1$. The share of consumers who visit the store for each sample size, n, hence the firm's probability of being sampled given a consumer search process Ψ^n is

$$\hat{\lambda}_j^n = \frac{1}{M}\sum_{i=1}^M \pi_{ji}^n.$$

It follows $\sum_{j=1}^N \hat{\lambda}_j^n = 1$. For consumers who sample one store, $n = 1$,we know that

$\hat{\lambda}_j^1 = \hat{\pi}_j$ from the data is the proportion of consumers whose first visit was to store j before each transaction. Using $\hat{\pi}_j$ I can recover $\hat{\lambda}_j^n$ for $n = 2, .., N$ assuming sampling without replacement with perfect recall. For example for $n = 2$

$$\hat{\lambda}_j^2 = \sum_{\substack{\omega^2 \in \Omega^2 \\ j \neq h}} \left[\hat{\pi}_j \frac{\hat{\pi}_h}{1 - \hat{\pi}_j} + \hat{\pi}_h \frac{\hat{\pi}_j}{1 - \hat{\pi}_h} \right].$$

Notice that in the case that a consumer randomly sample prices $\hat{\lambda}_j^n = \pi = 1/N$ for every n and j.

B Data Sample Construction

This appendix describes in detail the construction of the book dataset from the ComScore data. I excluded observations from firms that could not be identified as online bookstores, such as unidentified domains and auction sites. In total, 18 percent of the sample transactions were excluded; most of these were from Ebay.com (15 percent of transactions). Although the excluded transactions represent a large number of observations, they cannot be considered sales from an online bookstore because they are auctions of potentially different books, for example used books, autographed volumes, or auctioned items. A small number of transactions from international Amazon websites (in the United Kingdom, Canada, and Denmark) were also dropped. To avoid double counting, browsing activity from Borders.com is excluded. Although initially Borders operated Borders.com, in April 2001 it signed a commercial agreement giving Amazon control of customer service, fulfillment, and inventory operations. As a result all visits to Borders.com are redirected to Amazon.com.

I restrict the sample to book transactions and eliminate all non-book transactions (i.e. a large number of periodicals, and smaller numbers of videos, DVDs, calendars, CDs, and audio books). The main difficulty is in identifying identical books at different sellers given that in some cases product description differs across firms. For example, firms may add or omit the subtitle, author, series name, publisher, edition, or year in the book description. I attempted to match the books by name whenever possible using the information available, mainly by separating book descriptors. However, to reduce errors and homogenize the remaining book names, I corrected them by visual inspection in less than 2 percent of the sample. There were some irregular observations in the data. Observations with negative prices or price or quantity equal to zero were dropped from the sample. Also, books with a price less than $2 were dropped from the sample. Under these restrictions, 8 percent of the observations were excluded.

B.1 Current Population Survey

I use weighted data from the Internet and Computer Use Supplement of the Current Population Survey from October 2003. I restrict the sample to those who have Internet access at home, are 18 years of age or older, and who claimed to have made purchases online. The resulting sample contains users with greater income and education, without a significant change in the age distribution. Those claiming Hispanic ethnicity were categorized as Hispanic regardless of race. Broadband is defined as

having DSL, cable modem, or fixed wireless connection such as MMDS. For comparison purposes, households with 6 members or more (3 percent of the sample) were considered to have 6 members. Yearly income was estimated by multiplying weekly earnings by 52.

B.2 Forrester Data

I use the Forrester Consumer Technographics Survey 2003 conducted from December 2002 to February 2003.[17] This survey contains a large array of questions about the online activities of more than 60,000 Internet users and has been used to analyze other Internet-related issues. I restrict the sample to U.S. individuals who have Internet access at home, are 18 years of age or older, and who declare they have made a purchase online in the last 3 months. In this survey, education level is for the head of household and age is for the oldest member of the household. Broadband is defined as the user having an ISDN connection, cable modem, DSL, satellite, or fixed wireless. Household size was capped at 6 members.

B.3 Zip Code Data

I estimate the number of bookstores located in a 5-mile radius of each user in the dataset using the ZIP Code Business Patterns, 2004. This corresponds to the total number of establishments in the Bookstores category, defined as "establishments primarily engaged in retailing new books" (NAICS code 451211). I calculate the number of bookstores located in a ZIP code whose centroid is located within a 5-miles radius of the user's ZIP code centroid. The centroid information was obtained from Zip Code Tabulation Area for 2000 from the U.S. Census Bureau.

