The Pennsylvania State University
The Graduate School
Department of Economics

ESSAYS IN ONLINE RETAIL MARKETS AND AUCTION MARKETS

A Dissertation in
Economics
by
Jicheng Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2015

The dissertation of Jicheng Liu was reviewed and approved* by the following:

Mark J.Roberts
Professor of Economics
Dissertation Advisor,
Chair of Committee

Peter Newberry
Assistant Professor of Economics

Joris Pinkse
Professor of Economics

Hari Sridhar
Associate Professor of Marketing

Barry Ickes
Professor of Economics
Head, Department of Economics

*Signatures are on file in the Graduate School.

Abstract

Chapter 1: "**Does Information Change Bidders' Value Distribution in eBay Motor Auctions?**"

In eBay Motor auctions, sellers voluntarily disclose information of the car to increase the selling prices (Lewis(2010)). An increase in the number of potential bidders or a change of potential bidders' value distribution can explain this phenomenon. This paper first tests whether the information disclosure influences bidders' value distribution by proposing a likelihood ratio test in an eBay auction model with an application to eBay Motor auctions. The test result shows that the bidders' value distribution is not affected by different information disclosure levels measured by the number of photos provided by sellers in the application to eBay Motor auctions. By exploiting the bidders' value distribution estimates, I also calculate the consumer surplus, the mean of consumer surplus is 48% of the selling price which shows how much the online auction market can benefit consumers.

Chapter 2: "**Consumer Search in the Online Movie DVD Market**"
In this paper, I study the consumers' online search behavior that has been documented thoroughly in the industrial organization literature. Using unique features of Comscore consumer search behavior data and transaction data, I focus on the online movie DVD retail industry. I find several features of consumer search behavior in this specific market. Particularly, consumer makes her purchase decision on both price and other retailer's specific features or its quality.

I estimate search costs in a nonsequential search model describing the web browsing and purchasing behavior. Consumer knows the price distribution and she can ascertain the exact price of each retailer by searching on the firm website. Following the canonical search model, consumer has two decision steps in each period. First, each consumer makes a choice set decision on how many retailers' websites to visit. After gaining information on the exact price of each retailer in the choice set, consumer maximize

her utility over the retailers in her choice set. The estimation result shows the average search costs are around $2.73 for an individual to search a firm's website before each transaction.

Chapter 3: "**A Structural Model of Search with Learning: Application to the Online Movie DVD Retail Market**"
Much of the demand literature uses search costs to explain consumer behavior and firms' pricing strategy in the online retail industry. However, some dynamic features of consumer search behavior are not captured. For example, novel web behavioral data indicates that the consumers' search sets or "consideration sets" change over time. Such evidence of behavior change of consumers is consistent with a consumer learning model. Consumers are usually not certain about the firm's quality, for example the available service, before their visit to the firm's website. Therefore, the consumer's search decision is a trade-off between higher expected maximum utility from a larger search set and higher search cost. The decision depends on consumer's information about the quality of the firms.

In this paper I develop a dynamic discrete choice model where consumers choose their search sets and then make the purchase decision from the firms in their search sets in each period. Each consumer's search decision is a dynamic process which depends on her endogenous belief evolution on quality.

I estimate the dynamic model using the ComScore consumer level search and purchase behavior data for the movie DVD retail market. The estimation results suggest that search cost is sizable and there exists consumer heterogeneity in firms' quality belief. Consumers are learning about the firms quality very slowly which plays an important role in explaining the frequent search behavior change of consumers.

Counterfactual experiments suggest that search costs and consumer learning have a huge impact on the market shares of firms. The industry leader Amazon would have 12% less share absent search cost. In addition, I compare the effects of different pricing strategies of firms. The market share of Amazon shrinks by 2% if Amazon matches the price of ColumbiaHouse which has the second largest market share. The market share of the leading firm Amazon increases from 35.16% to 68.07% if consumer has one-search learning process.

Table of Contents

Chapter 3

List of Tables

Acknowledgments

I would like to thank my advisor, Mark J.Roberts, for all his help, insightful suggestions and Patience. His advice helped me to grow in many ways more than just research. I am also deeply indebted to Peter Newberry for his intellectual generosity, helpful suggestions and guidance along this way. I am also grateful to Joris Pinkse, Hari Sridhar for their helpful discussion and comments.

Dedication

I would like to dedicate this dissertation to my parents, Dongmei and Maoliang, and my wife Xue.

Chapter 1 | Does Information Change Bidders' Value Distribution in eBay Motor Auctions?

1.1 Introduction

Over the last decade, a huge emergence of economic and social activities on the Internet has been seen. The on-line auction site eBay is one of the most successful Internet platforms which generated 11.652 billion dollars in year 2011.[1] One reason for the success of eBay is that assembling sellers and buyers through the Internet requires much lower cost than assembling them in a traditional auction. In a general eBay auction of used vehicle, unlike in a traditional auction, buyers are not able to see the vehicle in person but the information provided by the sellers. This information asymmetry between sellers and buyers may lead to adverse selection. The pioneer work by Akerlof (1970) shows that adverse selection can deteriorate the used good market and finally eliminate it. At first glance, it seems to be a paradox between the quick growth of the online used car auction market and the information asymmetry between sellers and buyers. However, the information gap between sellers and buyers is not as large as one might think of since car sellers usually provide standard information of their selling cars such as maker, age, mileage, transmission type, VIN number and so on. Additional to the standard information, sellers also provide pictures and text descriptions to indicate the “quality" of their cars.[2]

Lewis (2010) argues that in the eBay auction environment, it is possible to partially contract on the quality of goods by sellers' voluntarily information disclosure. Sellers offer text and photos on the auction webpage which specifies a contract to potential buyers to

[1] Source: eBay Financial Releases.

[2] EBay charges sellers a fixed insertion fee $40 to post the standardized description of the car for selling and another $0.15 per photo and nothing on text and graphics in the data period. Comparing to the selling price, this monetary cost of providing information is negligible. EBay have the policy that listing is free but sellers would be charged $125 for a successful trade.

deliver the selling car described on the webpage. He empirically tests the hypothesis that the information disclosure by sellers is important in determining the selling prices (the second highest bid in the auction), by showing that photos and text posted by the seller significantly influence selling prices in a hedonic regression model.

There are two possible channels through which the information disclosure influences on the selling price of a good. An increase of the number of potential bidders and/or a change, for example, a shift outward of potential bidder's value distribution can both increase the selling prices, given the characteristics of the good. With the same value distribution, more potential bidders lead to more draws of value from the same distribution, thus the second highest value which is the selling price has a higher probability to be large. With the same number of potential bidders, a shift of value distribution also leads to a shift of each draw of bidder's value, thus the second highest value can also increase. In this paper, I examine empirically the effect of the voluntary information disclosure on potential bidder's value distribution. I propose a private value auction model with different type of bidders. Different type of bidders have different value distribution which is decided by the information disclosure level. Then I test whether the bidders' value distributions under different voluntarily information disclosure level are the same or not in a structural eBay auction model built on Song (2004).

Within the standard symmetric independent private value (IPV) framework, potential bidder's distribution is non-parametrically identified with any order statistic of the bids (for example, second highest bids) in a second price auction with known number of bidders (Athey and Haile (2002)). Unlike the standard second price auction, the number of potential bidders is actually unknown in an eBay auction. Song (2004) first notices that potential bidder's distribution is also identified from the joint distribution of any two order statistic of the bids (for example, second highest- and third highest bids) in a second price auction with an exogenous unknown number of potential bidders. The identification result holds even in the case that the number of potential bidders varies across auctions. The semi-nonparametric (SNP) maximum likelihood estimation method can naturally be applied to consistently estimate the private value density function from the data of second and third highest valuations of eBay auctions.

My study focuses on the eBay Motor auction dataset which is the same one used by Lewis (2010). This is a large dataset of 82,358 vehicle auctions and has many observed characteristics of selling vehicles such as mileage, make, model and so on. This paper extends Song(2004) to allow auction objects with observed heterogeneity in order to analyze this dataset properly. I use the classic hedonic model to specify the utility of a good, specifically the utility is a function of its characteristic and an additive or multiplicative value from a common distribution across goods. Haile, Hong and Shum (2003) use this technique in a first-price auction model framework.

This paper uses a semi-nonparametric estimator by employing the Laguerre polynomial series rather than the adjusted Hermite polynomial series in Song(2004). It only requires single integral rather than the double integral in Song's case which makes the implement of the estimator easier. The semi-nonparametric approach has been introduced to economics studies in the 1980s by Gallant (1987) and his coauthors. Instead of maximizing a likelihood function over an infinite-dimensional space (the space of all possible density function of bidder's value in auction models), one can maximize it over a sequence of less complex, usually parameter spaces (seives). The choice of seives is not unique. Chen (2007) gives some commonly used sieves with good approximation properties for different spaces in her econometrics handbook chapter. The bidder's value is assumed to be a positive number which is measured by the money willing to pay. It leads to my choice of Laguerre polynomial series since any function in $L_2[0, \infty)$ can be approximated by Laguerre polynomials.

To test whether two bidders' value distributions are same or not, I propose a likelihood ratio test since it has the highest power among all competitors[3]. Since a likelihood ratio test is a parametric test, I restrict the testing model to be a parametric model by assuming that the value distribution functions are in a parametric family. A natural way is to fix the length of the chosen seives. I then test the null hypothesis that the two value distributions are the same against the alternative hypothesis that the two value distributions are different. The test result shows that the bidders' value distribution is not affected by different information disclosure levels measured by the number of photos provided by sellers once the effect of observed characteristics has been controlled.

I also estimate the consumer surplus generated by the eBay Motors after the estimation of the structural model. In the baseline estimation, I control only the book values of the selling cars. The result shows that the median Consumer Surplus is 45.5% of the median of book value. In a general specification, I control more observed characteristics of selling vehicles such as age, mileage, maker as well as the sellers' reputation feedbacks. The mean (Median) estimated consumer surplus share is 48.2%(32.0%) and the mean(median) consumer surplus is $4,070\$(2,136\$)$. These findings are consistent with those of Song (2004) and (Giray et al.(2009)) that consumer gains a significant amount of surplus in eBay auctions.

The paper is organized as follows. In section 2, I describe the eBay Motor auction

[3] Neyman Pearson lemma.

background. Section 3 introduces the Econometric model. Section 4 shows the estimator and test procedure. Section 5 describes the data and the empirical results. Section 6 concludes.

1.2 Background of eBay Motors Auction

According to the news from the Senior Director of eBay Motors in the year 2007, eBay Motors became the largest automotive site on the Internet and approximately every minute a car or a truck was sold. Its trade volume exceeds other online competitors such as the largest classified automotive site Autotrader.com. An important feature of those trades on eBay Motors is that almost 75 percent of the final buyers of the vehicles are from the other states. Bidders generally gain information of the listing vehicles just through the information sellers provide on the webpage but not off-line interaction. Mandatory information includes car characteristics such as maker, model, mileage and also seller's information such as the percentage of positive feedback and number of feedbacks. Generally, a seller will voluntarily disclose more information such as text, graphics, photos subject to the charges $.15 per photo but nothing on text and graphics by eBay. In a standard listing of a vehicle, all mandatory information is on the top of all other disclosed information, such as photos.

