Consumer Search on the Internet

Babur De los Santos*

John E. Walker Department of Economics
Clemson University

November 2017

Abstract

This paper analyzes search frictions in online markets using data depicting the web browsing and purchasing behavior of a large panel of consumers. In this data, consumer search behavior is observed prior to a transaction. I use data on consumers shopping for books online to link prices and consumer search patterns at different bookstores, estimating consumer search costs in the context of a fixed-sample search model. The search patterns indicate that consumers visit relatively few firms and exhibit a strong search preference for prominent retailers. I control for search intensities at different retailers during consumers' search process and find that search cost estimates are lower than when assuming consumers sample equally among alternatives. Accounting for heterogeneity in consumer search intensities across retailers reduces search cost estimates from $2.30 to $1.24 per search. I examine search cost heterogeneity by using a rich set of consumer characteristics and relating them to search patterns and search costs estimates. I use a flexible random effects model in which the number and order of firms visited by the consumer are her optimal ordered choices, allowing search cost cutoffs to depend on regressors. The estimates indicate that consumer search costs are related to their observable characteristics, such as income, where individuals with income greater than $100,000 incur relatively higher search costs.

*I am grateful to the editor José Luis Moraga-González and two anonymous reviewers for their careful comments. I also benefited from comments from Ali Hortaçsu, Jean-Pierre Dubé, Lura Forcum, Jeremy Fox, Matt Gentzkow, Austan Goolsbee, Günter Hitsch, Steven Levitt, Jesse Shapiro, Chad Syverson, Matthijs Wildenbeest, and seminar participants at the Workshop on Consumer Search, the University of Chicago, Indiana University, University of Michigan, the Federal Reserve Board of Kansas City. I gratefully acknowledge financial support from the NET Institute (www.netinst.org), the Kauffman Foundation, and CONACYT.

1 Introduction

Growing evidence suggests consumer search and other informational frictions soften competition in many product markets, even if such markets are populated by a large number of seemingly undifferentiated competing sellers. As a result, it is often observed that the same product is sold for different prices across retailers. Early research that analyzed this phenomenon in online markets found that price dispersion was always present, even though competing retailers are only a click away.[1]

A strand of empirical literature on search frictions quantifies search costs, mainly in homogenous product markets. This literature proposes methodologies to leverage either aggregate or individual data on prices, quantity, and consumer search behavior. For instance, Hong and Shum (2006) use only price data on textbooks to structurally estimate parameters of the search cost distribution. In order to identify search costs using only prices, they propose a model that exploits market equilibrium conditions that rationalize prices set by competing stores. Moraga-González and Wildenbeest (2008) extend this model to the oligopoly case and propose a new methodology to estimate search cost distributions and estimate the model using online price data from the computer memory chip industry. In addition to price data, Hortaçsu and Syverson (2004) utilize market shares in the mutual fund industry to estimate search cost distributions using firms' optimality conditions. Kim, Albuquerque, and Bronnenberg (2010) employ aggregate product search data to estimate search and online demand for durable goods. Moraga-González, Sándor, and Wildenbeest (2015) use similar aggregate product search data to estimate a discrete-choice model which incorporates costly consumer search in the automobile market. A more recent strand of papers quantifies search costs when consumer search behavior and choice data are observed (De los Santos, Hortaçsu, and Wildenbeest, 2012, 2017; Koulayev, 2014; Honka, 2014; Pires, 2015; Seiler, 2013). A common assumption in this literature is that consumers search firms randomly, even when there is some degree of product differentiation.

This paper contributes to this empirical literature by extending the fixed-sample search model of Burdett and Judd (1983) to allow for unequal sampling of firms. I show that the proposed fixed-sample search model is suitable to fit an ordered choice model when individual search behavior is

[1] Baye and Morgan (2009) provide a good summary of this research.

observed. Individual consumer search data allow me to estimate search cost distributions without invoking the pricing equilibrium conditions necessary to identify the search cost distribution function of Hong and Shum (2006) or Hortaçsu and Syverson (2004). This empirical methodology also allows me to link individual search behavior to consumer characteristics and hence directly explore the determinants of search cost heterogeneity.

In Section 2, I present background information of the online book industry, which is the focus of this paper. The advantage of using the online book industry to study search frictions is that book titles across retailers are homogeneous, and book retailing is a mature online industry. However, there is substantial retailer heterogeneity. The industry is highly concentrated, with Amazon and Barnes & Noble capturing 83 percent of book sales. Also, consumers exhibit strong search preferences for these retailers. In only 25 percent of transactions did consumers visit more than one bookstore at any point in their search. Amazon was visited in 74 percent of all book transactions, and in only 17 percent of transactions did Amazon buyers browse any other bookstore. In contrast, Barnes & Noble buyers searched other bookstores 39 percent of the time.[2] A limitation of the data is that I only observe search behavior at the domain level and visits to product pages are not observed. In contrast, Bronnenberg, Kim, and Mela (2016) use a unique data set constructed to record all consumer browsing activities for cameras across- and within websites. They find prior to buying a digital camera, a consumer conducts an average of 14 online searches, over 3 brands, 6 models and 4 domains. They also find that search is confined to a surprisingly small region of the attribute space and consumer's path through the attribute space displays strong state dependence which is consistent with sequential search and learning about preferences.

In Section 3, I describe the data on the web browsing and purchasing behavior of a large panel of consumers. I employ these data to obtain consumer search patterns in the online book industry prior to a transaction. These search patterns indicate that consumers visit a very small number of firms and that consumers visit retailers at different rates. Only 25 percent of searches leading to a transaction are from consumers who 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 the transactions,

[2]Despite the limitations of studying search behavior in a concentrated industry, a large number of consumers in the data have bought books online. Perhaps more important for estimation, a number of book titles have been purchased by more than one individual. The advantage of observing transactions across consumers is that it allows me to obtain a better picture of the price distribution from purchases of the same book title at different retailers.

consumers visited another bookstore before completing their transaction at Amazon. In contrast, about 39 percent of Barnes & Noble's customers visited another bookstore, mainly Amazon. These patterns reject the standard random sampling assumption common within the empirical literature on search costs.

I also examine the sources of search cost heterogeneity by studying the relationship between a rich set of consumer demographic characteristics and search duration. An important source of search cost heterogeneity is the relative value of time for different individuals, as implied by their socio-demographic characteristics. I find that search duration decreases with income and is greater for retirement-age individuals. These results are consistent with the insight that the duration of search is an important mechanism related to an individual's search costs. For example, Aguiar and Hurst (2005) found that at the time of retirement, individuals reduce food expenditures without an equivalent reduction in the quantity or quality of food consumption. The discrepancy between expenditure and consumption is explained by retirees spending more time searching for food.

In Section 4, I present the fixed-sample equilibrium search model based on Burdett and Judd (1983) and generalized by Hong and Shum (2006). Of the two main search paradigms set forth in the literature following Stigler's (1961) original model, under fixed-sample search a consumer optimally decides the number of stores to search and buys from the lowest priced alternative of this set of stores. The choice of a fixed-sample search model instead of a sequential search model derives from analytical convenience and empirical support more than theoretical dominance. De los Santos, Hortaçsu, and Wildenbeest (2012) present evidence that supports the use of a fixed-sample search model over a sequential model in this industry. Empirically adapting Burdett and Judd's model requires the use of data on sales. This is important contrast to papers which use only price data to identify the search cost distribution.

In Section 5, I present estimates of search costs under two search behavior scenarios. First, I link prices and consumer search patterns at different online bookstores to estimate consumer search costs in the context of a search model assuming random sampling of retailers. Second, I estimate search costs under unequal sampling based on observed search intensities at different retailers during the consumer search process. I find that search costs estimated under unequal sampling are consistently smaller. The reason is that unequal sampling results in smaller gains from search than when we assume consumers randomly sample from available alternatives. Accounting for unequal

consumer search reduces search cost estimates from $2.30 to $1.24 per search. These results suggest that earlier estimation methods are more appropriate in symmetric market settings.