[17] See Brown and Goolsbee (2002) for a detailed description of dataset and Prince (2004) for an estimation of the demand of personal computers using the Forrester survey.

References

Aguiar, M., and E. Hurst (2005): "Consumption vs. expenditure," *Journal of Political Economy*, 113, 919–48.

Baye, M., and J. Morgan (2005): "Brand and price advertising in online markets," Working paper.

Baye, M. R., J. Morgan, and P. Scholten (2004): "Price dispersion in the small and in the large: Evidence from a price comparison site," *Journal of Industrial Organization*, 52, 463–96.

Baylis, K., and J. M. Perloff (2002): "Price dispersion on the Internet: Good firms and bad firms," *Review of Industrial Organization*, 21, 305–24.

Brown, J. R., and A. Goolsbee (2002): "Does the Internet make markets more competitive? Evidence from the life insurance industry," *Journal of Political Economy*, 110, 481–507.

Brynjolfsson, E., and M. Smith (2000): "Frictionless commmerce? A comparison of Internet and conventional retailers," *Management Science*, 46, 563–85.

Burdett, K., and K. L. Judd (1983): "Equilibrium price dispersion," *Econometrica*, 51, 955–70.

Chevalier, J., and A. Goolsbee (2003): "Measuring prices and price competition online: Amazon.com and Barnes and Noble.com," *Quantitative Marketing and Economics*, 1, 203–22.

Clay, K., R. Krishnan, and E. Wolff (2001): "Prices and price dispersion on the Web: Evidence from the online book industry," *Journal of Industrial Economics*, 49, 521–39.

Clay, K., R. Krishnan, E. Wolff, and D. Fernandez (2002): "Retail strategies on the Web: Price and non-price competition in the online book industry," *Journal of Industrial Economics*, 50, 351–67.

Ellison, G., and S. F. Ellison (2004): "Search, obfuscation, and price elasticities on the Internet," Working Paper 10570, National Bureau of Economic Research.

Evans, D. L., L. M. Leemis, and J. H. Drew (2006): "The distribution of order statistics for discrete random variables with applications to bootstrapping," *Journal on Computing*, 18, 19–30.

Gallant, A. R., and D. N. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica*, 55, 363–90.

Hong, H., and M. Shum (2006): "Using price distributions to estimate search costs," *RAND Journal of Economics*, 37, 257–75.

Janseen, M. C. W., and J. L. Moraga-Gonzalez (2004): "Strategic pricing, consumer search and the number of firms," *Review of Economic Studies*, 71, 1089–118.

Klein, R., and R. Spady (1993): "An efficient semiparametric estimator for discrete choice models," *Econometrica*, 61, 387–421.

Klein, R. W., and R. P. Sherman (2002): "Shift restrictions and semiparametric estimation in ordered response models," *Econometrica*, 70, 663–91.

MacMinn, R. D. (1980): "Search and market equilibrium," *Journal of Political Economy*, 88, 308–27.

Moraga-Gonzalez, J. L., and M. R. Wildenbeest (2007): "Maximum likelihood estimation of search costs," *European Economic Review*, forthcoming.

Morgan, P., and R. Manning (1985): "Optimal search," *Econometrica*, 53, 923–44.

Pan, X., B. T. Ratchford, and V. Shankar (2003): "Why aren't the prices of the same item the same at Me.com and You.com?: Drivers of price dispersion among e-tailers," Working paper, University of Maryland.

——— (2004): "Price dispersion on the Internet: A review and directions for future research," *Journal of Interactive Marketing*, 18, 116–35.

Prince, J. (2004): "Measuring the digital divide: Structural estimation of the demand for personal computers," Working paper.

Rob, R. (1985): "Equilibrium price distributions," *The Review of Economic Studies*, 52, 487–504.

Salop, S., and J. Stiglitz (1977): "Bargains and ripoffs: A model of monopolistically competitive price dispersion," *Review of Economic Studies*, 44, 493–510.

Scholten, P., and S. A. Smith (2002): "Price dispersion then and now: Evidence from retail and e-tail markets," Working paper.

Stahl, D. O. (1989): "Oligopolistic pricing with sequential consumer search," *American Economic Review*, 74, 700–12.