All eBay Motor auctions use an ascending-bid format having the following two features. First, a fixed ending time is chosen for each auctioneer. Second, a proxy bidding system is used. A typical eBay motor auction goes in the following way. The auctioneer posts a starting price and auction begins. The first arrived bidder submits a cutoff price (highest bid he will accept) to the proxy system. The standing price will be the starting price plus a minimum increment. Suppose a second bidder arrives, then she chooses to submit a cutoff price or not. If she chooses to submit, the standing price will be the minimum of the two cutoff price plus a minimum increment and the bidder with low cutoff price will be notified via e-mail that she is outbid. The auction goes on in this way until the ending time which is an auctioneer's choice from 3,5,7,10 days. As a result of the proxy bidding system, a winner pays the second-highest bid plus the minimum increment. As soon as an auction ends, the winner is notified to pay and the information of all bidders' cutoff price except the highest bidder's becomes public available. During the auction, all bidders' identities, bidding times and actual bids are available to the public but not their cutoff prices.

1.3 The econometric model

I consider a sample of T eBay auctions of a single object with observed auction heterogeneity to accommodate the fact that in many cases including my application on eBay Motors auctions, the goods for sale differ in observable characteristics. One naturally expects that the distribution of valuations of potential bidders may shift with these characteristics. Specifically, let N_t denote the number of potential bidders in each auction t ($t = 1, \cdots, T$) and N_t is not observed by either econometricians or potential bidders which is a key difference between eBay auctions and traditional auctions. Let $\mathbf{z_t}$ be a realization of auction specific covariates $\mathbf{Z}$ in tth auction. For example, these characteristics can be mileages, age, make and so on. These information usually can be used to get the book value of the vehicle. In auction t, denote the bidder i's ($i = 1, \cdots, N_t$) private value (the utility of the selling good) by u_{it}. Private values are assumed to have the following form,

$$\ln u_{it} = \Gamma(\mathbf{z_t}) + \ln v_{it} \tag{1}$$

where $\Gamma(\mathbf{z_t})$ is a common part of the private value that is determined by the characteristics of the selling vehicle or the seller. The bidder-specific private information v_{it} ($t = 1, \cdots, T$ and $i = 1, \cdots, N_t$) is an independent draw of the random variable V_{it} which has an absolutely continuous distribution $F_{d_t}(\cdot)$ with support $[0, +\infty)$. $F_{d_t}(\cdot)$ depends on a dummy variable d_t represents the information disclosure level and $F_{d_t}(\cdot)$ is independent of $\mathbf{z_t}$. Equivalently, v_{it} is independent on the observed characteristics z_t conditional on the information disclosure level d_t. In this paper, d_t denotes a dummy for two groups of different quantities of photos. This means information disclosure level decides the type of consumer private signal distribution.Here, the bidder-specific private information V_{it} is assumed to be independent of $\mathbf{Z_t}$. [4] From the specification of the private values, it is necessary to have a location normalization on $\Gamma(\mathbf{z_t})$ (or private information distribution $F_{d_t}(\cdot)$).

Assumption 1 *Let the common value $\Gamma(z_t)$ satisfies that, for $\mathbf{z_0}$,*

$$\Gamma(\mathbf{z_t}) = (z_t - z_0)'\beta \tag{2}$$

This assumption is more than a location normalization because it also assumes that the common value has a linear form. The common value function $\Gamma(\cdot)$ is non-parametrically

[4] The additivity separability structure of the private values suggested by Haile, Hong and Shum(2003) is a practical way to avoid the curse of dimensionality.

identified but the linear form assumption significantly reduces the computational burden in estimation.

Each auction is executed over a time period selected by the seller. The ith potential bidder in auction t has at least one chance to monitor the auctions. A rational bidder will not let other bidder win at a price under his value and he will not make a bid which is larger than his value. Under the assumption of rational bidders, each bidder will bid his value when it comes to his last chance to monitor if the standing price is not over his value yet. Moreover, the bidder will not bid his value if the standing price is already over it.

It is possible that the bidder with the third highest value might not be able to submit any bid when the two highest-valued potential bidders submit their cutoff price before him which leads to a standing price higher than his value. The observed third order statistic of the observed bids can be below the third order statistic of the valuation. This probability of bias will decrease as the first- or the second-highest bidder submitting his last bid closer to the end of the auction. To eliminate the probability of observing a third bid which is not the third highest potential bidder's value, I assume that in all auctions, the first- or the second-highest bid is submitted late enough such that the third highest bidder has no chance to monitor the auction after it. Therefore, the third highest bid equals to the third highest value of all potential bidders. In practice, I will choose only those auctions with a first- or a second-highest bid submitted less than 10 seconds before the ends of them.

1.3.1 Identification

This subsection discusses the identification of bidder-specific private information distribution $F_{d_t}(\cdot)$ and the common part of bidder's private value Γ is shown below.

Intuitively, by the normalization that the common part of value at certain characteristic is a known constant C, the difference between type d_t's private information distribution $F_{d_t}(\cdot)$ and the value distribution of potential bidders in the subset of auctions with characteristic z_0 and d_t is just a location difference of the constant C. Athey and Haile (2002) show that a distribution is identified whenever the distribution of any order statistic with a known sample size drawn from the distribution is determined. In eBay auctions, a key difference is that the number of potential bidders is unknown. Song(2004) proves that a distribution $F(\cdot)$ is non-parametrically identified from observations of any two order statistics from an i.i.d sample, even when the sample size is unknown. The idea behind this is as follows. The density of second order statistics conditional on the third order statistics from an unknown sample size will not depend on the unknown sample size. This conditional density is also proportional to the first order statistic of a sample with size two from the same parent distribution. Since knowing any order statistic of the distribution in a known sample size is equivalent to know the distribution itself, it will be also enough to know any two order statistics in an unknown sample size to determine the distribution itself.

The second highest value always equals to the second highest bid in all eBay auctions since the bidder with second highest value will never see a standing price which is higher than his value. The third highest value equal to the third highest value under the assumption of late the first- or the second highest bid. The above argument implies that the value distribution of potential bidders in the subset of auction with the same characteristics (including the one with characteristic z_0 and d_t) is identified. It follows that $F_{d_t}(\cdot)$ is also identified since it is the value distribution of potential bidders in auctions with characteristic z_0 and d_t with a location change determined by the constant C.

The auction with characteristic $(\mathbf{z_t}, d_t)$ has a potential bidder's value distribution $H(u_{it}|\mathbf{z_0}, d_t)$ which is identified according to the argument above. For the same reason, the value distribution $H(u_{it}|\mathbf{z_t}, d_t)$ is also identified. And by the definition of Γ, the equation $\Gamma(\mathbf{z_t}) = E(\ln u_{it}|\mathbf{z_t}, d_t) - E(\ln u_{it}|\mathbf{z_0}, d_t)$ holds. The right-hand side of the equation is just the difference between the means of log value of potential bidder's in the corresponding subsets of auctions. The mean of log value can be derived from the distribution of the value. Hence, $\Gamma(\mathbf{z_t})$ is also non-parametrically identified.

The identification procedure also suggests that I should only use data from auctions in which the first- or the second-highest bidder submit their last bid late.

1.4 Estimation and Test

In this section, I will first derive the likelihood function as a functional of the model primitive $f_{d_t}(\cdot)$, $\Big($the density function of $F_{d_t}(\cdot)\Big)$ and $\Gamma(\cdot)$. It is well documented that semi-nonparametric (SNP) methods developed by (Gallant and Nychka (1987) etc.) can be used to approximate an unknown function and an unknown density function. One can replace the space of the unknown function by the corresponding seive space(a finite dimensional space). Practically, to avoid the problem of curse of dimensionality, I assume $\Gamma(\mathbf{z_t}) = (z_t - z_0)'\beta + C$ and leave the density function $f_{d_t}(\cdot)$ unspecified which leads to a semi-parametric model $(f_{d_t}\Big(\cdot), \beta\Big)$. Then, I will employ the SNP method to give a consistent estimator of $f_{d_t}(\cdot)$ and β. One advantage of the semi-parametric approach is that the Maximum Likelihood estimator(MLE) of parameter β will convergence at the normal parametric rate $\sqrt{T}$.

Let $(x_t, y_t, \mathbf{z_t})$ denote the third- and second-highest bids, and auction observed characteristics in auction $t = 1, \cdots, T$. Let d_t be a dummy variable of different information disclosure strengh. In the application, I will use the group of different quantity of photos as a proxy of d_t. Denote the second highest value and the third highest value of v_{it} by $v_t^{(2)}$ and $v_t^{(3)}$, respectively. The density of Y_t conditional on X_t, Z_t and d_t, $p(y_t|X_t = x_t, Z_t = z_t, D_t = d_t)$, is calculated as

$$\begin{aligned}
p(y_t|X_t = x_t, Z_t = z_t, D_t = 1) &= p\Big(v_t^{(2)} = exp(lny_t - z_t'\beta)|v_t^{(3)} = exp(lnx_t - z_t'\beta), D_t = 1\Big) \\
&= \frac{2[1 - F_1\Big(exp(lny_t - z_t'\beta)\Big)]f_1(exp\Big(lny_t - z_t'\beta)\Big)}{[1 - F_1\Big(exp(lnx_t - z_t'\beta)\Big)]^2} \\
p(y_t|X_t = x_t, Z_t = z_t, D_t = 0) &= p\Big(v_t^{(2)} = exp(lny_t - z_t'\beta)|v_t^{(3)} = exp(lnx_t - z_t'\beta), D_t = 0\Big) \\
&= \frac{2[1 - F_0\Big(exp(lny_t - z_t'\beta)\Big)]f_0(exp\Big(lny_t - z_t'\beta)\Big)}{[1 - F_0\Big(exp(lnx_t - z_t'\beta)\Big)]^2}
\end{aligned}$$

Since the likelihood function the joint density of (Y_t, X_t, Z_t, D_t) is a function of the unknown number of bidders, I consider the sample counterpart the sample partial likelihood function which is derived as follows

$$L_T(\hat{f}_1, \hat{f}_0, \beta) = \frac{1}{T}\sum_{t=1}^{T} \ln \frac{2[1 - \hat{F}_{d_t}\Big(exp(lny_t - z_t'\beta)\Big)]\hat{f}_{d_t}\Big(exp(lny_t - z_t'\beta)\Big)}{[1 - \hat{F}_{d_t}\Big(exp(lnx_t - z_t'\beta)\Big)]^2}$$

where $\hat{F}_{d_t}(v) = \int_0^v \hat{f}_{d_t}(t)dt$. The density function $\hat{f}_i$ (i=0,1) is an unknown function in the functional space $L_2[0, \infty)$. I will apply the SNP method (Chen 2009) to specify the density function $\hat{f}_i$ in a sequence of parameter spaces (sieve space).