In Section 6, I examine the sources of consumer search cost heterogeneity. I use a rich set of consumer demographic characteristics to analyze the relationship of search patterns and estimated search costs to observables, such as consumers' income. I propose an ordered probit model in which the number of firms visited by the consumer relates to optimal ordered choices where covariates have the same effect across alternatives. I extend this specification to allow for unobserved heterogeneity by including random coefficients on price and the interaction of price and consumer characteristics. Furthermore, I extend the model to allow search cost cutoffs to depend on regressors.[3] The estimates indicate that consumer search costs are related to observable characteristics in an intuitive way. For example, individuals with income greater than $100,000 have relatively higher search costs, which translates to fewer stores visited and less time spent during each visit. Retired people, those with lower education levels, and minorities (with the exception of Hispanics) spent significantly more time searching for books online. The search cost distribution function, obtained from the ordered probit specification, is bimodal with modes around $1.25 and $2.25, which is consistent with the search cost estimates from the empirical search cost distribution. These results relate to those of Moraga-González, Sándor, and Wildenbeest (2015), who estimate search costs of consumers in the automobile market and find that automobile dealers' distances from a consumer's home play an important part in search costs. These authors also find a positive relationship between income and search costs. Additional evidence of search cost heterogeneity is provided by Nishida and Remer (2018) who find significant variation of search costs across local retail gasoline markets and across household income.

Finally, this paper is also related to a strand of research which documents search frictions in online markets. This research typically measures price dispersion in online price comparison platforms, which present consumers with information about prices and product availability for various online retailers.[4] Unfortunately these data typically do not have transaction or sales information. As a result, most studies of online price dispersion weigh prices from different firms equally and assume that sales occur at each observed price. For instance, using data from the eight largest book

[3] A more flexible example of ordered choice models is presented by Cunha, Heckman, and Navarro (2007).

[4] See, e.g., Clay et al. (2001, 2002), Baye, Morgan, and Scholten (2004), Brynjolfsson and Smith (2000), and Ellison and Ellison (2009).

retailers, Brynjolfsson and Smith (2000) found significantly lower price dispersion when controlling for firms' market shares. The observed consumer search patterns found in this paper do not support random sampling of retailers, as online markets are often highly concentrated with few dominant firms and numerous fringe retailers, which could lead to overestimation of search frictions when only price data are used.[5] This evidence suggests that search frictions in online markets might be smaller than prior studies suggested, when we use price and sales data of retailers visited by consumers. This follows Wildenbeest (2011), who shows that a large part of the observed price dispersion can be explained by firm heterogeneity rather than search frictions. Hence, ignoring the vertical product differentiation component could lead to overestimation of search frictions.

2 Online Book Industry

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 grew to represent 10 percent of the total sales of the $28 billion book industry in 2004 and 17 percent two years later.[7] In 2002, Amazon's sales of media (books, music, and DVDs) for North America totaled $1.87 billion. For the same year, eBays annualized gross merchandise net sales were 1.4 billion dollars and Barnes & Nobles total bookstore sales were $3.92 billion. 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 to 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 & Noble (17 percent).[8]

The expansion of e-commerce has motivated a large body of research that analyzes search frictions mainly through measures of price dispersion in online markets. There is a general notion in these studies, which use predominantly price comparison websites, that substantial price dispersion persists in a large number of online markets.[9] Brynjolfsson and Smith (2000) report that prices

[5] In fact, most online bookstores, which post prices in these online price comparison platforms, are ignored by consumers in their searches. Only 15 out of about 230 online bookstores listed in the Yahoo directory had book sale transactions in this sample, and visits to these 15 stores account for 98 percent of all consumer visits to bookstores.

[6] See, e.g., Clay, Krishnan, and Wolff (2001) for a review of this industry.

[7] Figures from Noam (2009).

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

[9] See Pan, Ratchford, and Shankar (2004) for an excellent review of the research that studies online price dispersion.

differ by an average of 33 percent for 20 books sold at the eight online bookstores with the largest number of visitors. Clay, Krishnan, and Wolff (2001) find substantial price dispersion using prices for 399 books by 32 online bookstores between August 1999 and January 2000. They found that price dispersion, measured as the standard deviation as a percentage of the average price, was 27.7 percent for New York Times Best Sellers and 12.9 percent for a random book in their sample. These studies show that online price dispersion is higher than dispersion among traditional brick and mortar retailers (e.g., Clay, Krishnan, Wolff, and Fernandez, 2002; Scholten and Smith, 2002; Pan, Ratchford, and Shankar, 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, Krishnan, and Wolff (2001) find that Amazon, Barnes & Noble, and Borders had the lowest standard deviation of price, in contrast to the large dispersion found in 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 pricing strategies of large bookstores. Their results indicate that the average price differences with Amazon for Barnes & Noble and Borders were -$0.19 and $0.09, respectively. Clay, Krishnan, and Wolff (2001) calculate that for a sample of 399 books, 77 percent of Barnes & Noble's prices and 75 percent of Borders' prices are different from Amazon's prices by 1 percent or less.[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. 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. Clay, Krishnan, and Wolff (2001) also show that small online bookstores have varied price strategies. Most of these stores set prices slightly lower (around $0.10 on average below Amazon's prices). In some cases, small bookstores set prices above Amazon's prices. 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

[10]For the market of consumer electronics, Baye, Morgan, and Scholten (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.

this limitation using prices and sales rank data from Amazon and Barnes & 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 suggests higher-quality firms command higher prices (see, e.g., Baylis and Perloff, 2002; Pan, Ratchford, and Shankar, 2003).

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

3 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 channels user's online activity through their proxy servers which allows it to record all Internet traffic, including information on website visits and secure online transactions. Consumer online browsing data include the date, time, and duration of a visit. If there is a transaction during the visit, the data includes the price, quantity, and description of each product purchased during the session.

The dataset is similar to De los Santos, Hortaçsu, and Wildenbeest (2012) and contains users' transactions for products and services from June 2002 to December 2002 and for the full year of 2004. The sample includes 35,587 book purchases from 15 online bookstores. If multiple copies of a book title were purchased during a session, it is recorded as one observation. The sample excludes book transactions 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).

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. In order to identify search behavior, I link consumer's browsing activity of online bookstores up to 7 days prior to a transaction, which results in 43,862 search

sessions. I believe that seven days is a large enough search time span to capture most of consumer search behavior related to a book purchase. As search behavior could be shorter than the 7 day window, it is likely that some visits that do not lead to a transaction might be incorrectly linked to the next transaction which would overestimate the amount of consumer search. Hence, I will also consider shorter search windows in the analysis. In addition, as the search window is meant to capture all search activity of a consumer for a particular transaction, the actual search span would be less if another transaction has occurred before 7 days. The average time between transactions is 2.9 days. Some consumer browsing may not be related to the next observed transaction, but to a later one. Since I cannot identify the product being searched when there is no transaction an intervening search cannot be linked to a later transaction. If a user searches for book A, but buys book B first, this search activity is linked to book B. If the consumer buys book A at a later date, only the search activity after buying book B is linked to book A.

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 & Noble in the market might explain the low levels of consumer search: on average users searched 1.2 bookstores in 2002 and 1.3 bookstores 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.

3.1 Consumer Search Patterns

Search behavior provides insight into the nature of consumer awareness, brand recognition, and preference for some firms. Amazon and Barnes & Noble captured 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 on average 1.29 online bookstores.

In order to analyze consumer search of online bookstores, I group small bookstores into two categories to create four firms: Amazon (63 percent of transactions), Barnes & 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 if 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, 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. As a result, search behavior within the Amazon platform is unobserved, in which case the search cost estimates might be considered a lower bound. However, there are several reasons why the potential bias is not likely to be large. First, these sellers did not likely represent a large share of transactions in 2002 and 2004. Amazon Marketplace was launched in November 2000 and according to Amazon's financial reports for the third quarter of 2002, third-party seller transactions represented 23 percent of North American sales units. This figure included new, used, and refurbished items in several product categories in addition to books. Second, in those years the book offerings of third-party sellers focused on used books. Amazon usually had the lowest prices on new books, especially bestsellers, which are heavily discounted (De los Santos and Wildenbeest, 2017).