——— (1996): "Oligopolistic pricing with heterogeneous consumer search," *International Journal of Industrial Organization*, 14, 243–68.

Stewart, M. (2005): "A comparison of semiparametric estimators for the ordered response model," *Computational Statistics and Data Analysis*, 49, 555–73.

Zwick, R., A. Rapoport, and A. King Chun Lo (2001): "Consumer search: Not enough or too much," Working paper.

Table 1: Consumer Browsing and Transaction Data Descriptive Statistics

	2002			2004		
	Mean	Std. Dev.	Boostrap Std. Err.	Mean	Std. Dev.	Boostrap Std. Err.
Duration of each website visit (in minutes)						
Visits not within 7 days of transaction	8.89	13.03	0.04	7.69	12.36	0.03
Visits within 7 days, excluding transactions	12.21	15.55	0.16	10.72	14.84	0.10
Visits within 7 days, excluding transactions	19.04	18.26	0.16	15.74	17.37	0.12
Transactions only	27.90	17.69	0.18	25.93	17.68	0.19
Total duration, excluding transaction visits	32.47	49.80	0.83	38.41	78.33	1.14
Total duration, including transaction visits	43.77	43.72	0.47	47.68	65.99	0.69
Number of firms searched	1.27	0.54	0.01	1.30	0.56	0.01
Number of books per transaction	2.38	2.10	0.02	2.20	1.95	0.02
Transaction expenditures (books only)	36.70	40.67	0.54	32.21	35.68	0.35
Number of books purchased	17,956			17,631		
Number of transaction sessions	7,559			8,002		
Number of visits within 7 days	18,349			25,513		
Number of visits not within 7 days	94,012			189,200		

Notes: This table presents descriptive measures of user browsing of online bookstores calculated from ComScore data for the period July–September 2002 and for the year 2004. The number of stores visited and the duration in minutes of user visits to each bookstore are summarized for the 7-day cutoff period prior to each book purchase.

Table 2: Browsing by Firm

		Book purchased from			
	All bookstores	Amazon	Barnes & Noble	Book clubs	Other bookstores
First firm searched (%)					
Amazon	65.4	91.1	23.6	19.1	28.6
Barnes & Noble	16.9	3.5	67.8	2.2	5.5
Book clubs	10.8	1.6	2.8	74.2	3.2
Other bookstores	6.9	3.8	5.8	4.5	62.7
	100	100	100	100	100
Firm searched (%)					
Amazon	73.6	--	31.3	24.0	37.6
Barnes & Noble	27.5	8.6	--	5.1	13.7
Book clubs	15.3	3.0	4.5	--	5.6
Other bookstores	13.1	8.5	11.1	8.1	--
Any other bookstore	--	17.3	39.0	30.6	45.7
Number of firms searched	1.29	1.20	1.47	1.37	1.57
	(0.56)	(0.47)	(0.64)	(0.62)	(0.70)
Number of books	35,587	22,226	7,441	4,356	1,564

Notes: This table presents search patterns related to book transactions. All searches are linked to the next transaction and are limited to a maximum of 7 days prior to each book purchase. The mean and standard deviation are presented for the number of firms searched. Book clubs include the following sites (.com): Christianbook, Doubledaybookclub, Eharlequin, Literaryguild, and Mysteryguild. Other bookstores include (.com): 1bookstreet, Allbooks4less, Alldirect, Booksamillion, Ecampus, Powells, Varsitybooks, and Walmart.

Table 3: Search and Transaction Behavior by Length of Search Period

	All bookstores	Book purchased from: Amazon	Barnes & Noble	Book clubs	Other bookstores
One firm visited					
7 days	0.76	0.82	0.61	0.71	0.53
5 days	0.79	0.85	0.65	0.76	0.57
3 days	0.82	0.87	0.70	0.81	0.61
1 day	0.86	0.90	0.76	0.88	0.66
Transaction day	0.90	0.93	0.83	0.93	0.74
Two or more firms visited					
Purchased from the last firm visited					
7 days	0.65	0.50	0.80	0.82	0.77
5 days	0.63	0.48	0.77	0.77	0.74
3 days	0.61	0.46	0.75	0.72	0.73
1 day	0.60	0.46	0.74	0.69	0.73
Transaction day	0.61	0.47	0.75	0.70	0.74
Purchased from a previously visited firm					
7 days	0.35	0.50	0.20	0.18	0.23
5 days	0.37	0.52	0.23	0.23	0.26
3 days	0.39	0.54	0.25	0.28	0.27
1 day	0.40	0.54	0.26	0.31	0.27
Transaction day	0.39	0.53	0.25	0.30	0.26
Number of transaction sessions	15,561	10,197	3,042	1,653	669