Hereafter, I will use the following specification of $\hat{f}_{d_t}(v)$,

$$\hat{f}_{d_t}(v) = \frac{[\sum_{k=0}^{J_T} a_{k,d_t} L_k(v) \exp(-\frac{v}{2})]^2}{\sum_{k=0}^{J_T} a_{k,d_t}^2}, v \geq 0$$

where $L_k(v)$ is the kth Laguerre polynomials and $a_{0,d_t} = 1$. Laguerre polynomial series is an orthonormal basis of $L_2[0, \infty)$. The series length parameter J_T satisfies that $\lim_{T\to\infty} J_T = \infty$. The maximum likelihood estimator $(\hat{a}_1, \hat{a}_0, \hat{\beta})$ is the unique maximizer of the sample quasi-likelihood function, defined as

$$(\hat{a}_{1,1}, \hat{a}_{2,1}, \cdots, \hat{a}_{J_T,1}, \hat{a}_{1,0}, \hat{a}_{2,0}, \cdots, \hat{a}_{J_T,0}, \hat{\beta}) = \arg \max_{(a_{1,1}, a_{2,1}, \cdots, a_{J_T,1}, a_{1,0}, a_{2,0}, \cdots, a_{J_T,0}, \beta)} L_T(\hat{f}_1, \hat{f}_0, \beta)$$

The density estimator is

$$\hat{f}_{d_t}(v) = \frac{[\sum_{k=0}^{J_T} \hat{a}_{k,d_t} L_k(v) \exp(-\frac{v}{2})]^2}{\sum_{k=0}^{J_T} \hat{a}_{k,d_t}^2}$$

In application, since I propose to use a parametric test, thus the series length parameter J_T is chosen to be a fixed number. I also assume the parametric family of distribution is correct specified.

To test whether the bidders from the two groups of auctions have the same private value distribution(or density) or not, I employ the likelihood ratio test. The null hypothesis is that the two private value distributions are the same, or $H_0 : \forall i = 1, 2, \cdot, J_T, a_{i,d_0} = a_{i,d_1}$. The alternative hypothesis is that two private value distribution are different, or $H_1 : \exists i \in \{1, 2, \cdot, J_T\}$, such that $a_{i,d_0} \neq a_{i,d_1}$. The likelihood ratio test procedure is stated as follows. The first step is to estimate the model under the null hypothesis and the alternative hypothesis. The second step is to construct the likelihood ratio test statistic which is twice the ratio of log-likelihoods of the alternative model and the null model. By comparing to the $100 \times (1 - \alpha)$ percentile critical value of a Chi Square distribution with k (k is the dimension of restricted variables) degree of freedom, one rejects the null hypothesis if and only if the test statistic is larger than the critical value at significant level α.

1.5 Application to eBay Motor Auctions

1.5.1 Data

The data is originally collected by Lewis from the eBay auction webpages of completed used car auctions over an 8 months period. The whole dataset consists of 82538 observations of 18 models of vehicles. The variables of interest are collected from the HTML codes of those webpages. In each auction, some vehicle characteristics are mandatory and standard such as model, year, mileage, transmission type. The seller's feedback which is a proxy for seller's reputation is also observed. The magnitude of information disclosure in the item listing can be measured by the number of photos. On the other side of the market, all bidders' bids, thus the cutoff prices of all bidders except the highest value bidder in each auction are observed. Another important observation is the exact timing of each auction including the beginning time, ending time, and the bidders' bidding time. In my application, I will only use the sample with a first-highest bid or a second-highest bid within 10 seconds before the end of the auction. It is quite plausible that the third highest value bidder actually bids in the auction in my sample since there is little chance for her to see a higher standing price which prevent her from bidding. The resulting sample size is 3552 with the corresponding summary statistics in Table 1.

Table 1.1: Summary Statistics

Variables	Sample Mean	Std.Dev.	Min	Max
Book Value($)	9,862.4	7,650.7	889	45,097
Miles	91,410	67,311	1	500,000
Age(year)	8.2	4.3	1	17
% of Positive Feedback	98.5	3.8	60	100
Second bid	9,877.5	9,,094.2	112.5	77,600
Third bid	9,762.9	9,053.0	110	77,500
#of Photos	17.6	11.0	1	80
%Automatic Transmission	71.7			
% Pickup Truck	31.0			

This table provides summary statistics for using covariates.

The summary statistics show that on average the cars are old (17 years) and heavily traveled (over 90,000 miles). This indicates that the market might subject to the adverse selection effect of asymmetric information between sellers and buyers. Clearly, substantial variation exists in the number of photos that sellers choose to provide. Sellers have generally high reputation with an average 98.5% percentage positive feedback.

1.5.2 Baseline Estimation and Test

In the baseline estimation, I will choose log book value as the only exogenous characteristic $\mathbf{z_t}$. The implicit assumption is that the book value can capture all observed heterogeneity of cars. The required normalization in equation (2) with a log-linear specification of $\Gamma(\mathbf{z_t}) = \mathbf{z_t}\beta$ is equivalent to normalize $\beta = \frac{C}{\mathbf{z_0}}$ (for a constant C) or just $\beta = 1($ when $C = \mathbf{z_0})$. The estimating sample consists of 500 auctions with most photos and other 500 auctions with fewest photos. Let the dummy variable $d_t = 1$ if the auction is in top 500 group, otherwise

$d_t = 0$. I assume the distribution of private value is in the parametric family of length 3 Laguerre polynomials series. The assumption is given as

Assumption 2 *The private value density function $f_{d_t}(t = 0, 1)$ has the following form,*

$$\hat{f}_{d_t}(v) = \frac{[\Big(1 + a_{1,d_t}L_1(v) + a_{2,d_t}L_2(v) + a_{3,d_t}L_3(v)\Big)\exp(-\frac{v}{2})]^2}{1 + a_{1,d_t}^2 + a_{2,d_t}^2 + a_{3,d_t}^2}, v \geq 0$$

where $L_k(v)(k = 1, 2, 3)$ is the kth Laguerre polynomials such that

$$L_1(v) = 1 - v, L_2(v) = \frac{1}{2}(v^2 - 4v + 2), L_3(v) = \frac{1}{6}(-v^3 + 9v^2 - 18v + 6)$$

and a_{k,d_t} is the parameter.

To test whether the private value distributions of the two groups are same or not, I use the likelihood ratio test as follows. The null hypothesis is that the two private value distributions are the same, or H_0 : for all $i = 1, 2, \cdot, J_T, a_{i,d_0} = a_{i,d_1}$. The alternative hypothesis is that two private value distributions are different, or H_1 : there exists $i \in \{1, 2, \cdot, J_T\}$, such that $a_{i,d_0} \neq a_{i,d_1}$. The test procedure is stated as follows. The first step is to estimate the model under the null hypothesis and the alternative hypothesis. The second step is to construct the likelihood ratio test statistic and compare to the relative chi-square critical value. The estimates and log-likelihood of the full model (under alternative hypothesis) and the restricted model (under the null hypothesis) is shown in Table 2.

Table 1.2: Estimates of Distributions

	Full Model		Restricted Model	
	Coefficient	Std. Dev.	Coefficient	Std. Dev.
Distribution Parameter				
a_{1,d_0}	0.559	0.0055	0.558	0.0036
a_{2,d_0}	0.282	0.0060	0.287	0.0042
a_{3,d_0}	0.119	0.0063	0.125	0.0058
a_{1,d_1}	0.558	0.0042	0.558	0.0036
a_{2,d_1}	0.290	0.0048	0.287	0.0042
a_{3,d_1}	0.129	0.0060	0.125	0.0058
log-likelihood	1063.7		1062.2	

The test statistic is twice the ratio of log-likelihoods of the alternative model and the null model which is 2.003. Comparing to the critical value 4.60 of significant level 10% of a χ^2

distribution with 2 degrees of freedom. The test fails to reject the null or different amount of photos does not significantly change the distribution of private value. It also implies that the increase of number of potential bidders explains the increasing selling price.[5] According to the test result, I will use the estimates of the restricted model that the two distributions of different information disclosure levels are the same to compute the consumer surplus.

The consumer's surplus equals to the valuation of the winner minus the selling price which is the second highest bid. The consumer surplus in auction t is calculated as,

$$\begin{aligned} CS_t &= U_t^{(1)} - U_t^{(2)} \\ &= z_t(V_t^{(1)} - V_t^{(2)}) \end{aligned}$$

where $V_t^{(i)}$ is the i th highest private value and z_t is the characteristic of the car and seller in auction t. The expected consumer's surplus is calculated as follows:

$$E[CS_t | U_t^{(2)} = u_{2t}] \quad = \quad z_t \cdot \Big(\int_{v_t^{(2)}}^{\infty} \frac{f_{d_t}(v)}{1 - F_{d_t}(v_t^{(2)})} \cdot v dv \Big) - u_{2t}$$

where the second valuation satisfies that $u_{2t} = z_t v_t^{(2)}$. I also estimate the expected share of consumer's surplus comparing to the selling price as $E[\frac{CS_t}{U^{(2)}} | U_t^{(2)} = u_{2t}]$. The summary statistics of the estimate of consumer's surplus and share of consumer surplus at each auction are shown in Table 3. The sum of consumer's surplus (1000 auctions) is 5.9 Million dollars.

Table 1.3: Consumer Surplus and Share of Consumer Surplus

	Mean	Std. Dev.	Median	Min	Max
CS($)	5935.4	8655.1	3911.6	287.7	141726.4
CS share	61.6%	36.2%	41.8%	40.8%	357.2%

Comparing to the sample book value median $8596, the median Consumer Surplus is 45.5% which is quite significant because in the model shipping costs and the opportunity costs of monitoring the auctions are not counted. If these costs are considered, the consumer surplus estimates are overestimated. The potential gain of using the eBay platform could be even larger because of the potential supplier' surplus.

[5] It is not possible to identify the number of potential bidders unless more assumptions on the bidding behavior are made. But still in the data, there exists a positive correlation (or the number of actual bids) between the number of actual bidders and the number of photos.

1.5.3 Estimation and Test with Observed Characteristics

To capture the heterogeneity of cars that are auctioned, I will use the following six characteristic variables: log mileage, age, transmission type (Auto or Manual), dummy for classic car, dummy for pickup truck and percentage of seller's good feedback[6]. The estimating sample is chosen the same way which includes two groups: 500 auctions with most photos and other 500 auctions with fewest photos. The variable d_t is the group dummy such that $d_t = 1$ if the auction is in the top 500 group. For the following estimation result in Table 4, I choose the series length J_T to be 2. In order to get the confidence interval of estimates, I assume the distribution of private value is in the parametric family of length 2 Laguerre polynomials series. [7] It is stated as follows,

Assumption 3 *The private value density function $f_{d_t}(t = 0, 1)$ has the following form,*

$$\hat{f}_{d_t}(v) = \frac{[\big(1 + a_{1,d_t} L_1(v) + a_{2,d_t} L_2(v)\big)\exp(-\frac{v}{2})]^2}{1 + a_{1,d_t}^2 + a_{2,d_t}^2}, v \geq 0$$

where $L_k(v)(k = 1, 2)$ is the kth Laguerre polynomials such that

$$L_1(v) = 1 - v, L_2(v) = \frac{1}{2}(v^2 - 4v + 2),$$

and a_{k,d_t} is the parameter.