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 to the firm's market shares: Amazon was visited first in 65 percent of the sample; Barnes & 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 & Noble buyers who visited Barnes & Noble first. A significant share of consumers of Barnes & Noble, Book clubs and Other bookstores visit Amazon first in their search process (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 of all book transactions, and in only 17 percent of transactions did Amazon buyers browse any other bookstore. In contrast, Barnes & Noble buyers searched other bookstores (mainly Amazon) 39 percent of the time; Book club shoppers, 31 percent of the time; and Other bookstore shoppers, 46 percent of the time. The limited search process is reflected in the number of stores that consumers search during each transaction. On average, Amazon buyers search 1.2 bookstores, compared to Barnes & Noble (1.5), Book clubs (1.4), and Other bookstores (1.6). The asymmetric nature if search patterns in this industry suggest that there might be important retailer differences. Although I assume that books are homogenous products for modeling convenience. The same book title purchased at different retailers might be considered a differentiated product if there is variation in retailer quality (e.g. transaction convenience, service, shipping, returns, etc.). For instance, De los Santos, Hortaçsu, and Wildenbeest (2012) allow product differentiation for books that mainly arise from retailer heterogeneity.

3.2 Patterns of the Search Stopping Decision

I use observed patterns of consumer search to shed some light on the differences in features of common search rules found in the literature. In particular, following De los Santos, Hortaçsu, and Wildenbeest (2012), 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 to 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 is 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 have visited all stores without observing a price below the reservation price. In contrast, in a fixed-sample 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, particularly a consumer's 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, 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 search window lengths. 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 day prior to each book purchase, or to the day of the transaction.[11]

The table reiterates the previous findings that consumer search in this industry is very limited and that consumers do not randomly sample different retailers. The first column shows that in 76 percent of all transactions, consumers visited one bookstore when a search was performed 7 days prior to a transaction. The proportion of people that search a store increases as we shorten the length of the search period. When we consider search on the transaction day, consumers visited only one store in 90 percent of cases. 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 a 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, either by purchasing a book from a last firm they visited or by purchasing a book from their

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

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 recalled a firm they have previously visited.

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 & Noble's consumers recall its price in only 20 to 26 percent of the time. 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 & 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 when 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 exhaust their search by visiting all their known stores—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 that use sequential search models have found similar results. This systematic recall supports the fixed-sample search process presented here. However, there are other models that might explain this, such as directed search.[12] A more extensive analysis of the contrasting implications of fixed-sample and sequential search models are tested by De los Santos, Hortaçsu, and Wildenbeest (2012) using data on consumer search.

[12] Zwick, Rapoport, Lo, King, and Muthukrishnan (2003) 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.

3.3 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.[13] 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 discrepancy in education level is due to the large proportion of college students (those with "Some college but no degree") in the ComScore sample. Online users are predominantly white, ComScore oversamples Hispanics and Forrester oversamples whites compared to the CPS.

As shown in 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.

[13]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.

3.4 Search Duration

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 explored above; 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 5 presents regression estimates of the total time spent searching for a book based on consumer characteristics. The total duration of search is presented in column (1), and column (2) excludes those visits where consumers complete the transaction. The same distinction is made in columns (3) and (4), but for the average time spent per book bought.

There are interesting patterns in Table 5 that indicate some consumers might enjoy shopping. We would expect consumers to spend less time visiting retailers where they had made transactions in the past if consumers' objectives are 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 (columns (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 be derived from a consumer's 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 for 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 5 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 inversely monotonic relationship between education level and search duration. Those with lower education levels spend more time searching. This could be explained by the relatively higher opportunity cost of more educated people, but also because more educated people might be more efficient in their search. 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.

The patterns of search across retailers and the time spent online suggest that search might encompass searching for a better product match in terms of observable or unobservable product characteristics. Evidence of this is provided by Bronnenberg, Kim and Mela (2016) who analyze consumer search behavior prior to consumer purchase of cameras. The advantage of their data is that they observe behavior within and across websites, which allows them to track product characteristics other than price (zoom, pixels, etc.) of each searched product through the search process. Another example is De los Santos, Hortaçsu, and Wildenbeest (2017) who assume that consumers searching website for a electronic product sample other available products available from the retailer. As such, the search process also involves different products to learn the various product offerings with different characteristics.

4 Model

The model is based on Burdett and Judd's (1983) fixed-sample equilibrium search model.[14] Of the two main search paradigms set forth in the literature following Stigler's (1961) original model, under fixed-sample search a consumer decide optimally the number of stores to search and buys from the lowest priced alternative of this set of stores. In general, fixed-sample search is preferred when there are fixed costs for conducting a search. This might be the case when online consumers face time constraints for their Internet shopping and have to stop when time runs out.

A strand of the literature, starting with McCall (1970) and Mortensen (1970), argues that a

[14]See Moraga-González and Wildenbeest (2008) for an application of Hong and Shum's (2006) model in an oligopolistic setting.

sequential search model provides a better description of actual consumer search. Intuitively, this literature points out that consumers cannot commit to a fixed-sample size search strategy when the expected marginal benefit of an extra search exceeds the marginal cost.[15] However, there is no theoretical agreement as to the search paradigm. Feinberg and Johnson (1977) show that under certain search cost magnitudes, a fixed-sample search rule is preferable to the sequential search rule. In fact, Morgan and Manning (1985) show that there is no clear advantage of a sequential search model over a fixed-sample search model. Their analysis shows that an optimal search rule combines the elements of fixed-sample search with the flexibility of sequential search. De los Santos, Hortaçsu, and Wildenbeest (2012) is one of the few papers that empirically compares these two search paradigms using individual search data. They show that the fixed-sample models better characterize online consumer search behavior than sequential search models in this industry.

A major advantage of observing search behavior prior to a transaction, in addition to price and sales data, is that Burdett and Judd's framework can be extended to allow unequal sampling of retailers. As the search patterns described above indicate, random sampling is likely rejected in this setting.

Following Burdett and Judd (1983), I assume that consumers inelastically demand one unit of a homogenous good. Under a fixed-sample 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. This is a common assumption in the theoretical literature (e.g., Stahl, 1989). Empirically it allows the use of data on consumer search behavior which leads to a transaction. If the first price quote was costly, consumers with sufficiently high search costs would not visit any retailer. Hence we would need to use search activity that does not lead to a transaction and hence the model would need to explicitly take this "outside option" into account.[16] 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

[15] Examples of sequential search models include Axell (1977); Reinganum (1979); Carlson and McAfee (1983); Rob (1985), and Stahl (1989).

[16] See Janseen and Moraga-Gonzalez (2004) for an oligopolistic model that analyzes the implications of deviations from this assumption.

distribution of the market with N firms, described by a probability mass function, be given by

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

where $\pi_j > 0$ for $j = 1, \ldots, N$ 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

$$Ep_{(1)}^{n} = E\left[\min\left\{p_1, \ldots, p_n\right\}; n\right] = \sum_{j=1}^{n} p_j f_{p_{(1)}}^{n}(p_j), \tag{2}$$

where $f_{p_{(1)}}^{n}(p)$ denotes the p.m.f. of the minimum order statistic when consumers sample n prices without replacement.[17] In Appendix A, I describe in detail the methodology to compute $f_{p_{(1)}}^{n}(p_j)$ 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

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

Given that $\Delta_j \geq 0$ for $n = 1, \ldots, N$, and the sequence of expected savings $\{\Delta_n\}_{n=1}^{N}$ is non-increasing, the optimal sample size, n^*, satisfies

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

Notice that we can reinterpret Δ_n as the largest search cost of a consumer who is indifferent

[17] Notice the implicit assumption 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 is to draw a price from an 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 $M(p)$. 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. This assumption more accurately reflects markets where consumers have a great deal of information. For example, in the case of a finite number of multi-product firms, consumers learn about the relative price distribution through repeated transactions with firms. This is particularly important in cases where a firm's relative prices for a range of products are stable over time.

between searching n^* and $n^* - 1$ firms.[18] Hence, Δ_{n^*} can be used as cutoff values that generate partitions of the search cost distribution $G(c)$. The proportion of consumers who sample $n = 1, \ldots, N$ prices is given by

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

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 (Moraga-González and Wildenbeest, 2008).