Notes: This table presents search patterns related to book transactions. All transaction session data fall into the category "one firm visited" or "two or more firms visited." All searches are linked to the next transaction and are limited to a maximum of 7, 5, 3, or 1 days prior to each book purchase, or to the same day of the transaction. The number in the first panel reflects the proportion of transaction sessions where consumers visited one firm for each of the lengthts of search periods considered. The subgroup "two or more firms visited" is further divided according to consumers' transaction strategy. For those who searched more than one firm, the numbers represent the proportion of transactions where consumers bought from the last firm they visited or the proportion that recalled a price by buying from a previously visited firm.

Table 4: Demographic Characteristics of Internet Users

Variable	ComScore Book Sample	ComScore 2002	ComScore 2004	Forrester 2003	Population CPS 2003
Number of users	9,446	38,193	24,834	28,716	10,504,092
Broadband connection	0.42	0.44	0.34	0.28	0.47
	(0.49)	(0.50)	(0.47)	(0.45)	(0.50)
Household size	2.94	3.05	2.95	2.69	3.02
	(1.36)	(1.36)	(1.34)	(1.25)	(1.29)
Children present in household	0.41	0.46	0.37	0.36	0.43
	(0.49)	(0.50)	(0.48)	(0.48)	(0.49)
Age distribution (%)					
18–20	1.9	2.7	0.7	0.2	3.3
21–24	4.2	5.3	3.6	1.2	7.7
25–29	6.4	6.9	6.4	4.9	11.3
30–34	9.5	9.8	10.7	8.2	13.9
35–39	9.0	8.6	11.1	11.7	13.9
40–44	11.9	10.7	14.8	14.4	14.4
45–49	16.0	17.7	15.2	13.3	13.7
50–54	15.9	16.2	13.4	14.1	10.5
55–59	9.1	8.0	8.9	12.2	7.1
60–64	7.1	6.6	6.2	8.0	3.1
65 and over	9.0	7.6	8.9	11.7	1.1
Household income distribution (%)					
Less than $15,000	4.7	5.1	5.7	3.9	12.8
$15,000 – $24,999	8.5	9.9	8.9	5.7	15.9
$25,000 – $34,999	13.7	14.8	15.7	7.8	15.8
$35,000 – $49,999	19.8	20.1	20.8	12.8	22.6
$50,000 – $74,999	26.3	26.1	25.3	18.3	18.8
$75,000 – $99,999	13.2	12.1	12.2	26.7	8.0
More than $100,000	13.9	12.1	11.3	24.9	6.2
Education distribution (%)					
Number of observations	6,573	27,148	16,108	28,716	10,504,092
Less than high school	1.29	1.5	2.7	1.7	1.8
High school diploma or GED	13.91	16.0	22.0	11.6	17.8
Some college but no degree	30.15	36.6	31.6	18.5	20.2
Associate degree	10.86	10.5	12.3	7.0	10.7
Bachelor's degree	26.18	22.1	20.6	32.0	32.6
Graduate degree	17.60	13.4	10.9	29.3	16.9
Race (%)					
White	81.5	81.3	74.7	88.3	81.3
Black	4.3	4.7	7.3	4.4	5.7
Hispanic	8.8	7.9	13.1	4.6	5.8
Asian	3.1	3.3	2.6	2.2	5.9
Other	2.4	2.8	2.3	0.5	1.2
Region of residence (%)					
Northeast	21.5	19.3	19.2	21.7	21.9
Midwest	22.0	24.4	22.6	24.4	23.4
South	32.7	34.3	35.7	32.5	31.0
West	23.8	22.0	22.6	21.5	23.7

Sources: ComScore Web-Behavior Panel dataset (June 2002–December 2002); 2003 Forrester Technographics Consumer Survey; and the Internet and Computer Use Supplement, CPS October 2003.