My primary empirical object is to test whether the bidders from the two groups of auctions have the same private value distribution or not. I use the likelihood ratio test as follows. The null hypothesis is that the two private value distributions are the same, or $H_0 : a_{1,d_0} = a_{1,d_1}$ and $a_{2,d_0} = a_{2,d_1}$. The alternative is that two private value distribution are different, or $H_1 : a_{1,d_0} \neq a_{1,d_1}$ or $a_{2,d_0} \neq a_{2,d_1}$. The test procedure is implemented in the following way. The first step is to estimate the model under the null hypothesis and the alternative hypothesis. The second step is to construct the likelihood ratio test statistic and compare to the relative chi-square critical value. The estimates and log-likelihood of the full model (under alternative hypothesis) and the restricted model (under the null hypothesis) is shown in Table 4.

The test statistic is twice the ratio of log-likelihoods of the alternative model and the null model which is 2.002. Comparing to the critical value 4.60 of significant level 10% of a

[6] The cars are divided into three groups:"reliable" cars (Japanese cars including Honda, Toyota), "classic" cars (vintage and muscle cars), "pickup" truck (e.g. Ford F-series, Dodge Ram). See Lewis (2010).

[7] There is no existing result for the confidence interval of the semi-nonparametric estimator.

Table 1.4: Estimates of Distributions and Characteristic Parameters.

	Full Model		Restricted Model	
	Coefficient	Std. Dev.	Coefficient	Std. Dev.
Distribution Parameter				
a_{1,d_0}	0.348	0.113	0.348	0.093
a_{2,d_0}	0.091	0.065	0.088	0.054
a_{1,d_1}	0.348	0.096	0.348	0.093
a_{2,d_1}	0.087	0.054	0.088	0.054
Characteristic Parameter				
log miles	-0.083	0.017***	-0.081	0.006***
Age	-0.094	0.035***	-0.095	0.007***
Good feedback (%)	0.013	0.001***	0.012	0.002***
Dummy, Auto	0.043	0.003***	0.043	0.042
Dummy, classic	-0.751	0.033***	-0.751	0.027***
Dummy, pickup	-0.259	0.059***	-0.258	0.018***
Constant	9.155	0.171***	9.145	0.347***
log-likelihood	659.7		659.1	

χ^2 distribution with 2 degrees of freedom. The test fails to reject the null which means a higher information disclosure level does not change the distribution of private value. Hence, the reason for the increase of selling price should be an increase in the number of potential bidders. According to the test result, I will use the estimates of the restricted model that the two distributions of different information disclosure levels are the same in the following analysis.

It is natural to expect that consumer prefers a car with low mileage, age and a seller with a better reputation. The estimates confirm the conjecture as it shows that mileage and age of a car has a significant negative effect on bidder's utility. Seller's higher percentage feedback increases bidder's utility and consumer prefers "Reliable" Japanese cars. Transmission type has an insignificant positive effect on the value of consumers.

The consumer surplus in auction t is calculated as,

$$\begin{aligned} CS_t &= U_t^{(1)} - U_t^{(2)} \\ &= exp(z_t\beta)(V_t^{(1)} - V_t^{(2)}) \end{aligned}$$

where $V_t^{(i)}$ is the i th highest private value and z_t is the book value of the item in auction t. Since the highest bid $V_t^{(i)}$ is not observed, I estimate the expected consumer's surplus as follows:

$$E[CS_t|U_t^{(2)} = u_{2t}] \quad = \quad exp(z_t\beta) \cdot \Big(\int_{v_t^{(2)}}^{\infty} \frac{f_{d_t}(v)}{1 - F_{d_t}(v_t^{(2)})} \cdot vdv \Big) - u_{2t}$$

where the second valuation satisfies that $u_{2t} = exp(z_t\beta)v_t^{(2)}$. The expected share of consumer's surplus comparing to the selling price $E[\frac{CS_t}{U^{(2)}}|U_t^{(2)} = u_{2t}]$ is also estimated. The summary statistics of the estimate of consumer's surplus and share of consumer surplus at each auction are shown in Table 5. The sum of estimated consumer's surplus (1000 auctions) is 4.1 million dollars.

Table 1.5: Consumer Surplus and Share of Consumer Surplus

	Mean	Std. Dev.	Median	Min	Max
CS($)	4070.3	6136.5	2136.5	449.1	67669.4
CS share	48.2%	47.0%	32.0%	19.7%	511.4%

This result is consistent with the finding in Song (2004) and (Giray et al.(2009)) that consumer gains a significant amount of surplus in eBay auctions.[8] The mean (Median) estimated consumer surplus share is 48.2%(32.0%) and the mean(median) consumer surplus is $4,070$\$(2,136$\$).The mean (and median) consumer surplus is of the same magnitude but smaller than the mean (and median) consumer surplus estimate in the baseline estimation. The real consumer surplus could be actually less than the estimate from the model since the existence of unknown shipping costs and possible entry costs.

1.6 Conclusion

It is important to understand how online auction market works given the rapid growth of this market. Sellers voluntarily disclose their information of selling cars to increase selling prices. Two reasons can explain the effect of information on selling price: An increase of the number of bidders or/and a change of bidder's value distribution. This paper tests if the value distribution is influenced by the information level under the first-price auction framework with observed heterogeneity. The testing result shows that the value distribution does not

[8] Song (2004) estimates $36.26 an average comparing to an average selling price $31.3 in eBay yearbook auctions. Giray et al.(2009) estimate an mean consumer surplus share around 20 to 30 percent in various specifications of model in eBay computer auctions.

vary across different information level in the dataset of eBay motor auction. It implies that the reason for the increasing selling price is an increase of the number of potential bidders.

It seems to be a controversy that the number of bidders increase but the value distribution does not change. It would be more interesting to test whether an endogenous entry auction model is more favorable to explain the online auction data.

The information to disclose is actually endogenously chosen by auctioneers. One direction for future research is to incorporate that auctioneers optimally choose the amount of information to disclose comparing the potential gain from selling price and the cost of disclosing information.

1.7 References

[1] Akerlof, G.(1970). “The Market for ‘Lemon’: Quality Uncertainty and the Market Mechanism.”. *The Quaterly Journal of Eocnomics*, 84: 488-500.

[2] Athey, S and Haile, P.(2002). “Identification of Standard Auction Models.”. *Econometrica*, 70: 2107-2140.

[3] Athey, S and Haile, P.(2007). “Nonparametric Approaches to Auctions.”. *Econometrica*, 70: 2107-2140.

[4] Bajari,P and Hortacsu, A.(2007). “The Winner’s Curse, Reserve Prices and Endogenous Entry: Empirical Insights from eBay auctions.”. *Rand Journal of Economics*, 2003: 329-355.

[5] Bierens, H. and Song, H. (2006). “Semi-nonparametric Estimation of First-price Auctions with Auction-specificÂĚc Heterogeneity using Simulated Method of Moments, ”. *Working Paper*.

[6] Chen, X.(2007). “Large Sample Sieve Estimation of Semi-nonparametrci Models.”. *Handbook of Econometrics*, 76.

[7] Gallant,R. and Nychka, D.(1987). “Semi-nonparametric Maximum Likelihood Estimation.”. *Econometrica*, 55: 363-390.

[8] Giray, T., Hasker, K., Jiang, B. and Sickles, R. (2009). “Estimating Consumer Surplus in eBay Computer Monitor Auctions.”. *Working Paper*.

[9] Guerre, E. Perrigne, I. and Vuong, Q. (2000). “Optimal Nonparametric Estimation of First-price Auctions.”. *Econometrica*, 68: 525-574.

[10] Haile, P. Hong, H. and Shum, M. (2003).“Nonparametric Tests for Common Values at First-price Sealed Bid Auctions.”. *NBER working paper*.

[11] Hoel,G, Port,C, and Stone,J.“Testing Hypothesis”. *Introduction to Statistic Theory*, Chapter 3.

[12] Krishna,V.(2002). “Auction Theory.”. *San Diego: Academic Press*.

[13] Lewis,G.(2010). “Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors.”. *American Economic Review forthcoming*.

[14] Newey, W and McFadden,D.(1994). “Large Sample Estimation and Hypothesis Testing.”. *Handbook of Econometrics*, 35.

[15] Ockenfels, A and Roth,A.(2002). “Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet.”. *American Economics Review*, 92:1093-1103.

[16] Song,U.(2004).“Nonparametric Estimation of an eBay Auction Model with an Un-

known Number of Bidders.". *Working Paper*.

Chapter 2
Consumer Search in the Online Movie DVD Market

2.1 Introduction

This paper aims to analyze search frictions in online movie DVD retail market within the framework of consumer search model. This paper utilizes novel and detailed data on the web browsing and purchasing behavior of a large panel of consumers from the ComScore Web Behavior Database. The dataset is novel in that it provides detailed information tracking of all consumer searches prior to each transaction. Using observed search and transaction data, I estimate search costs in the movie DVD retail market industry. These estimates can help explain reasons of search cost heterogeneity.

In online market, many empirical works show that the price dispersion persists.[1] This motivates the empirical search application to online markets. Hong and Shum(2006) rationalize the prices setting by online textbook stores with search models. They find the magnitude of search cost in a sequential model is larger than that in a fixed sample size search model (or nonsequential search model). De Los Santos, Hortacsu, and Wildenbeest (2012) empirically test the two classical search models using observed online book purchasers' search behavior in ComScore Web-Behavior Panel dataset. They find that fixed sample size search model provides a more accurate description in this particular market of online bookstores. They also show evidence that some consumers do not buy form the lowest priced store in their sample and some stores (e.g. Amazon) are searched more often. Both observations lead to that bookstores differ in aspects (e.g. ease of using interface) other than price.

My demand model is closely related to a stream of papers that utilize discrete choice search models of differentiated products. In marketing literature, Mehta, Rajiv and Srinivasan are the pioneers to link the "consideration set" model to nonsequential search model. Gonzalez,Sandor and Wildenbeest (2011) use aggregate data to estimate their model. However, consumers'

[1] See Brynjolfsson and Smith(2000), Clay and Wolff(2001).

search behavior and choice sets decisions are not observed in their data. My demand model is close to Honka's (2010) which exploits data on individual consumers' choice sets. In another stream of papers, Kim, Albuquerque and Bronnenberg (2010) and Koulayev (2009) estimate sequential search models of demand. Kim, Albuquerque and Bronnenberg (2010) utilize aggregate information on search behavior at Amazon. On the other hand, Koulayev (2009) exploits individual search histories on hotel price comparison engine. Bronnenberg, Kim and Mela(2014) find that consumers search for more attributes rather than price, and consumers tend to purchase from the later searched websites. These facts are consistent with the sequential search model.

The rest of the paper is organized as follows. Section 2 describes the data and discuss the search patterns. Section 3 presents the structural model and estimates the search costs in the market. Section 4 concludes with implications for future research.