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 $n = 1, \ldots, N$ without imposing firms' equilibrium conditions. From these values, using equation (5), I recover the search cost distribution, $G(c)$, which is evaluated at cutoff points Δ_n, for $n = 1, .., N-1$.

In order to account for consumers' preferences for searching certain retailers, 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. I estimate gains Δ_i from the empirical distribution of transaction prices $f_p(p) = \pi_j$ for $p = p_j,\ j = 1, \ldots, N$, using data on consumer search histories for retailers to estimate the weights π_j. One advantage of search data and transaction prices is that we can approximate the equilibrium price distribution in the presence of firm heterogeneity. Appendix A details the empirical methodology to estimate the weight based on the sampling probabilities. It

[18] 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 $n = 1, \ldots, N$. See the work of Burdett and Judd (1983), Hong and Shum (2006) and MacMinn (1980) for a derivation of these models.

follows to compare the search cost distributions that rationalize the gains of search from asymmetric consumer search with those that result from a random sampling rule, $f_p(p) = \pi = 1/N$.

Regarding the equilibrium implications of unequal sampling. It is unlikely that firms are playing mixed strategies in the setting with unequal sampling probabilities. Retailers which are more likely to be visited will price differently than firms visited later in the search process. This is a similar to the concept of prominence retailers as in Armstrong, Vickers, and Zhou (2009) who find that the prominent firm, the one visited first by a consumer, will charge a lower price that its less prominent rivals, even when there are no systematic quality differences among the firms. This is also related to Baye, De los Santos, and Wildenbeest (2016) who show that individuals are more likely to select the more prominent retailer from organic search results at the online major search engine (e.g. Google or Bing). However, my estimation strategy does not rely on the equilibrium assumptions of the model, particularly as both prices and transactions are observed.

Another modeling approach could involve relaxing the assumption of search cost uniformity across retailers. However, given the large differences in the share of consumers visit among retailers, one would think that the main source of search cost heterogeneity across retailers is visiting the retailer itself. This retailer heterogeneity is taken into account when consumers optimally choose the number of stores to search by minimizing the net cost of search under unequal sampling probabilities. In this case, consumers are more likely to find a prominent retailer when searching which is equivalent to visiting a retailer with a low search cost. Unfortunately, I cannot separately identify heterogeneous search costs from unequal sampling as they both capture the fact that consumers are more likely to visit a particular retailer.

4.1 Consumer Search Cost Heterogeneity

In this section, I show how to use consumer search pattern data to analyze the sources of search cost heterogeneity based on consumers' observable characteristics. The fixed-sample search model is suitable to fit an ordered choice model, given that I observe the number of firms a consumer samples before making a purchase, but do not directly observe the search cost for each consumer. Define Y_i as the number of firms that consumer i samples, which can take the values of $n = 1, \ldots, N$. Consumer search costs are

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

where x_i is a vector of a consumer's observable characteristics, 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 sampled by consumers when shopping for a particular item. As shown before, a consumer with search cost c_i will optimally search n firms if $\Delta_{n+1} < c_i \leq \Delta_n$. Hence we can define the cutoffs of the search cost distribution based on estimated gains Δ_n. For instance, in the case of the empirical application $N = 4$, and we observe four possible outcomes:

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

The probabilities of each of these outcomes are related to consumer characteristics x_i by the linear separable search costs defined above:

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

We can characterize consumer choice of the number of stores to search with ordered choice models. They require the distribution H to be fully specified, e.g., in the case ε_i is normally distributed, use the standard ordered probit model. The standard or traditional ordered choice models impose several limitations on the estimates of marginal effects across all outcomes. In particular, the relative marginal effects are constant across individuals and outcomes, which leads to the parallel regression effect, i.e., the marginal effect of explanatory variables being the same across pairs of outcomes. First proposed by Terza (1985) and Maddala (1983), a generalized threshold model

relaxes this assumption by making threshold values depend on regressors.[19] For our purposes, outcome-specific regressors can be incorporated in the threshold values, e.g., $\Delta_n = \tilde{\Delta}_n + x_i\gamma_n$, where γ_n is an outcome-specific parameter, or in the search cost equations $c_{in} = x_i\beta_n + \varepsilon_i$. This specification will allow us to capture more of the different effects across outcomes. An extended specification is provided by Cunha, Heckman, and Navarro (2007) who allow the thresholds to depend on outcome-specific unobservables, e.g., $\Delta_n = \tilde{\Delta}_n + x_i\gamma_n + \upsilon_n$. The conditions for identification are an additively separable model in observables and unobservables, and the unobservables must satisfy a stochastic monotonicity assumption to preserve the ordered nature of the model.

5 Search Costs

In this section, I present estimates of the search cost distribution implied by the fixed-sample search model outlined in section 4. 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 confirming 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 6 displays the 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 the 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 the transactions, consumers visit one or two bookstores.

[19] See Boes and Winkelmann (2006) for an example of a generalized ordered choice model with regressor-dependent thresholds.

Table 7 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 uniform 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 4 that 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 $\Delta_1 = \$1.12$. The lower bound of search costs of those searching only once, as measured by averaging Δ_1 across titles is \$1.80 and \$0.89 for those who search more than once in 2004. For book titles in 2002 the average search cost cutoffs are lower: on average Δ_1 is \$1.39 and Δ_2 is \$0.35.

Second, I take into account the preference exhibited by consumers' search patterns who are more likely to visit certain retailers. As shown in Table 3 above, consumers visit Amazon at higher rates than other booksellers. Also, consumers are more likely to visit Amazon first during their search. This is taken into account with the unequal sampling probabilities of each retailer as more people are more likely to find Amazon when searching. Heterogeneity in search costs across retailers and asymmetric sampling cannot be separately identified as they both capture the fact that consumers are more likely to visit a particular retailer. Firms' unequal sampling probabilities are calculated assuming sampling without replacement with perfect recall, using the proportion of people that visit each firm as the relevant consumer sampling weight (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 7 report search cost cutoffs using unequal sampling probabilities. The cutoffs of the search cost distribution are substantially reduced in all cases. For book titles in 2004, the average Δ_1 is 0.90, which is half of the magnitude estimated under uniform sampling. The difference of the search cost cutoff Δ_2 is ninefold, the average is \$0.10 compared to \$0.89 under uniform sampling. Averaging Δ_1 of the titles in both years yields \$2.30 under uniform sampling and \$1.24 under unequal sampling. The proportional reduction is much higher for Δ_1. For example, under uniform 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 a search cost below

$0.55. It follows that the search cost distribution under uniform 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 overestimated using uniform sampling.

The results indicate that the benefits of search are much smaller once I control for the asymmetric nature of search. Since incentives for search are small, search intensity is low as well, which is 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. These no-searchers are able to buy at the first retailer as it is assumed that the first price quote is free. In order to identify search cost above this threshold we would need to characterize the purchasing decision and use observations of searches that do not lead to a transaction. Alternatively, we could relax the first price observation for free assumption, but the search cost would need to be scaled proportionally for those consumers who visit one or more firms, as consumers will optimally choose the number of firms that minimizes the expected cost of search $c(n) + Ep^n_{(1)}$ instead of $c(n-1) + Ep^n_{(1)}$.

To address some of these limitations, in the next section I fit an ordered choice model that exploits search data to recover search cost distributions from consumer characteristics, I also address the limitations of this methodology with respect to the identification of those consumers who only visit one firm.

One approach to aid with the identification of the search cost distribution function is provided by Moraga-González, Sándor, and Wildenbeest (2013). They show that combining data from various markets with similar search behavior and market characteristics (valuations, number of firms, and costs) help on the identification of the full support of the distribution as it will generate a larger and distinct sequence of cutoff points. The current setting of combining prices of several unique book titles might be a good application as the sellers, search behavior and costs are similar. However, we need sufficiently large price variation across titles to generate different cutoff points. Another interesting application of combining price data is De los Santos, Hortaçsu, and Wildenbeest (2017). In their setting consumers search for several electronic products at multi-product retailers. The large price variation across products and the identical setting help in the identification of the search cost distribution function.