Notes: Standard deviations are shown in parentheses. The sample is restricted to users located in the United States who access the Internet at home. CPS data is weighted. Those claiming Hispanic ethnicity were categorized as Hispanic regardless of race. For ComScore and Forrester data, education level is for the head of household, not necessarily the oldest member; for CPS education level is the respondent's. For ComScore and Forrester, age refers to the oldest member of household; for CPS it is the age of the respondent. For expositional simplicity, households with 6 members or more (3 percent of the sample) were considered to have 6 members.

Table 5: Descriptive Statistics for Selected Books

		Prices ($)				Consumers by Sample Size (%)		
Product name	Obs.	Mean	Std. Dev.	Min	Max	q_1	q_2	q_3
Best sellers 2004								
The Da Vinci Code	52	14.3	1.1	12.7	15.0	0.712	0.250	0.038
Trace	21	16.2	2.5	13.9	18.9	0.524	0.429	0.048
R Is for Ricochet	21	16.9	4.7	11.7	20.9	0.667	0.286	0.048
3rd Degree	18	16.8	2.1	14.7	18.9	0.833	0.111	0.056
Key of Valor	10	7.5	2.6	4.9	11.0	0.600	0.300	0.100
State of Fear	9	16.0	0.8	15.2	16.8	0.556	0.333	0.111
Best sellers 2002								
From a Buick 8	37	17.5	2.2	15.1	19.6	0.730	0.243	0.027
Four Blind Mice	35	17.1	2.3	15.1	19.2	0.800	0.171	0.029
Nights in Rodanthe	17	13.9	1.1	12.9	15.1	0.824	0.059	0.118
The Lovely Bones	14	14.0	4.8	10.0	21.0	0.643	0.214	0.143
Harry Potter and the Goblet of Fire	14	10.9	4.7	6.0	16.0	0.786	0.143	0.071
Visions of Sugar Plums	14	13.2	0.6	12.6	13.8	0.857	0.000	0.143

Notes: This table presents descriptive statistics for the books with the largest online sales in the sample each year. N represents the maximum number of bookstores with price data. q_n represents the proportion of consumers that visited n=1,...,N bookstores.

Table 6: Empirical Non-Sequential Search Cost CDF for Selected Books

	Equal Sampling		Unequal Sampling			
Product name	Δ_1	Δ_2	Δ_1	Δ_2	$G(\Delta_1)$	$G(\Delta_2)$
Best sellers 2004						
The Da Vinci Code	1.124	0.512	0.557	0.019	0.288	0.038
Trace	1.980	0.988	1.012	0.108	0.476	0.048
R Is for Ricochet	4.129	2.061	2.102	0.127	0.333	0.048
3rd Degree	1.714	0.855	0.875	0.078	0.167	0.056
Key of Valor	1.170	0.587	0.316	0.144	0.400	0.100
State of Fear	0.697	0.348	0.520	0.100	0.444	0.111
Best sellers 2002						
From a Buick 8	3.261	0.753	1.242	0.083	0.270	0.027
Four Blind Mice	3.284	0.650	1.224	0.034	0.200	0.029
Nights in Rodanthe	0.884	0.441	0.779	0.194	0.176	0.118
The Lovely Bones	5.914	0.679	4.557	1.016	0.357	0.143
Harry Potter and the Goblet of Fire	2.858	1.432	1.595	1.091	0.214	0.071
Visions of Sugar Plums	0.538	0.270	0.111	0.020	0.143	0.143

Notes: For each of the products listed, the number of firms searched is N=*3*. I can only report the quantile estimates of the search cost distribution for *n=1,2* defined by $G(\Delta_1)=1-q_1$, $G(\Delta_2)=1-q_1-q_2$, $G(\Delta_3)=0$. The search cutoffs are the expected gain of additional search Δ_n. Each expected price was measured by averaging over 100,000 iterations. The minimum of N prices was obtained by sampling without replacement from the empirical price distribution.