2.2 Data

This paper uses a dataset comes from the ComScore Web Behavior Database including detailed online searching and transaction information as well as the panelists' demographic information. ComScore is a public company and a leading provider of online browsing and purchasing behavior information of users across the United States. It maintains a panel of more than 2 million global Internet users and were chosen at random by ComScore. Each user's online activity is recorded through ComScore proxy servers. The data includes date, time, duration of visit, price, quantity, description of purchased product during each session. [2] A random sample of 1,760,165 distinct transactions of products and services from 61 categories by 104,107 users was chosen in the ComScore Web Behavior Database in 2006 and 2007.[3]

The dataset used in the empirical analysis is constructed from two parts. The first part is the transaction data which includes price, quantity, product category and name of each product purchased as well as the demographics of consumers. The second part is the browsing/searching behavior data and includes information on website browsing history

[2] Monitoring softwares are installed on the computers of the panelists, with brands including Permission-Research, OpinionSquare and VoiceFive Networks. In exchange for joining the comScore research panels, users are presented with various benefits, such as computer security software, Internet data storage, virus scanning and chances to win cash or prizes. *Source: Wikipedia.com*

[3] On average, each consumer made 16.9 purchases, from 4.4 categories, bought from 4.0 web stores and spent 46.3 US dollars for each transaction.

including websites visited (including referral websites), date, time, duration, and pages viewed of each visit. Specifically, each consumer is identified as one computer id in the dataset. Thus, it can capture the browsing and transaction history of the whole family. Demographic information of these families includes household head education level, household income, household size, racial background, census region and so on. Comparing to the Current Population Survey, the ComScore sample are representative of online buyers in the United States.[4] The dataset also contains detailed information of duration and the number of pages viewed in each visit. In traditional markets, in addition to the customer service and consumer idiosyncratic preference for a given retailer, products from different retailers are often differentiated by location, and in turn search cost includes the monetary costs of acquiring information and the opportunity cost of the time spent. With the development of e-retail, products are no longer differentiated by their location and search costs constitute of mostly opportunity cost. The time spent on each search is a good approximate of such opportunity cost, given consumers' demographic information.[5]

2.2.1 Summary Statistics

I focus on the online movie DVD retail market in this empirical analysis. The benefit is that, in this category, the good with the same product name from different retailers can be treated as homogeneous good.(e.g. Spider Man 3 DVD from Amazon and Columbiahouse) Since the sample of users is randomly chosen in each year, I will examine the sample of users appearing in both year 2006 and year 2007 to study the consumer's search behavior overtime.[6] Each transaction observation records the detail web browsing history of a single purchase possibly multiple units of the same products. According to the pattern in the data, in each transaction session, a consumer can have several transaction of different products(e.g. Spider Man 3 DVD and Spider Man 2 DVD). Different products are recorded in different observations. In the browsing/searching behavior data, I can only identify the website name but not the product category if there is no transaction happened in this search session. A consumer may browse a website just to surf the Internet or to look for products of different categories if a website is not specialized in a particular category. I exclude observations that

[4] See De Los Santos (2008) for a more detailed decription of the data, and De Los Santos, Hortacsu, and Wildenbeest (2012) for discussion about the representativeness of the dataset.

[5] The Internet connection speed may influence the duration of search and the connection types (broadband or not broadband) is a good indicator of the connection speed in the data.

[6] This treatment gives me a longer panel which is good to check the behavior pattern over time. However, this limits my sample with fewer observations.

could not be identified as online movies and videos stores such as web portal yahoo.com. According to the dataset, a total number of 757 users made at least one transaction from 13 different websites with a total number of 2,624 transaction sessions in this category.

From the above argument, the number of visits of each website is an upper bound of the number of visits related to a particular transaction in movie DVD category. Following the discussion on related search behavior in De los Santos, Hortacsu, and Wildenbeest (2012), I define relevant searches as the browsing history (on relevant websites) up to seven days before a transaction, if no other transaction has occurred within these seven days. Otherwise, a search history can contain less than seven days is up to the most recent transaction in the past. To show the summary statistics of the sample, I also use an alternative of relevant searches definition of browsing history during 1 day period before a transaction.

Table 1 displays the summary statistics of this sample. The first column summarizes numbers of transactions from each site. In total, we have 13 online retailers (.com) and the top 5 retailers shared 90 percent of the market: Amazon (44 percent of transactions), Columbia House (30 percent), Go (9 percent), Walmart (4 percent), Overstock (3 percent). Total Numbers of all visits and relevant visits (in 7-days and 1-day periods) are summarized in the second, third and fourth columns. The last three columns shows that the ranking of transactions among online retailers is not perfectly correlated with the ranking of visits to these sites. A consumer may not need to visit a website multiple times to make a purchase if she has a strong preference to this particular website and may visit a website many times that she does not know very well and still not buy from it. Table 2 and Table 3 summarize the transaction and search behavior of average consumer from this sample. On average, each consumer has made 3.5 transactions from 1.3 different movie online retailers. Each transaction on average costs 14.51$. Each web retailer visit on average takes 9.7 minutes for visits happened without the 7-day search period, 13.2 minutes for visits without transactions happened within the 7-day search period, and 28.2 minutes for visits with transactions. This may be due to the fact that choosing products and checking out incur extra time on visits with transactions. The second panel of Table 3 shows that for each 7-day search period, 63.6% of the time is spent on visiting the websites of the stores with realized transactions by an average consumer. The third panel summarizes a number of visits during each 7-day search period. An average online shopper visited 4.4 websites (including repeated visits to the same online retailer). Even though on average a consumer is aware of 8 among the 13 online retailers, but only 2.4 of them are visited each time.[7] Table 4 summarizes the distribution of

[7] Given the large number of online retailers relative to the small number of stores actually visited, I assume

the number of visited retailers conditional on the number of consumer's aware stores. For those shoppers knowing more than 1 store, no more than 0.1% of the times they will visit all their aware stores. This pattern is consistent with the situation when a search is costly.

In the dataset, the price of a product is available only if it appears in a transaction. To recover missing prices for the other visited stores of a product, I will follow the method proposed by De los Santos, Hortacsu, and Wildenbeest (2012) to use the most recent transaction prices of the same products on those stores. The relative homogeneity among products in movies DVD market allow me to extrapolate a large portion of the price data. Table 5 shows the summary statistics of the matched price. In the average 2.4 visited stores, the price of the same product can be found in 1.4 of those stores. All the prices of the visited websites can be matched in 1,652 (63%) of the total 2,624 transaction and visit periods. Table [8] Table 6-8 displays descriptive statistics for 10 best-selling movie DVD in the year 2006 and year 2007 transaction sample. The mean prices are quite similar for each movie DVD across websites except for Harry Potter 5. The selling price of a movie DVD on Amazon is not always higher than those prices on Columbiahouse or other retailers' websites.

2.2.2 Empirical Evidence

This subsection discusses some empirical evidence regarding consumers' searching and purchasing behavior in the online movies DVD retail market. According to the pattern in the data, consumer sometimes chooses to buy from a retailer with a higher price even she visits at least one other retailer with a lower price. (234 out of the 1,652 transaction periods with all visited retailers' price matched) This fact shows that consumer makes her purchase decision on both price and other retailer's specific features or its quality.

In e-retail markets, as discussed above, products are no longer differentiated by their location and search costs constitute of mostly opportunity cost. The time spent on searching is a good measure of search cost given consumers' demographic information since it represents the opportunity cost of time. I explore the determinants of consumer search cost heterogeneity using this measure of search cost and display the results in Table 9. To exclude the unrelated website browsing behavior, I only include the time spent on searching in the 7-day search period defined above in this category. Table 9 presents estimates of regression of the total time spent on searching before a transaction of movie or video on household characteristics.

that the consumer is aware of a given retailer if she has ever visited the store.

[8] The number of periods when all visited retailers' price is matched will increase to 2,325 by dividing all 13 websites into 3 groups (Amazon, Columbiahouse, and other websites.)

A consumer on average spends 40 more minutes on an extra firm. A consumer spends less time (11 minutes) in searching when they had repeated purchases on the same retailer as we would expect. Search cost is also influenced by individual characteristics. A larger family tends to spend more time on searching. Retired consumers tend to search longer which is also found in different scenarios.[9] It is interesting to notice that individuals search 3 minutes more in their next search on average. This may be explained by heterogeneity in intrinsic shopping preference since the people who shop more often may enjoy shopping and spend more time on shopping.

In search models, the search cost is usually measured by the number of stores visited. Table 10 shows estimates of regression of the number of visited retailers in 7-day search period on household characteristics. Shoppers on average visit 0.01 more retailers with an extra quantity of purchase. Consumer visits 0.25 fewer retailers if they purchase from the same retailer in last period of purchase. A larger size household and consumer younger than 21 tends to visits more retailers.

2.3 Model

In this section, I develop a demand model that captures consumer's search behavior. Consumer demand one unit of a homogeneous good. Firstly, with the knowledge of price distribution, consumers choose a subset (search set or choice set) of all the firms to search in order to know their exact prices. I assume that the consumers use a non-sequential search strategy which means they will visit all the firms in the search set not in a particular order.[10] Secondly, a consumer makes a purchase among the searched firms once they have the information on prices. Consumers are assumed to make the search decisions to maximize their expected utilities. The purchase decision of consumer is to pick the highest utility firm product after all the prices of the firms in the search set are known.

Consider an online retail market with $j = 1, \cdots, J$ online firms selling a homogeneous good. I allow consumers to have heterogeneous store preferences. Thus, I assume consumer i's $(i = 1, \cdots, I)$ utility from store j is given by

$$u_{ij} = \mu_j + X_i\beta_j - \alpha_i p_j + \epsilon_{ij}, \tag{1}$$

[9] Aguiara et al (2005) finds that retirement reduce one's unit spend on food consumption which can be explained by more time spent on searching food prices.

[10] Consumers purchase from firms that are searched earlier in the sequence of searched firms. Such recall behavior by consumer is consistent with a non-sequential search model.

where α_i is a consumer-specific price coefficient and ϵ_{ij} is an idiosyncratic utility shock. Before visit the store website, consumers know their utility but not the price from each store. Consumers need to search the store website to learn the price.

Consumers decide to choose a subset of the total J firms to search and then purchase the product from one of the sampled firm. The search is costly based on the empirical evidence. I also allow a general form for the search cost of consumer i,c_i, to depend on her characteristics, $c_i = c + X_i\beta$. The expected utility of consumer i searching a subset S of all stores, denoted by U_{iS}, is given by

$$U_{iS} = E_\epsilon[\max_{j \in S} u_{ij}] - ||S||c_i, \tag{2}$$

where $||S||$ is the number of firms in subset S. To smooth the choice set probabilities, I add a type I extreme value distributed search set specific noise δ_{iS} with scale parameter σ_δ to U_{iS}. This noise can be explained as the individual assessment errors of the expected utility. Therefore, the probability of consumer i choose a search set S, denoted by P_{iS} has the following logit form,

$$P_{iS} = \frac{\exp[U_{iS}/\sigma_\delta]}{\sum_{S' \in S} \exp[U_{iS'}/\sigma_\delta]}. \tag{3}$$

After making the search set decision, the consumers knows the price of all stores in their search sets. Since consumer has no uncertainty in this stage, she simply purchases from the highest utility firm. The purchase probability $P_{ij|S}$ is then,

$$P_{ij|S} = \Pr(u_{ij} > u_{ik}, \forall k \neq j \in S_i). \tag{4}$$

If a consumer has a higher utility shock on a firm, the firm is more likely in consumer's search set. Hence, conditional on the search set decision, the utility shock ϵ_{ij} is not i.i.d. This means the conditional probability of purchasing in equation (4) has no closed-form solution.