6 Sources of Search Cost Heterogeneity

In this section, I explore the determinants of consumer search cost heterogeneity by using information on the number of firms visited by a consumer during their search. I relate the number of firms visited to consumer's observable characteristics using an ordered probit specification where covariates have the same effect across alternatives. In addition, I control for the unobserved heterogeneity of search costs by including random price coefficients. I extend this analysis by relaxing the assumption that the effect of the covariates is the same across alternatives and estimate a random effect generalized ordered probit model.

Table 8 displays ordered probit estimates of where the dependent variable captures the number of retailers searched by a consumer and I control for a rich set of household demographic characteristics and transaction variables. Column (1) presents ordered probit estimates assuming that the effect of covariates is the same across alternatives. The largest effect is when the consumer buys from the same bookstore as in her previous transaction. The strong negative effect of the consumer's first transaction in the panel indicates a similar effect, since the consumer's first observation is likely to be from the same unobserved bookstore. As the expected gains from search increases for higher priced books, consumers are more likely to search. There is a concern that prices might be endogenous. In particular, in equilibrium we should expect that large search frictions will increase the ability of retailers to charge higher prices, in which case we expect few firms to be visited in equilibrium. Hence, the coefficient of price might be underestimated. Ideally, I would address this with a two-stage control function approach as Petrin and Train (2010). However, I lack good instruments for price: BLP-type instruments, which would be based on book characteristics are difficult to construct given the data, and Haussman instruments cannot be constructed since we do not observe price data from other markets.

In terms of the effect of consumer characteristics in Table 8, more avid readers are more likely to search, and 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 ages 55 to 64 exhibit relatively lower search costs. An interesting case is that of individuals with the largest opportunity cost of search, those with income greater than $100,000

are less likely to search. For reference, the marginal effects of the standard ordered probit estimates are reported in Table A-1.

Figure 1 presents the implied search cost density derived from the ordered probit specification from column (1) of Table 8. The search cost density function appears to be bimodal, with an average of $2.04 and modes around $1.25 and $2.25. These results support the observed behavior of a large proportion of consumers searching only one store and are consistent with the search cutoff estimates presented in Table 7.

In order to further explore the determinants of search costs and control for unobserved heterogeneity, I estimate an ordered probit with random coefficients in two different specifications. Column (2) adds a normally distributed random price coefficient to the ordered probit estimation. The estimated variance of the random coefficient is positive and significant. Column (3) includes an interaction of price coefficient with a continuous income variable (constructed at the mid-point of each category interval). The inclusion of the interaction price coefficient with consumer characteristics to capture unobserved heterogeneity follows Petrin (2002).

In order to account for unobserved heterogeneity and dependencies of search outcomes across transactions, I relax the parallel regression assumption of the ordered probit. Table 9 presents marginal effects of the random effect generalized ordered choice model that allows the effect of regressors to vary across outcomes by specifying thresholds that depend on covariates.[20] Each column of the table represents the ordered outcomes of a consumer's search from one to four bookstores. Notice that the table presents the marginal effects of each ordered outcome ($Y_i = 1, \ldots, 4$) and as such the effects are distributed and add up to zero across all outcomes. Qualitatively and in order of magnitude, the effects of searching one store are similar to the decision of whether to search one or more stores, which is presented in the probit model. This table provides the breakdown of the effect across outcomes. For example, consumers are more likely to search more than once for higher prices. In this case, the price coefficient of one search (search = 1) is -0.009, which has its counterpart in consumers searching twice, where the price coefficient is 0.006; and to

[20] The model is approximated numerically using a Gauss-Hermite quadrature approximation. See Boes (2006) for a detailed explanation of the econometric specification. Other alternatives to a probit specification are provided by Klein and Spady (1993), using semiparametric approximation to the distribution of the binary response models, extended by Klein and Sherman (2002) 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).

a lesser extent, in consumers searching three times, where the price coefficient is 0.003. As most people searched one or two stores in the sample, the marginal effects are largest on searching one or two stores.

7 Conclusions

This paper extends the fixed-sample framework of Burdett and Judd (1983) to allow for unequal sampling of retailers and proposes an ordered choice model that relates the number of firms sampled to consumer characteristics. It uses individual search activity from consumer online browsing and transactions to structurally estimate this model. Search patterns in the online book industry indicate limited consumer search and a strong preference for particular retailers. I find that the standard assumption of uniform sampling of firms can lead to biased estimates of search costs. Search model estimates which incorporate unequal sampling by a consumer lead to substantially lower search cost estimates. This evidence suggests that search frictions in online markets are lower than other studies suggested.

In addition, I show that an ordered choice model as applied to individual search behavior is a parsimonious way to estimate a fixed-sample model and to analyze the sources of consumer search cost heterogeneity. Estimates of search cost on consumer characteristics show a strong substitution between time and expenditures in online markets.

References

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

Armstrong, M., J. Vickers, and J. Zhou (2009): "Prominence and Consumer Search," *The RAND Journal of Economics*, 40(2), 209–233.

Axell, B. (1977): "Search Market Equilibrium," *The Scandinavian Journal of Economics*, 79(1), 20–40.

Baye, M. R., B. De los Santos, and M. R. Wildenbeest (2016): "Search Engine Optimization: What Drives Organic Traffic to Retail Sites?," *Journal of Economics & Management Strategy*, 25(1), 6–31.

Baye, M. R., and J. Morgan (2009): "Brand and Price Advertising in Online Markets," *Management Science*, 55(7), 1139–1151.

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.

Boes, S. (2006): "Three Essays on the Econometric Analysis of Discrete Dependent Variables," Ph.D. thesis, University of Zurich.

Boes, S., and R. Winkelmann (2006): "Ordered response models," in *Modern Econometric Analysis*, pp. 167–181. Springer.

Bronnenberg, B. J., J. B. Kim, and C. F. Mela (2016): "Zooming in on choice: How do consumers search for cameras online?," *Marketing Science*, 35(5), 693–712.

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.

Carlson, J. A., and R. P. McAfee (1983): "Discrete Equilibrium Price Dispersion," *Journal of Political Economy*, 91(3), 480–93.

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.

Cunha, F., J. J. Heckman, and S. Navarro (2007): "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds," *International Economic Review*, 48(4), 1273–1309.

De los Santos, B., A. Hortaçsu, and M. R. Wildenbeest (2012): "Testing Models of Consumer Search using Data on Web Browsing and Purchasing Behavior," *American Economic Review*, 102(2), 2955–2980.

——— (2017): "Search with Learning for Differentiated Products: Evidence from E-Commerce," *Journal of Business & Economic Statistics*, 35(4), 626–641.

De los Santos, B., and M. R. Wildenbeest (2017): "E-book Pricing and Vertical Restraints," *Quantitative Marketing and Economics*, 15(2), 85–122.

Ellison, G., and S. F. Ellison (2009): "Search, Obfuscation, and Price Elasticities on the Internet," *Econometrica*, 77, 427–452.

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.

Feinberg, R. M., and W. R. Johnson (1977): "The Superiority of Sequential Search: A Calculation," *Southern Economic Journal*, 43(4), 1594–1598.

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.

Honka, E. (2014): "Quantifying Search and Switching Costs in the U.S. Auto Insurance Industry," *RAND Journal of Economics*, 45(4), 847–884.

Hortaçsu, A., and C. Syverson (2004): "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 Index Funds," *Quarterly Journal of Economics*, 119, 403–56.

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.

Kim, J. B., P. Albuquerque, and B. J. Bronnenberg (2010): "Online demand under limited consumer search," *Marketing science*, 29(6), 1001–1023.

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.

Koulayev, S. (2014): "Search for Differentiated Products: Identification and Estimation," *RAND Journal of Economics*, 45(3), 553–575.

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

Maddala, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs.

McCall, J. J. (1970): "Economics of Information and Job Search," *The Quarterly Journal of Economics*, 84(1), 113–26.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2013): "Semi-Nonparametric Estimation Of Consumer Search Costs," *Journal of Applied Econometrics*, 28(7), 1205–1223.

Moraga-González, J. L., Z. Sándor, and M. R. Wildenbeest (2015): "Consumer search and prices in the automobile market," working paper.