Table 7: Regression of Search Duration on Household Characterisitcs

	(1)		(2)		(3)		(4)	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Number of unique firms visited	29.447	(0.779)***	18.217	(1.121)***	18.551	(0.619)***	12.007	(0.883)***
Number of books bought	3.215	(0.221)***	1.896	(0.375)***				
First transaction indicator	7.149	(1.647)***	5.041	(2.641)*	5.301	(1.307)***	3.777	(2.080)*
Same bookstore as previous transaction	7.932	(1.688)***	11.147	(2.704)***	6.508	(1.340)***	9.083	(2.130)***
Free shipping (sales ≥ $25)	6.513	(0.937)***	3.287	(1.610)**	-6.722	(0.700)***	-6.882	(1.198)***
Cumulative book transactions	0.399	(0.037)***	0.467	(0.053)***	0.309	(0.030)***	0.363	(0.042)***
Household size	1.275	(0.431)***	1.570	(0.721)**	0.636	(0.343)*	0.643	(0.568)
Broadband connection	-0.683	(0.867)	0.260	(1.474)	-0.613	(0.689)	-0.008	(1.162)
Children present in household	-0.734	(1.170)	-1.287	(1.978)	-0.472	(0.929)	-0.496	(1.559)
Age								
18–20	-5.211	(3.651)	-4.799	(6.280)	-1.460	(2.899)	-2.034	(4.949)
21–24	-1.247	(2.530)	-0.903	(4.389)	-0.964	(2.009)	-1.884	(3.458)
25–29	-1.343	(2.072)	-0.101	(3.528)	-0.380	(1.645)	-0.095	(2.780)
30–34	1.449	(1.780)	3.714	(2.996)	1.039	(1.414)	2.606	(2.361)
35–39	1.220	(1.791)	1.853	(3.045)	-0.321	(1.422)	-0.241	(2.399)
40–44	-1.036	(1.629)	0.142	(2.774)	-1.085	(1.293)	-0.595	(2.186)
50–54	-1.045	(1.504)	-1.213	(2.557)	-0.803	(1.194)	-0.290	(2.015)
55–59	0.333	(1.720)	1.200	(2.931)	0.480	(1.366)	0.987	(2.310)
60–64	4.989	(1.870)***	11.162	(3.216)***	1.181	(1.484)	4.009	(2.533)
65 and over	5.588	(1.720)***	8.500	(2.932)***	3.610	(1.366)***	5.866	(2.310)**
Household income								
Less than $15,000	1.320	(2.182)	-1.820	(3.678)	1.719	(1.733)	-0.260	(2.899)
$15,000 – $24,999	6.162	(1.756)***	7.145	(2.977)**	3.511	(1.394)**	3.053	(2.346)
$25,000 – $34,999	4.310	(1.405)***	3.725	(2.401)	2.490	(1.115)**	1.413	(1.892)
$35,000 – $49,999	2.392	(1.262)*	3.672	(2.163)*	1.583	(1.002)	2.527	(1.704)
$75,000 – $99,999	2.694	(1.417)*	4.623	(2.410)*	1.631	(1.125)	2.786	(1.899)
More than $100,000	-4.174	(1.399)***	-3.116	(2.415)	-3.005	(1.111)***	-2.997	(1.903)
Education								
Less than high school	15.141	(5.061)***	13.440	(8.070)*	11.528	(4.019)***	12.338	(6.360)*
High school diploma or GED	5.935	(1.682)***	7.577	(2.888)***	2.295	(1.335)*	2.603	(2.276)
Some college but no degree	3.225	(1.367)**	3.884	(2.360)*	0.825	(1.086)	0.867	(1.860)
Associate degree	1.595	(1.838)	2.906	(3.148)	1.542	(1.460)	1.692	(2.479)
Graduate degree	-1.107	(1.525)	-0.941	(2.636)	-0.510	(1.211)	0.010	(2.077)
Race								
Black	7.540	(2.203)***	7.552	(3.771)**	6.058	(1.750)***	5.381	(2.972)*
Hispanic	-0.747	(1.527)	-2.808	(2.588)	-0.641	(1.212)	-2.434	(2.039)
Asian	13.671	(2.580)***	18.748	(4.241)***	9.681	(2.048)***	11.805	(3.341)***
Other	6.450	(2.841)**	10.327	(4.838)**	6.327	(2.256)***	9.336	(3.813)**
Region of residence								
Northeast	1.380	(1.181)	0.997	(1.991)	1.846	(0.938)**	1.770	(1.569)
Midwest	1.680	(1.172)	3.310	(2.007)*	1.116	(0.931)	2.702	(1.582)*
West	1.302	(1.150)	1.099	(1.986)	0.785	(0.913)	0.997	(1.565)
Constant	-22.941	(2.826)***	-23.716	(4.797)***	-4.624	(2.229)**	-7.391	(3.750)**
Years	2002, 2004		2002, 2004		2002, 2004		2002, 2004	
Transaction visits	Yes		No		Yes		No	
Multiple books	Yes		Yes		Avg. duration/book		Avg. duration/book	
R-squared	0.13		0.06		0.08		0.05	
Number of individuals	15561		8206		15561		8206	