The probability of observing search set S and purchasing product j,denoted as P_{ijS} will be $P_{ijS} = P_{iS}P_{ij|S}$.

2.3.1 Estimation

I estimate the model using a two-stage estimation method based on that of De Los Santos, Hortacsu, and Wildenbeest (2012). In the first-stage estimation, I recover the parameters of price distributions by utilizing the observed prices. The utility parameters and search cost parameters are recovered in the second-stage of the estimation by maximum likelihood method.

I follow Mehta, Rajiv and Srinivasan(2003)'s assumption on price distribution in order to have a closed-form expression for the expected utility of a search set S, $E_\epsilon[\max_{j\in S} u_{ij}]$ which speeds up my estimation process. A closed-form expression allows me to avoid the potential numerical integration in the process of deriving the expected utilities. [11] By assuming the prices are type I extreme value distributed with store specific location parameter γ_j and a common scale parameter σ_γ, the expected utility of consumer i searching a subset S,U_{iS} will have a closed-form expression as follow,

$$\begin{aligned} U_{iS} &= E_\epsilon[\max_{j\in S} u_{ij}] - ||S||c_i \\ &= \alpha_i\sigma_\gamma \log(\sum_{j\in S} \exp[\frac{\mu_j + X_i\beta_j - \alpha_i p_j + \epsilon_{ij} + \alpha_i\gamma_j}{\alpha_i\sigma_\gamma}]) - ||S||c_i. \end{aligned} \tag{5}$$

The parameters of the type I extreme value price distributions are estimated by maximum likelihood method, using the de-meaning prices for all transactions in the data to control the unobserved differences in movie DVD characteristics. [12] Denote the observed search set of consumer i by S_i and the purchase firm by d_i. This leads to the following likelihood function:

$$\begin{aligned} L(\beta) &= \sum_i \log P_{ijS_i} \\ &= \sum_i \log(P_{iS_i} P_{id_i|S_i}) \\ &= \sum_i \log(\frac{\exp[U_{iS_i}/\sigma_\delta]}{\sum_{S'\in S}\exp[U_{iS'}/\sigma_\delta]} \Pr(u_{id_i} > u_{ik}, \forall k \neq d_i \in S_i)) \end{aligned} \tag{6}$$

The utility parameters will influence on consumer's utilities u_{ij} for each firm and also the consumer's expected utilities of selecting a particular choice set S. Hence, utility parameters will influence on both p_{iS} and $p_{ij|S}$. However, the search cost only has influence on p_{iS}. As in the second stage of the purchase decision, the search cost is already paid as a sunk cost.

The probability of observing search set S_i, P_{iS_i} has a closed-form expression. However, I will rely on simulation to get the probability of purchase from firm d_i out of the search set S_i. I simulate 100 consumers for each observation with their utility shocks follow a type I extreme value distribution. To be noticed, to identify the demand coefficients in discrete choice models,

[11] This assumption help me to avoid the numerical integration which can be introduced by the other distribution assumptions. For example, a normal distribution assumption on price distribution will slow down the estimation process substantially.

[12] The estimated location parameters are 24.722 for Amazon,4.8035 for Columbiahouse and 46.8542 for other firms.

I choose to normalize the variance of the utility shock distribution and estimate the variance of the stochastic searching optimization error σ_δ.

I model consumer i's search and purchase decision period as each of her observed transaction through the 2 years of data. Because of the price matching issues discussed in the data section, I restrict the number of firms to be 3. The three firms are Amazon, ColumbiaHouse and the other firms.

Following the above maximum likelihood method procedure, I obtain the estimates of the model and report them in Table 2.11. The price coefficients for three different income groups (more than $\$75,000$, less than $\$35,000$, and between $\$35,000$ and $\$75,000$) are estimated. The estimates suggest that the lower income group is more price sensitive. The estimated search cost is substantially significant. Normalizing by the price estimates, the search costs are around \$2.73.[13] The search cost decreases as a household has a broadband connection. An additional household member increases the search cost.

The store fixed effect estimates are significantly different across firms. Without any surprise, Amazon has the highest firm fixed effect since it is the highest searched firm. In additional, Columbiahouse has larger firm fixed effects than the Other Movie DVD retailers. The estimates also indicate that preferences on firms are influenced by the consumer demographics. For example, compared to buying at Other Movie DVD retailers, adding a household member, the marginal utility of buying at Amazon and Columbiahouse decreases. Meanwhile, the marginal utility of purchasing at Amazon is increasing faster when the household income pass over $\$50,000$, comparing to other Movie DVD retailers and the opposite happens for the marginal utility of purchasing at Columbiahouse.

The consumers with ages over 30 prefer Amazon more than the consumers with an ages less than 30. This result is consistent with that old consumers may have stronger store preference than younger consumers. The consumers with income over $\$50,000$ prefers Amazon and Columbiahouse more than the consumers with income less than $\$50,000$. This estimation result indicates that high income consumers may have stronger preference on particular brands.

[13] The estimated average search cost is less than the average estimated search cost \$1.67 in online book industry.(Santos Hortacsu Wildenbeest(2012)) The reason can be that the movie DVD industry is relatively more concentrated than the book industry.

2.4 Conclusion

In this paper, I study the consumers' online search behavior that has been documented thoroughly in the industrial organization literature. Using unique features of Comscore consumer search behavior data and transaction data, I focus on the online movie DVD retail industry. I find several features of consumer search behavior in this specific market. Particularly, consumer makes her purchase decision on both price and other retailer's specific features or its quality.

I estimate search costs in a nonsequential search model describing the web browsing and purchasing behavior. Consumer knows the price distribution and she can ascertain the exact price of each retailer by searching on the firm website. Following the canonical search model, a consumer has two decision steps in each period. First, each consumer makes a choice set decision on how many retailers' websites to visit. After gaining information on the exact price of each retailer in the choice set, a consumer maximizes her utility over the retailers in her choice set. The estimation result shows the average search costs are around $2.73 for an individual to search a firm's website before each transaction.[14]

Throughout the paper, I try to show what are the online consumers' search behavior and what are the possible factors influence on them. I consider that developing a structural model that includes details of consumer's searching behavior with learning on firm's quality is a necessary step for explaining the online DVD retail markets. I plan to follow this line of research in my future work.

2.5 References

[14] The estimates of search costs are $1.67 in online book retail industry in Santos Hortacsu Wildenbeest(2012). The results confirm the conjecture that book retail industry is relatively more competitive than the movie DVD retail market.

[1] Ackerberg, D. (2003). “Advertising Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination”. *International Economic Review*, 44: 1007-1040.

[2] Bronnenberg, Bart J., Byeong Jun Kim, and Carl F. Mela. "Zooming In on Choice: How Do Consumers Search for Cameras Online?."

[3] Brynjolfsson, E., and Smith,M. (2000). “Frictionless commerce? A comparison of Internet and conventional retailers”. *ÂŤManagement Science*, 46: 563-ÂŰ85.

[4] Burdett, K., and Judd,K. (1983). “Equilibrium Price Dispersion”. *Econometrica*,51: 955-969.

[5] Clay, K., Krishnan R., and Wolff, E. (2001). “ÂŞPrices and price dispersion on the Web: Evidence from the online book industry”. *ÂŤJournal of Industrial Economics*, 49, 521-ÂŰ39.

[6] De los Santos, B. (2008). “Consumer Search on the Internet”. *NET Institute Working Paper*, #08-15.

[7] De los Santos, B., Hortacsu, and A., Wildenbeest, M. (2012). “Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior”. *American Economic Review*, forthcoming.

[8] Erdem,T., and Keane, P. (1996). “Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets”. *Marketing Science*, 15: 1-20.

[9] Goeree, M. (2008). “Limited Information and Advertising in the U.S. Personal Computer Industry”. *Econometrica*, 76:1017-74.

[10] Hong, H., and Shum, M. (2006). “Using Price Distributions to Estimate Search Costs”. *RAND Journal of Economics*, 37: 257-275.

[11] Israel,M. (2005). “Services as Experience Goods: An Empirical Examination of Consumer Learning in Automobile Insurance”. *American Economic Review*, 95: 1444-63.

[12] McCall, J. (1970). “Economics of Information and Job Search”. *Quarterly Journal of Economics*, 84: 113-126.

[13] Mehta, N., Rajiv,S., Srinivasan, K. (2003). “Price Uncertainty and Consumer Search: a Structural Model of Consideration Set Formation”. *Marketing Science*, 22: 58-84.

[14] Nelson.P. (1971). “Information and Consumer Behavior”. *Journal of Political Economy*, 78: 311-29.

[15] Stigler, G. (1961). “The Economics of Information”. *Journal of Political Economy*, 69: 213-225.

Table 2.1: Transactions and Visits by Movies DVD Store

	Transaction		All Visits		Relevant Visits(7 days)		Relevant Visits(1 day)	
Website	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Amazon.com	1,174	44.74	50,921	30.63	4,319	37.27	1,884	40.52
Columbiahouse.com	776	29.57	9,216	5.54	1,726	14.90	1,009	21.70
Barnesandnoble.com	73	2.78	3,567	2.15	275	2.37	137	2.95
Bestbuy.com	64	2.44	6,359	3.82	457	3.94	164	3.53
Buy.com	3	0.11	2,047	1.23	117	1.01	26	0.56
Cduniverse.com	14	0.53	775	0.47	59	0.50	22	0.47
Christianbook.com	16	0.61	914	0.55	64	0.55	42	0.90
Circuitcity.com	45	1.71	3,952	2.38	243	2.10	102	2.19
Dvdempire.com	11	0.42	747	0.45	54	0.47	22	0.47
Go.com	229	8.73	46,382	27.90	1,917	16.54	535	11.51
Overstock.com	90	3.43	10,734	6.46	626	5.40	219	4.71
Target.com	26	0.99	14,284	8.59	676	5.83	174	3.74
Walmart.com	103	3.93	16,362	9.84	1054	9.10	313	6.73
Total	2,624		166,260		11587		4,649	

Table 2.2: Summary Statistics of Movie DVD Transactions

	Mean	Std.Dev.	10%	median	90%
Transactions of consumer(#)	3.47	4.18	1	2	7
Web retailers purchased from(#)	1.32	0.57	1	1	2
Quantity of each transaction	1.02	0.29	1	1	1
Transaction expenditure (US dollars)	14.51	7.63	5.98	14.48	21.95

Table 2.3: Summary Statistics of Movie DVD Store Visits

	Mean	Std.Dev.
Duration of each website visit (in minutes)		
Visits not within 7 days of transaction	9.3	22.5
Visits within 7 days, exclude transactions	13.2	46.8
Visits with transactions	28.2	40.8
Total duration of each 7-day search period (in minutes)		
On web retailers without transaction	27.2	184.6
On web retailers with transaction	47.6	74.2
Number of visits within each 7-day search period		
Number of visits (repeated sites included)	4.4	6.5
Number of aware stores	8.0	2.1
Number of different web stores visited	2.4	1.5