Moraga-González, J. L., and M. R. Wildenbeest (2008): "Maximum Likelihood Estimation of Search Costs," *European Economic Review*, 52(5), 820–48.

Morgan, P., and R. Manning (1985): "Optimal Search," *Econometrica*, 53(4), 923–44.

Mortensen, D. T. (1970): "Job Search, the Duration of Unemployment, and the Phillips Curve," *The American Economic Review*, 60(5), 847–862.

Nishida, M., and M. Remer (2018): "The Determinants and Consequences of Search Cost Heterogeneity: Evidence from Local Gasoline Markets," *Journal of Marketing Research*, forthcoming.

Noam, E. (2009): *Media Ownership and Concentration in America*. Oxford University Press.

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," *Ssrn Electronic Journal*.

Pan, X., B. T. Ratchford, and V. Shankar (2004): "Price Dispersion on the Internet: A Review and Directions for Future Research," *Journal of Interactive Marketing*, 18, 116–135.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110(4), 705–729.

Petrin, A., and K. Train (2010): "A Control Function Approach to Endogeneity in Consumer Choice Models," *Journal of Marketing Research*, 47(1), 3–13.

Pires, T. (2015): "Costly Search and Consideration Sets in Storable Goods Markets," working paper.

Prince, J. (2008): "Repeat Purchases amid Rapid Quality Improvement: Structural Estimation of the Demand for Personal Computers," *Journal of Economics and Management Strategy*, 17, 1–33.

Reinganum, J. F. (1979): "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy*, 87(4), 851.

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," *Advances in Applied Microeconomics*, 11, 63–88.

Seiler, S. (2013): "The impact of search costs on consumer behavior: A dynamic approach," *Quantitative Marketing and Economics*, 11(2), 155–203.

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.

Stigler, G. J. (1961): "The economics of information," *Journal of Political Economy*, 69, 213–25.

Wildenbeest, M. R. (2011): "An empirical model of search with vertically differentiated products," *The RAND Journal of Economics*, 42(4), 729–757.

Zhang, X., T. Chan, and Y. Xie (2016): "Price Search and Periodic Price Discounts," working paper.

Zwick, R., A. Rapoport, C. Lo, A. King, and A. V. Muthukrishnan (2003): "Consumer Sequential Search: Not Enough or Too Much?," *Marketing Science*, 22, 503–519.

Figure 1: Search Cost Distribution by Number of Firms Searched

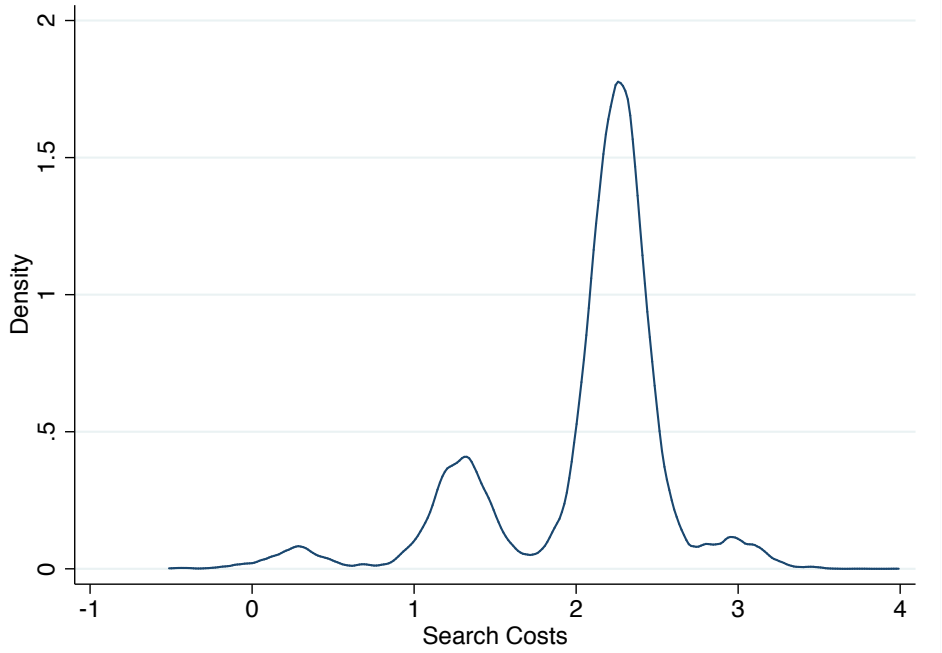

Table 1: Consumer Browsing and Transaction Data Descriptive Statistics

	2002			2004		
	Mean	Std. Dev.	Bootstrap Std. Err.	Mean	Std. Dev.	Bootstrap 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, including 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 behavior of online bookstores calculated from ComScore data for the period July to 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

	All bookstores	Book purchased from: 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

Search window	All bookstores	Book purchased from: Amazon	Barnes & Noble	Book clubs	Other bookstores
		Only 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 lengths 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: Regression of Search Duration on Household Characteristics

	(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	7.932	(1.688)***	11.147	(2.704)***	6.508	(1.340)***	9.083	(2.130)***
Free shipping (sales $\geq$ \$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	
One book title purchased	Yes		Yes		Avg. duration/book		Avg. duration/book	
R-squared	0.13		0.06		0.08		0.05	
Number of observations	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. "Same bookstore" indicates the transaction was completed at the same bookstore as the previous transaction. "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 6: Descriptive Statistics

		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
00 The Lovely Bones	14	14.0	4.8	10.0	21.0	0.643	0.214	0.143
Harry Potter (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 $n = 1, 2, \ldots, N$ bookstores.

Table 7: Empirical Search Cost CDF

	Uniform 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
Average 2004 ($)	$1.80	0.89	0.90	0.10		
Best sellers 2002						
From a Buick 8	3.261	0.753	1.242	0.083	0.270	0.027
Four Blind Mice	3.284	0.650	1.224	0.034	0.200	0.029
Nights in Rodanthe	0.884	0.441	0.779	0.194	0.176	0.118
The Lovely Bones	5.914	0.679	4.557	1.016	0.357	0.143
Harry Potter (Goblet of Fire)	2.858	1.432	1.595	1.091	0.214	0.071
Visions of Sugar Plums	0.538	0.270	0.111	0.020	0.143	0.143
Average 2002 ($)	1.39	0.35	0.79	0.2		
Overall Average ($)	2.30	0.80	1.24	0.25		

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 $G(\Delta 1) = 1 - q1$, $G(\Delta_2) = 1 - q1 - q2$, $G(\Delta_3) = 0$. 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 8: Ordered Probit Estimates of the Number of Visited Firms on Consumer Characteristics