Notes: This table presents regression estimates of the duration of search on consumer characteristics. The dependent variable is the duration in minutes of user visits to each bookstore for the 7-day cutoff period prior to each book purchase. The number of firms visited is the unique number of bookstores browsed during this period, 1 to 4 firms. "First transaction" indicates the first observation in the dataset for the user. "Cumulative book transactions" are the number of book purchases prior to the current one. "Number of nearby bookstores" corresponds to the total number of bricks and mortar bookstores located in a ZIP code within a 5-mile radius of the user's ZIP code address obtained from ZIP Business Patterns, 2004. The estimation includes non-reported state of residency indicator variables. Standard errors are robust clustered by user. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 8: Ordered Probit Estimates of Sample Size on Household Characterisitcs

	(1)		(2)		(3)		(4)	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
First transaction indicator			-0.612	(0.040)***	-0.541	(0.051)***	-0.461	(0.075)***
Same bookstore as previous transaction			-0.705	(0.042)***	-0.632	(0.054)***	-0.510	(0.082)***
Free shipping (sales ≥ $25)			0.012	(0.027)	0.048	(0.037)	-0.008	(0.053)
Multiple books			-0.012	(0.026)	0.063	(0.034)*		
Price							0.003	(0.001)**
Number of nearby bookstores					-0.047	(0.047)	-0.032	(0.058)
Cumulative book transactions	0.009	(0.002)***	0.008	(0.002)***	0.006	(0.002)***	0.004	(0.002)***
Household size	0.066	(0.018)***	0.064	(0.017)***	0.086	(0.024)***	0.056	(0.031)*
Broadband connection	0.111	(0.029)***	0.108	(0.028)***	0.154	(0.041)***	0.135	(0.053)**
Children present in household	-0.030	(0.046)	-0.030	(0.044)	-0.074	(0.062)	0.063	(0.079)
Age								
18–20	0.018	(0.104)	0.014	(0.102)	0.015	(0.300)	-0.039	(0.534)
21–24	0.028	(0.074)	0.010	(0.071)	-0.061	(0.103)	-0.029	(0.142)
25–29	0.016	(0.063)	0.010	(0.061)	0.012	(0.089)	0.141	(0.112)
30–34	0.127	(0.058)**	0.117	(0.055)**	0.171	(0.077)**	0.231	(0.098)**
35–39	0.127	(0.059)**	0.112	(0.057)*	0.148	(0.076)*	0.134	(0.098)
40–44	0.094	(0.053)*	0.087	(0.052)*	0.051	(0.069)	0.007	(0.088)
50–54	0.053	(0.049)	0.042	(0.048)	0.021	(0.072)	0.094	(0.092)
55–59	0.147	(0.059)**	0.135	(0.057)**	0.117	(0.084)	0.127	(0.106)
60–64	0.144	(0.081)*	0.128	(0.076)*	0.105	(0.119)	0.167	(0.158)
65 and over	-0.018	(0.061)	-0.023	(0.057)	-0.079	(0.074)	-0.045	(0.099)
Household income								
Less than $15,000	0.004	(0.065)	0.007	(0.063)	-0.048	(0.085)	-0.107	(0.115)
$15,000 – $24,999	0.086	(0.053)	0.085	(0.052)	0.061	(0.072)	0.051	(0.091)
$25,000 – $34,999	0.017	(0.050)	0.025	(0.048)	0.044	(0.064)	0.138	(0.080)*
$35,000 – $49,999	0.062	(0.045)	0.052	(0.043)	0.058	(0.059)	0.118	(0.075)
$75,000 – $99,999	0.017	(0.048)	0.027	(0.047)	-0.034	(0.064)	-0.011	(0.080)
More than $100,000	-0.115	(0.048)**	-0.100	(0.047)**	-0.109	(0.067)	-0.170	(0.087)**
Education								
Less than high school	0.065	(0.149)	0.082	(0.143)	0.098	(0.191)	0.390	(0.245)
High school diploma or GED	0.055	(0.066)	0.054	(0.063)	0.125	(0.082)	0.127	(0.103)
Some college but no degree	0.049	(0.044)	0.059	(0.042)	0.045	(0.060)	0.014	(0.079)
Associate degree	-0.035	(0.064)	-0.007	(0.063)	-0.089	(0.088)	-0.092	(0.112)
Graduate degree	-0.023	(0.053)	-0.018	(0.051)	0.037	(0.074)	0.107	(0.094)
Race								
Black	0.092	(0.064)	0.093	(0.063)	0.077	(0.089)	-0.011	(0.116)
Hispanic	-0.071	(0.053)	-0.051	(0.050)	-0.067	(0.065)	-0.013	(0.081)
Asian	0.174	(0.090)*	0.171	(0.084)**	0.208	(0.127)	-0.025	(0.141)
Other	0.075	(0.092)	0.070	(0.089)	-0.011	(0.139)	-0.023	(0.146)
Region of residence								
Northeast	0.045	(0.038)	0.042	(0.037)				
Midwest	0.017	(0.042)	0.019	(0.040)				
West	-0.133	(0.040)***	-0.129	(0.038)***				
Years	2002, 2004		2002, 2004		2004		2004	
State dummies	No		No		Yes		Yes	
Region dummies	Yes		Yes		No		No	
Multiple books	Yes		Yes		Yes		No	
Pseudo R2	0.0149		0.0315		0.0402		0.0404	
Number of individuals	8985		8985		4157		2528	
Number of observations	15561		15561		7962		3731	