Table 2.4: Distribution of Visited Websites by Aware Store Number

		Number of Visited Websites						
Aware Store#	Total	1	2	3	4	5	6	7 or More
2	9	9	0	0	0	0	0	0
3	29	26	3	0	0	0	0	0
4	120	96	18	5	1	0	0	0
5	154	118	27	8	1	0	0	0
6	250	163	62	24	1	0	0	0
7	345	193	89	44	12	5	2	0
8	563	275	156	88	30	9	4	1
9	393	154	108	70	39	16	3	3
10	376	132	115	59	44	19	6	1
11	288	78	74	61	40	23	8	4
12	64	9	9	20	13	8	4	1
13	33	7	12	7	5	1	0	1
All	2624							

Table 2.5: Summary Statistics of Movie DVD Price

	Mean	Std.Dev.
Total number of transactions	2624	
Number of different web stores with matched price	4.2	2.4
Number of different web stores visited	2.4	1.5
Number of different visited web stores with matched price	1.4	0.8
Number of transactions with all visited web stores price matched	1652	
Number of transactions not from minimum price web store	234	

Table 2.6: Descripitive Statistics of Best Selling Movie DVDs on Amazon

	Price($) on Amazon					
Best Selling Movie DVD	Obs.	Mean	Std.Dev.	10%	median	90%
Harry Potter 4	72	19.9	0.8	19.95	19.95	19.95
Over The Hedge	29	16.1	1.4	15.73	15.73	19.95
Happy Feet	32	18.6	0.7	18.26	18.26	19.95
Shrek 3	57	19.6	1.2	19.95	19.95	19.95
Pirates of The Caribbean 2	95	23.5	1.8	24	24	24
Harry Potter 5	60	21.5	1.8	21.99	21.99	21.99
Cars	81	15.3	1.2	14.99	14.99	14.99
King Kong	28	16.3	2.3	15.99	15.99	15.99
The Little Mermaid	58	13.3	2.2	12.65	12.65	12.65
Pirates of The Caribbean 3	59	22.8	1.07	22.99	22.99	22.99

Table 2.7: Descripitive Statistics of Best Selling Movie DVDs on Columbiahouse

	Price($) on Columbiahouse					
Best Selling Movie DVD	Obs.	Mean	Std.Dev.	10%	median	90%
Harry Potter 4	147	22.3	2.5	22.95	22.95	22.95
Over The Hedge	165	13.4	1.7	12.99	12.99	12.99
Happy Feet	151	19.5	2.2	19.95	19.95	19.95
Shrek 3	96	15.4	1.4	14.99	14.99	15.81
Pirates of The Caribbean 2	0					
Harry Potter 5	64	10.8	4.0	9.18	9.18	19.95
Cars	0					
King Kong	94	18.5	2.5	18.96	18.96	18.96
The Little Mermaid	0					
Pirates of The Caribbean 3	0					

Table 2.8: Descripitive Statistics of Best Selling Movie DVDs on Others

	Price($) on Others					
Best Selling Movie DVD	Obs.	Mean	Std.Dev.	10%	median	90%
Harry Potter 4	10	22.4	1.3	19.95	22.95	22.95
Over The Hedge	24	13.4	1.5	12.99	12.99	12.99
Happy Feet	19	19.95	0	19.95	19.95	19.95
Shrek 3	40	15.0	0.2	14.99	14.99	14.99
Pirates of The Caribbean 2	82	19.9	0.9	19.99	19.99	19.99
Harry Potter 5	30	10.2	2.9	9.18	9.18	13.74
Cars	70	15.4	1.8	14.91	14.91	16.43
King Kong	18	18.5	2.6	12.98	18.96	18.96
The Little Mermaid	78	19.91	0.7	19.99	19.99	19.99
Pirates of The Caribbean 3	75	15.2	1.2	14.99	14.99	14.99

Table 2.9: Duration as Search cost on Household Characteristics

Total Time Spent on Searching	Coef.	Std.Dev.
Number of retailers visited	40.06	(1.89)***
Quantity of purchase	1.10	(0.71)
Bought from last purchase store	-11.37	(4.79)***
Transaction periods	3.30	(0.49)***
Household size	7.07	(1.95)***
Broadband connection	-8.97	(5.54)
Children present in household	-6.52	(5.50)
Age		
21-24	34.7	(35.4)
25-29	26.56	(32.23)
30-34	29.41	(31.12)
35-39	23.30	(31.48)
40-44	29.44	(31.12)
45-49	18.50	(31.36)
50-54	32.33	(31.30)
55-59	15.47	(31.52)
60-64	26.64	(31.66)
65 and over	73.89	(31.36)*
Household income		
$15,000-$24,999	-21.46	(10.88)
$25,000-$34,999	-10.45	(9.93)
$35,000-$49,999	14.63	(8.75)*
$50,000-$74,999	-10.67	(8.42)
$75,000-$99,999	-7.95	(9.32)
More than$100,000	-3.90	(9.29)
Education		
High school	-51.24	(119.46)
College dropout	-104.63	(119.45)
Associate degree	-95.55	(121.91)
Bachelor degree	-93.20	(119.69)
Graduate degree	-88.94	(119.63)
Other	-94.26	(119.28)
Race		
Black	9.91	(12.56)
Asian	-1.39	(24.08)
Other	57.87	(23.6)**
Region of residence		
North Central	3.92	(6.32)
South	-2.55	(5.74)
West	-1.23	(6.37)
Constant	54.96	(123.41)
R-squared	10.3	
Number of individuals	6,472	

∗ significant at 10%; ∗∗ significant at 5%; ∗ ∗ ∗ significant at 1%.

Table 2.10: Number of Visited Retailers on Household Characteristics

Number of retailers visited	Coef.	Std.Dev.
Quantity of purchase	0.01	(0.005)***
Bought from last purchase store	-0.25	(0.003)***
Transaction periods	0.02	(0.003)***
Household size	0.05	(0.013)***
Broadband connection	-0.30	(0.036)***
Children present in household	-0.02	(0.036)
Age		
21-24	-1.06	(0.232)***
25-29	-1.17	(0.212)***
30-34	-1.09	(0.205)***
35-39	-1.07	(0.206)***
40-44	-1.14	(0.204)***
45-49	-1.11	(0.206)***
50-54	-1.18	(0.206)***
55-59	-1.08	(0.207)***
60-64	-1.15	(0.208)***
65 and over	-1.03	(0.206)***
Household income		
$15,000-$24,999	-0.01	(0.072)
$25,000-$34,999	0.02	(0.065)
$35,000-$49,999	-0.01	(0.058)
$50,000-$74,999	-0.04	(0.055)
$75,000-$99,999	-0.06	(0.061)
More than$100,000	-0.08	(0.061)
Education		
High school	0.65	(0.787)
College dropout	0.70	(0.787)
Associate degree	0.79	(0.802)
Bachelor degree	0.72	(0.788)
Graduate degree	0.70	(0.788)
Other	0.49	(0.786)
Race		
Black	-0.06	(0.08)
Asian	0.11	(0.159)
Other	-0.42	(0.156)***
Region of residence		
North Central	-0.11	(0.042)***
South	-0.11	(0.038)***
West	-0.28	(0.042)***
Constant	2.34	(0.812)***
R-squared	0.05	
Number of individuals	6,472	

∗ significant at 10%; ∗∗ significant at 5%; ∗ ∗ ∗ significant at 1%.

Table 2.11: Estimation Result

Variable	Coef.	Std.Dev.
Price		
Income less than $\$35,000$	-0.285	(0.019)***
Income $\$35,000$-$\$50,000$	-0.269	(0.019)***
Income more than $\$50,000$	-0.108	(0.019)***
Search cost		
Constant	0.660	(0.057)***
household size	0.056	(0.025)**
Broadband connection	-0.124	(0.045)**
Age 30 and over	2.160	(0.030)***
Store fixed effect		
Amazon	1.494	(0.029)***
Columbiahouse	0.984	(0.021)***
Broadband connection		
Amazon	-0.337	(0.021)***
Columbiahouse	1.057	(0.051)***
Age 30 and over		
Amazon	1.096	(0.060)***
Columbiahouse	-0.356	(0.180)***
Household income $\geq \$50,000$		
Amazon	1.834	(0.041)***
Columbiahouse	0.567	(0.020)***
Household size		
Amazon	-0.722	(0.032)***
Columbiahouse	-0.216	(0.059)***
Scale (σ_δ)	7.029	(0.051)***
Number of observations	2624	

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Chapter 3
A Structural Model of Search with Learning: Application to the Online Movie DVD Retail Market

3.1 Introduction

Stigler (1961) notices that in a market with price dispersion consumers may not have full information of products' prices and have to acquire the price information by costly search in these markets. A strand of the literature following explains many facts in incomplete competitive markets including price dispersion in product markets and frictional unemployment in labor markets by using search models. Generally two types of search are considered. The first is fixed sample size search where consumers search over a fixed number of sellers and choose to buy from the one giving her the highest utility.[1] The alternative is sequential search, which is initially proposed by McCall(1977) to model the job search behavior. The sequential search decision is characterized in terms of the reservation price in the framework of consumption good search. The consumer's optimal strategy is to buy at any price lower than the reservation price and continue to search with a price higher than it.

Consumers not only lack full information about the prices of goods, their information is probably even poorer about the quality of varies brands. Nelson in the seminar paper (1971) points out that especially for experienced goods, a consumer may not learn the quality of a brand unless he purchases it. Hence, the information of quality can be acquired through consumer learning. Erdem and Keane(1996) firstly empirically study consumer learning in the case of laundry detergent. Ackerberg(2003) also considers learning in frequent purchased good market(yogurt). Their findings show that consumers have accurate prior information and fast speed of learning. Israel (2005) examines the learning about the services of automobile insurance as experience goods. However, he has a contrast finding that consumer has poor prior information and the learning speed in this market is limited by the low claim probability. These empirical studies show that in different markets, the information acquisition processes

[1] See also Burdett and Judd (1983)

are quite different.

As the rapid expansion of the online retail market, a large literature has tried to explain the consumer behavior in this market. Many empirical works show that the price dispersion across firms persists as in other traditional retail markets.[2] This motivates the empirical search applications to online markets. Hong and Shum(2006) rationalize the prices setting by online textbook stores with sequential search model as well as a non-sequential search model. They find the magnitude of search cost in a sequential model is larger than that in a fixed sample size search model (or a nonsequential search model). De los Santos, Hortacsu, and Wildenbeest (2012) empirically test the two classical search models using observed online book purchasers' search behavior in ComScore Web-Behavior Panel dataset. They find that a fixed sample size search model provides a more accurate description in this particular market of online bookstores. Bronnenberg, Kim and Mela(2014) find that consumers search for more attributes rather than price, and consumers tend to purchase from the later searched websites. These facts are consistent with sequential search behavior.

In online retail market, before visiting the webpage of a firm, consumers are usually not certain about the firm's specific quality of the product such as the rating of the product and the usage condition and also the quality of services (e.g. shipping services and other customer services) from customer reviews provided by other consumers. This is especially true for experiential products such as books, movies, and music since these are consumed mainly for the pleasure they provide (Senecal and Nantel 2004). Such information about reviews and services is now widely available on the websites of firms for all their products. Recent research also provides evidence that the qualities of products have an impact on product sales basing on aggregate-level analysis. For example, Chevalier and Mayzlin (2006) shows a positive relationship between book sales and consumer book ratings. Liu(2006) finds that the volume of consumer reviews is positively related to its box office revenue. These studies all support the basic premise that the information of quality acquired through learning by searching on the firms' websites has an impact on consumer's final purchase decision.