	(1)		(2)		(3)	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Price ($10)	0.025	(0.006)***	0.023	(0.007)***	0.044	(0.013)***
Var(Price)			0.030	(0.008)***		
Price ($1000) × Income					-0.023	(0.018)
Var(Price ($1000) × Income)					0.022	(0.009)**
First transaction indicator	-0.613	(0.040)***	-0.606	(0.037)***	-0.615	(0.037)***
Same bookstore	-0.702	(0.042)***	-0.701	(0.039)***	-0.707	(0.038)***
Free shipping (sales ≥ $25)	-0.026	(0.029)	-0.015	(0.028)	-0.025	(0.028)
One book title purchased	-0.043	(0.027)	-0.037	(0.026)	-0.042	(0.026)
Cumulative book transactions	0.009	(0.002)***	0.009	(0.001)***	0.009	(0.001)***
Household size	0.064	(0.017)***	0.063	(0.011)***	0.064	(0.011)***
Broadband connection	0.106	(0.028)***	0.105	(0.022)***	0.104	(0.022)***
Children present in household	-0.028	(0.044)	-0.029	(0.030)	-0.031	(0.030)
Age						
18-20	0.010	(0.102)	0.002	(0.096)	0.015	(0.095)
21-24	0.007	(0.072)	0.005	(0.067)	0.009	(0.066)
25-29	0.005	(0.061)	0.004	(0.055)	0.009	(0.055)
30-34	0.116	(0.056)**	0.109	(0.046)**	0.115	(0.046)**
35-39	0.113	(0.058)**	0.116	(0.046)**	0.116	(0.046)**
40-44	0.089	(0.052)*	0.087	(0.042)**	0.092	(0.042)**
50-54	0.042	(0.048)	0.042	(0.039)	0.043	(0.039)
55-59	0.137	(0.057)**	0.141	(0.044)***	0.139	(0.044)***
60-64	0.130	(0.076)*	0.132	(0.048)***	0.134	(0.048)***
65 and over	-0.022	(0.057)	-0.024	(0.046)	-0.020	(0.046)
Household income						
Less than $15,000	0.007	(0.063)	0.002	(0.057)	-0.029	(0.061)
$15,000 - $24,999	0.087	(0.052)*	0.089	(0.045)**	0.065	(0.049)
$25,000 - $34,999	0.026	(0.048)	0.022	(0.036)	-0.001	(0.040)
$35,000 - $49,999	0.054	(0.043)	0.053	(0.032)	0.044	(0.035)
$75,000 - $99,999	0.031	(0.047)	0.031	(0.037)	0.045	(0.039)
More than $100,000	-0.104	(0.047)**	-0.100	(0.037)***	-0.077	(0.043)*
Education						
Less than high school	0.087	(0.144)	0.063	(0.130)	0.076	(0.130)
High school diploma or GED	0.057	(0.063)	0.049	(0.043)	0.053	(0.043)
Some college but no degree	0.059	(0.042)	0.052	(0.035)	0.056	(0.035)
Associate degree	-0.007	(0.064)	-0.014	(0.048)	-0.009	(0.048)
Graduate degree	-0.021	(0.051)	-0.021	(0.040)	-0.024	(0.040)
Race						
Black	0.090	(0.063)	0.085	(0.055)	0.087	(0.055)
Hispanic	-0.053	(0.050)	-0.056	(0.040)	-0.053	(0.040)
Asian	0.161	(0.084)*	0.164	(0.064)**	0.162	(0.064)**
Other	0.068	(0.089)	0.059	(0.072)	0.068	(0.072)
Region of residence						
Northeast	0.042	(0.037)	0.044	(0.030)	0.041	(0.030)
Midwest	0.019	(0.040)	0.021	(0.030)	0.020	(0.030)
West	-0.128	(0.038)***	-0.122	(0.030)***	-0.129	(0.030)***
Cuttoff Δ_1	0.487	(0.088)***	0.478	(0.067)***	0.485	(0.067)***
Cuttoff Δ_2	1.552	(0.087)***	1.558	(0.069)***	1.562	(0.068)***
Cuttoff Δ_3	2.566	(0.095)***	2.583	(0.081)***	2.587	(0.081)***
Random Coefficients	No		Yes		Yes	
Log-likelihood	-10,232.2		-10,212.931		-10,227.977	

Notes: The table presents the coefficients of an ordered probit model ($N = 15,561$) using as a dependent variable the number of bookstores visited by a consumer and assuming that the effect of each covariate is the same across alternatives (also referred as the parallel regressions assumptions). For each transaction, consumers searched either 1,2, 3 or 4 bookstores. 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. "Same bookstore" indicates the transaction was completed at the same bookstore as the previous transaction. "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: Average Marginal Effect s of the Number of Visited Firms on Consumer Characteristics (Random Effects Generalized Ordered Probit)

	Search = 1		Search = 2		Search = 3		Search = 4	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Price ($10)	-0.009	(0.002)***	0.006	(0.002)***	0.003	(0.001)***	0.000	(0.000)
First transaction indicator	0.131	(0.012)***	-0.102	(0.011)***	-0.025	(0.004)***	-0.004	(0.001)***
Same bookstore	0.154	(0.012)***	-0.117	(0.011)***	-0.035	(0.005)***	-0.002	(0.001)
Free shipping (sales $\geq$ $25)	0.011	(0.008)	-0.008	(0.008)	-0.003	(0.004)	-0.001	(0.001)
One book title purchased	0.014	(0.008)*	-0.015	(0.007)*	-0.001	(0.003)	0.001	(0.001)
Cumulative book transactions	-0.003	(0.000)***	0.002	(0.000)***	0.001	(0.000)***	0.000	(0.000)***
Household size	-0.015	(0.004)***	0.008	(0.003)**	0.006	(0.002)***	0.001	(0.000)
Broadband connection	-0.021	(0.008)***	0.011	(0.007)	0.010	(0.003)***	-0.000	(0.001)
Children present in household	0.000	(0.010)	-0.000	(0.009)	-0.001	(0.004)	0.001	(0.001)
Age								
18-20	-0.011	(0.031)	0.026	(0.030)	-0.027	(0.014)*	0.012	(0.004)***
21-24	-0.010	(0.022)	0.015	(0.020)	0.027	(3.388)	-0.032	(3.388)
25-29	0.001	(0.019)	0.011	(0.017)	0.020	(2.063)	-0.031	(2.063)
30-34	-0.030	(0.016)*	0.034	(0.014)**	-0.003	(0.007)	-0.001	(0.002)
35-39	-0.026	(0.016)*	0.025	(0.014)*	-0.000	(0.006)	0.001	(0.002)
40-44	-0.024	(0.015)*	0.027	(0.013)**	-0.004	(0.006)	0.002	(0.002)
50-54	-0.010	(0.014)	0.020	(0.012)*	-0.009	(0.006)	-0.001	(0.002)
55-59	-0.038	(0.015)**	0.036	(0.014)***	0.000	(0.006)	0.002	(0.002)
60-64	-0.012	(0.017)	0.005	(0.015)	0.005	(0.007)	0.002	(0.002)
65 and over	-0.008	(0.016)	0.016	(0.014)	-0.006	(0.007)	-0.002	(0.002)
Household income								
Less than $15,000	-0.001	(0.020)	-0.004	(0.018)	0.002	(0.008)	0.002	(0.003)
$15,000 - $24,999	-0.020	(0.015)	0.014	(0.014)	0.005	(0.006)	0.001	(0.002)
$25,000 - $34,999	-0.003	(0.013)	0.001	(0.011)	0.004	(0.005)	-0.002	(0.002)
$35,000 - $49,999	-0.009	(0.011)	0.003	(0.010)	0.008	(0.005)	-0.001	(0.001)
$75,000 - $99,999	-0.002	(0.013)	-0.000	(0.011)	0.005	(0.005)	-0.002	(0.002)
More than $100,000	0.028	(0.013)**	-0.020	(0.012)*	-0.004	(0.006)	-0.005	(0.002)*
Education								
Less than high school	-0.035	(0.043)	0.084	(0.048)*	-0.020	(7.209)	-0.029	(7.209)
High school diploma or GED	-0.006	(0.015)	0.006	(0.014)	-0.004	(0.006)	0.003	(0.002)*
Some college but no degree	-0.021	(0.012)*	0.014	(0.011)	0.005	(0.005)	0.003	(0.002)*
Associate degree	0.009	(0.017)	-0.015	(0.015)	0.006	(0.007)	-0.001	(0.002)
Graduate degree	0.017	(0.014)	-0.017	(0.013)	-0.003	(0.006)	0.003	(0.002)
Race								
Black	-0.028	(0.019)	0.021	(0.017)	0.006	(0.008)	0.002	(0.002)
Hispanic	-0.003	(0.014)	0.002	(0.012)	0.001	(0.006)	0.000	(0.002)
Asian	-0.042	(0.022)*	0.025	(0.020)	0.046	(2.883)	-0.029	(2.883)
Other	-0.006	(0.026)	0.009	(0.023)	0.003	(0.011)	-0.006	(0.004)
Region of residence								
Northeast	-0.014	(0.011)	0.013	(0.009)	-0.001	(0.004)	0.002	(0.001)
Midwest	-0.000	(0.010)	0.003	(0.009)	-0.004	(0.004)	0.001	(0.001)
West	0.036	(0.010)***	-0.025	(0.009)***	-0.011	(0.005)**	-0.000	(0.002)

Notes: The table presents the average marginal effects of a random effects generalized ordered probit model ($N = 15,561$, *log-likelihood* $= -9876.3$) using as a dependent variable the number of bookstores visited by a consumer. For each transaction, consumers searched either 1,2, 3 or 4 bookstores. 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. "Same bookstore" indicates the transaction was completed at the same bookstore as the previous transaction. "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%.