Notes: The dependent variable is the number of bookstores visited for each transaction (*n=1,...,4*). All searches are linked to the next transaction and occur no more than 7 days prior to each book purchase. "First transaction" indicates the first observation in the dataset for the user. "Cumulative book transactions" are the number of book purchases prior to the current one. "Number of nearby bookstores" corresponds to the total number of bricks and mortar bookstores located in a ZIP code within a 5-mile radius of the user's ZIP code address obtained from ZIP Business Patterns, 2004. The estimation includes non-reported state of residency indicator variables. Standard errors are robust clustered by user. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 9: Geographic Distribution by State of Internet Users

State	ComScore 2004	Population CPS 2003	State	ComScore 2004	Population CPS 2003
Alabama	1.1	0.9	Montana	0.4	0.3
Alaska	0.3	0.3	Nebraska	0.6	0.8
Arizona	1.6	1.9	Nevada	0.9	0.8
Arkansas	0.9	0.6	New Hampshire	0.6	0.7
California	11.5	11.9	New Jersey	3.2	3.6
Colorado	2.0	2.1	New Mexico	0.9	0.4
Connecticut	1.2	1.7	New York	7.6	7.0
Delaware	0.4	0.3	North Carolina	3.0	2.5
District Of Columbia	0.3	0.4	North Dakota	0.2	0.3
Florida	6.2	5.1	Ohio	4.1	4.0
Georgia	2.7	2.9	Oklahoma	1.3	0.8
Hawaii	0.3	0.4	Oregon	2.0	1.7
Idaho	0.5	0.6	Pennsylvania	5.4	4.6
Illinois	3.7	3.9	Rhode Island	0.3	0.5
Indiana	1.7	1.5	South Carolina	1.4	1.1
Iowa	1.2	1.1	South Dakota	0.2	0.3
Kansas	0.9	1.0	Tennessee	2.0	1.5
Kentucky	1.1	1.6	Texas	5.7	5.9
Louisiana	1.2	0.9	Utah	0.8	0.9
Maine	0.7	0.4	Vermont	0.3	0.3
Maryland	1.7	2.4	Virginia	3.0	3.1
Massachusetts	2.1	3.1	Washington	2.3	2.3
Michigan	3.2	3.6	West Virginia	0.6	0.5
Minnesota	1.8	2.3	Wisconsin	1.8	2.5
Mississippi	0.7	0.7	Wyoming	0.3	0.2
Missouri	1.9	2.2			
Number of observations	4157	10,504,092			

Sources: ComScore Web-Behavior Panel dataset (June 2002–December 2002), 2003 Forrester Technographic Consumer Survey and the Internet and Computer Use Supplement, CPS October 2003.

Notes: The sample is restricted to users located in the United States who access the Internet at home. CPS data is weighted.