A Estimation of Minimum Prices under Different Sampling Assumptions

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

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

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 size of n is given by

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

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 of uniform probability sampling without replacement, $f_p(p) = \pi = 1/N$ for $p = p_1, \ldots, p_N$. 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

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

where $x \in [1, 2, \ldots, N + 1 - n]$ denotes the reordered support of $p \in [p_1, p_2, \ldots, p_{N+1-n}]$ (see Evans, Leemis, and Drew, 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 price reduces to

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

Next, if all the stores are sampled, $n = N$, the minimum price is observed as $f_{p_1}^{N}(p) = 1$. Finally,

$2 < n < N-2$ is a non-trivial case, with no closed form. It is calculated from combinatorial procedures.

For simplicity, Ω^n is defined as the set containing the combination of n prices sampled from $p_1, \ldots, p_n$. for the case of $N = 4$ and $n = 2$:

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

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, which is 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

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

and equivalent probabilities for $p_2, \ldots, p_n$. 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:

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

The objective is to estimate the probabilities λ_j^n from consumer search data. The probability

that consumer i samples p_j when optimally sampling n prices is defined 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 that $\sum_{j=1}^N \hat{\lambda}_j^n = 1$. For consumers who sample one store, $n = 1$, we know from the data that $\hat{\lambda}_j^1 = \hat{\pi}_j$ 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 $n = 2$:

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

B Data Sample Construction

This appendix describes in detail the construction of the book dataset from the ComScore data. I restrict the sample to book transactions (i.e., exclude periodicals, 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 attempt 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 correct 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 quantity that equals 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.

I only include transactions from online bookstores. I exclude transactions from Ebay.com which represent 15 percent of the total number of book transactions . An additional 3 percent of book transactions are also excluded were from websites that could not be identified as online bookstores, such as auction sites and domains that could not be identified as a bookstore. Search behavior and the products are between these sites and online bookstores are fundamentally different as products in auctions sites are sold by third-party seller and list used books, autographed volumes, or auctioned items. A fringe number of observations from international Amazon websites in the United Kingdom, Canada, and Denmark were also excluded.

Search activity from Borders.com, a major brick-and-mortar bookstore which is now closed, is excluded. The reason is that 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 were redirected to Amazon.com and already observed in the data.

B.1 Current Population Survey

I use the 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,

who are 18 years of age or older, and who claim 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

For this paper, I use the Forrester Consumer Technographics Survey 2003, which is conducted from December 2002 to February 2003.[21] 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, which is defined as "establishments primarily engaged in retailing new books" (NAICS code 451211). I calculated 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.

[21] See Brown and Goolsbee (2002) for a detailed description of the dataset and Prince (2008) for an estimation of the demand of personal computers using the Forrester survey.

Table A-1: Average Marginal Effect s of the Number of Visited Firms on Consumer Characteristics (Ordered Probit Model)

	Search = 1		Search = 2		Search = 3		Search = 4	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Price ($10)	-0.008	(0.002)***	0.006	(0.001)***	0.002	(0.000)***	0.000	(0.000)***
First transaction indicator	0.184	(0.012)***	-0.131	(0.008)***	-0.047	(0.003)***	-0.006	(0.001)***
Same bookstore	0.211	(0.012)***	-0.150	(0.009)***	-0.053	(0.004)***	-0.007	(0.001)***
Free shipping (sales $\geq$ $25)	0.009	(0.009)	-0.006	(0.006)	-0.002	(0.002)	-0.000	(0.000)
One book title purchased	0.013	(0.008)*	-0.010	(0.006)	-0.003	(0.002)	-0.000	(0.000)
Cumulative book transactions	-0.003	(0.001)***	0.002	(0.000)***	0.001	(0.000)***	0.000	(0.000)***
Household size	-0.019	(0.005)***	0.014	(0.004)***	0.005	(0.001)***	0.001	(0.000)**
Broadband connection	-0.032	(0.009)***	0.023	(0.006)***	0.008	(0.002)***	0.001	(0.000)***
Children present in household	0.009	(0.013)	-0.006	(0.009)	-0.002	(0.003)	-0.000	(0.000)
Age								
18-20	-0.003	(0.031)	0.002	(0.022)	0.001	(0.008)	0.000	(0.001)
21-24	-0.003	(0.022)	0.002	(0.015)	0.001	(0.005)	0.000	(0.001)
25-29	-0.002	(0.018)	0.001	(0.013)	0.000	(0.005)	0.000	(0.001)
30-34	-0.035	(0.017)**	0.025	(0.012)**	0.009	(0.004)**	0.001	(0.001)*
35-39	-0.034	(0.017)**	0.025	(0.012)**	0.009	(0.004)**	0.001	(0.001)*
40-44	-0.027	(0.016)*	0.019	(0.011)*	0.007	(0.004)*	0.001	(0.001)
50-54	-0.013	(0.015)	0.009	(0.010)	0.003	(0.004)	0.000	(0.001)
55-59	-0.041	(0.017)**	0.030	(0.012)**	0.010	(0.004)**	0.001	(0.001)**
60-64	-0.040	(0.023)*	0.028	(0.016)*	0.010	(0.006)*	0.001	(0.001)
65 and over	0.007	(0.017)	-0.005	(0.012)	-0.002	(0.004)	-0.000	(0.001)
Household income								
Less than $15,000	-0.002	(0.019)	0.001	(0.013)	0.000	(0.005)	0.000	(0.001)
$15,000 - $24,999	-0.026	(0.016)*	0.019	(0.011)*	0.007	(0.004)*	0.001	(0.001)
$25,000 - $34,999	-0.008	(0.014)	0.006	(0.010)	0.002	(0.004)	0.000	(0.000)
$35,000 - $49,999	-0.016	(0.013)	0.011	(0.009)	0.004	(0.003)	0.001	(0.000)
$75,000 - $99,999	-0.009	(0.014)	0.007	(0.010)	0.002	(0.004)	0.000	(0.000)
More than $100,000	0.031	(0.014)**	-0.022	(0.010)**	-0.008	(0.004)**	-0.001	(0.001)**
Education								
Less than high school	-0.027	(0.043)	0.019	(0.031)	0.007	(0.011)	0.001	(0.002)
High school diploma or GED	-0.017	(0.019)	0.012	(0.013)	0.004	(0.005)	0.001	(0.001)
Some college but no degree	-0.018	(0.013)	0.013	(0.009)	0.004	(0.003)	0.001	(0.000)
Associate degree	0.002	(0.019)	-0.001	(0.014)	-0.000	(0.005)	-0.000	(0.001)
Graduate degree	0.006	(0.015)	-0.004	(0.011)	-0.002	(0.004)	-0.000	(0.001)
Race								
Black	-0.027	(0.019)	0.019	(0.014)	0.007	(0.005)	0.001	(0.001)
Hispanic	0.016	(0.015)	-0.011	(0.011)	-0.004	(0.004)	-0.001	(0.001)
Asian	-0.048	(0.025)*	0.034	(0.018)*	0.012	(0.006)*	0.002	(0.001)*
Other	-0.020	(0.027)	0.014	(0.019)	0.005	(0.007)	0.001	(0.001)
Region of residence								
Northeast	-0.013	(0.011)	0.009	(0.008)	0.003	(0.003)	0.000	(0.000)
Midwest	-0.006	(0.012)	0.004	(0.009)	0.001	(0.003)	0.000	(0.000)
West	0.038	(0.011)***	-0.027	(0.008)***	-0.010	(0.003)***	-0.001	(0.000)***

Notes: The table presents the average marginal effects of an ordered probit model ($N = 15,561$, *log likelihood*$= -10232.2$) using as a dependent variable the number of bookstores visited by a consumer. For each transaction, consumers searched either 1,2, 3 or 4 bookstores. 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. "Same bookstore" indicates the transaction was completed at the same bookstore as the previous transaction. "Cumulative book transactions" are the number of book purchases prior to the current one. 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%.