As in the online retail market, consumers do have limited information on both selling prices and quality of firms, it is appropriate to consider both learning and search behavior of consumers. In this paper, I develop a model with two stages of decisions by consumers: a dynamic discrete choice decision on choosing search sets and a static purchase decision on purchasing from the firms in their search set in each period. The basic idea is that consumers are imperfectly informed and uncertain about the true quality of a firm. To cope with such uncertainty, consumers may learn based on their past experiences and the new information

[2] See Brynjolfsson and Smith(2000), Clay and Wolff(2001).

through searching on the websites of the firm.[3] They then update their beliefs of firms' qualities in a Bayesian fashion. They also learn the exact prices of firms by visiting the firms. When determining which firm to purchase from, the consumers use their information on both firms' qualities and prices.

This paper utilizes novel and detailed data on the web browsing and purchasing behavior of a large panel of consumers from the ComScore Web Behavior Database. The dataset is novel in that it provides detailed information tracking of all consumer searches prior to each transaction. In the empirical analysis, I focus on the online movie DVD retail market. Several features of this market would help us identify the learning behavior of the consumers. A panel of consumers can be constructed since the consumers in this sample generally have multiple purchases in movie DVD category. According to the pattern in the dataset, individuals in the market adjust their choice sets and purchase decisions over time. In my data, 13.8% of the consumers do not buy from the lowest priced store in their choice set. Consumers are also more likely to search on certain stores. Both facts indicate quality differences between movie DVD retailers other than price.

I apply the proposed model to a panel data set constructed from ComScore consumer level search and purchase behavior data of the online movie DVD retail market. The empirical analysis leads to the following findings. First, I find in this market, search cost is sizable. Second, the hypothesis of perfect information can be easily rejected and there exists consumer heterogeneity in firms' quality beliefs. Finally, consumers are learning about the firms quality very slowly which plays an important role in explaining the frequent search behavior change of consumers appearing in the data.

Next I conduct three counterfactual experiments. By simulating the market without search costs based on the parameter estimates of the structural model, I find that search costs have a huge impact on the market shares of firms. The industry leader Amazon would have 12% less share absent search costs. In addition, I compare the effects of different pricing strategies of firms. The market share of Amazon shrinks by 2% if Amazon matches the price of ColumbiaHouse which has the second largest market share. By simulating the market with one search learning behavior consumers, I find that market becomes more concentrate such that the leading firm Amazon market share increases from 35.16% to 68.07%.

My paper relates to several papers which estimate search costs with consumer learning. Mehta, Rajiv, and Srinivasan (2003) construct a structural search model with consumer learning. In their model, consumers are uncertain about prevailing prices of the brands although they are aware of the price distributions. Consumer makes a trade-off between higher

[3] Consumers usually know more about a retailer's quality such as the available consumer services from browsing its website.

expected utility from a more extensive price search and the cost of search. They empirically study the scanner data of liquid detergent where search cost comes from the search over different brands in the same retail store. De los Santos, Hortacsu, and Wildenbeest (2012) relaxes the assumption that consumer knows the price distribution of the firms, and propose a search model with a Bayesian learning process on the price distribution. My paper is also related to a literature that estimates Bayesian learning models. Ackerberg(2003) studies the effect of advertising in an experienced good market which is subject to consumer learning. Crawford and Shum(2005), Narayanan and Manchanda (2009) model physician of the quality of drugs. Section 2 introduces the data used in this paper. Section 3 describes the model of consumer search with learning behavior. Section 4 presents my estimation methods and its results. In Section 5, I perform counterfactual experiments on search cost structure and pricing strategies and Section 6 concludes.

3.2 Data

This paper utilizes a dataset from the ComScore Web Behavior Database including detailed online searching and transaction information as well as the panelists' demographic information. The dataset used in the empirical analysis is constructed from two parts. The first part is the transaction data which includes price, quantity, product category and name of each product purchased as well as the demographics of consumers. The second part is the browsing/searching behavior data and includes information on website browsing history including websites visited (including referral websites), date, time, duration, and pages viewed of each visit. Specifically, each consumer is identified as one computer id in the dataset. Thus, it can capture the browsing and transaction history of the whole family. Demographic information of these families includes household head education level, household income, household size, racial background, census region and so on. Comparing to the Current Population Survey, the ComScore sample are representative of online buyers in the United States.[4]

3.2.1 Summary Statistics

I focus on the online movie DVD retail market in this empirical analysis. The benefit is that, in this category, the good with the same product name from different retailers can be treated as homogeneous good.(e.g. Spider Man 3 DVD from Amazon and Columbiahouse) Since the

[4] See De los Santos (2008) for a more detailed decription of the data, and De los Santos, Hortacsu, and Wildenbeest (2012) for discussion about the representativeness of the dataset.

sample of users is randomly chosen in each year, I will examine the sample of users appearing in both year 2006 and year 2007 to study the consumer's search behavior overtime.[5] Each transaction observation records the detail web browsing history of a single purchase possibly multiple units of the same products. According to the pattern in the data, in each transaction session, consumer can have several transaction of different products(e.g. Spider Man 3 DVD and Spider Man 2 DVD). Different products are recorded in different observations. In the browsing/searching behavior data, I can only identify the website name but not the product category if there is no transaction happened in this search session. Consumer may browse a website just to surf the Internet or to look for products of different categories if a website is not specialized in a particular category. I exclude observations that could not be identified as online movies and videos stores such as web portal yahoo.com. According to the dataset, a total number of 757 users made at least one transaction from 13 different websites with a total number of 2,624 transaction sessions in this category.

From the above argument, the number of visits of each website is an upper bound of the number of visits related to a particular transaction in movie DVD category. Following the discussion on related search behavior in De los Santos, Hortacsu, and Wildenbeest (2012), I define relevant searches as the browsing history (on relevant websites) up to seven days before a transaction, if no other transaction has occurred within these seven days. Otherwise, a search history can contain less than seven days is up to the most recent transaction in the past.

Table 1 and Table 2 summarize the transaction and search behavior of average consumer from this sample. On average, each consumer has made 3.5 transactions from 1.3 different movie online retailers. Each transaction on average costs 14.51$. Each web retailer visit on average takes 9.7 minutes for visits happened without the 7-day search period, 13.2 minutes for visits without transactions happened within the 7-day search period, and 28.2 minutes for visits with transactions. This may be due to the fact that choosing products and checking out incur extra time on visits with transactions. The second panel of Table 2 shows that for each 7-day search period, 63.6% of the time is spent on visiting the websites of the stores with realized transactions by an average consumer. The third panel summarizes the number of visits during each 7-day search period. An average online shopper visited 4.4 websites (including repeated visits to the same online retailer). Even though on average a consumer is aware of 8 among the 13 online retailers, but only 2.4 of them are visited each time.[6]

In the dataset, the price of a product is available only if it appears in a transaction. To

[5] This treatment gives me a longer panel which is good to check the behavior pattern overtime. However, this limits my sample with fewer observations.

[6] Given the large number of online retailers relative to the small number of stores actually visited, I assume that the consumer is aware of a given retailer if she has ever visited the store.

recover missing prices for the other visited stores of a product, I will follow the method proposed by De los Santos, Hortacsu, and Wildenbeest (2012) to use the most recent transaction prices of the same products on those stores. The relative homogeneity among products in movie DVD market allow me to extrapolate a large portion of the price data. Table 3 shows the summary statistics of the matched price. In the average 2.4 visited stores, the price of the same product can be found in 1.4 of those stores. All the prices of the visited websites can be matched in 1,652 (63%) of the total 2,624 transactions and visit periods.[7]

3.2.2 Empirical Evidence

This subsection discusses some empirical evidence regarding consumers' searching and learning behavior in the online movie DVD retail market. According to the pattern in the data, a consumer sometimes chooses to buy from a retailer with a higher price even she visits at least one other retailer with a lower price. (234 out of the 1,652 transaction periods with all visited retailers' price matched) This fact shows that consumer makes her purchase decision on both price and other retailer's specific features or its quality.

I also examine the consumer's formation of search set. Specifically, I consider the determinants of the number of newly searched online retailers. Newly searched online retailers are the stores visited in this 7-day search period but not in the last one. Consumer is expected to expand their search set and search more new retailers if she had a not so good experience last time. I use two different measures for experience: the indicator of whether the consumer purchased from the store of last purchase this period and the indicator of whether consumer visited the store of last purchase this period. I also include consumer characteristics such as connection type and household size. Table 4 displays the regression results of the number of newly searched online retailers on these two purchase experience measure in specification (A) and (C). In specification (B), I use consumer fixed effect instead of demographic characteristics. These results show that a good experience in last purchase will decrease the number of new websites searched which is consistent with the expectation that the consumers make the choice set decisions based on their experience. Consumer will search more new websites if they have a fast connecting speed since it decrease their unit time cost. It is also not surprising to find that a family also tends to search more new websites if their family size is large.

I also check why consumer decides not to search some retailers that they have searched in previous consumption periods. Table 5 summarizes the regression results of the number of the retailers dropped from the previous choice set on the two experience measure. These

[7] The number of periods when all visited retailers' price is matched will increase to 2,325 by dividing all 13 websites into 3 groups (Amazon, Columbiahouse, and other websites.)

results show that a good experience will also decrease the number of retailers dropped from previous choice set. This also confirms the choice set decision of consumers are influenced by experience.

3.3 Model

In this section, I develop a structural model that captures consumer's search behavior with learning. Firstly, in each period with the knowledge of quality beliefs and price distribution, consumers choose a subset (search set or choice set) of all the firms to search in order to know their exact prices and quality signals. I assume that the consumers use a non-sequential search strategy which means they will visit all the firms in the search set not in a particular order. Secondly, consumer makes a purchase among the searched firms once they have the information on prices and quality signals. Consumers are assumed to make the search decisions to maximize their expected discounted sum of future utilities since the search decisions lead to updates on quality beliefs which has impact on the future search decisions. However, the purchase decisions are static decisions since there is no learning happening in this stage.

Consider an online retail market with J different firms selling a good differentiated by firm specific qualities. In each period t, consumer i knows the price distribution $F_j(\cdot)$ of firm j. Consumer can only know the exact price p_{ijt} of each firm through a costly search on it. The exact prices are allowed to vary across both consumers and time. The search cost per firm for consumer i is assumed to be c_i. Consumer i also holds prior beliefs about the quality of each firm which is updated only if the consumer search the good on its website. After the search, consumer receives a quality signal which is used to form the posterior quality beliefs. The quality refers to the uncertain firm's attributes such as its services, and the quality signal refers to the gathered information from searching the website (i.e. customer reviews).

I allow for a more general learning process similar to Bayesian learning Models such as Erdem and Keane (1996), and Ackerberg(2003). The received quality signal μ_{ijt} is given by:

$$\mu_{ijt} = \mu_{ij} + \nu_{ijt}, \text{ where } \nu_{ijt} \sim iid\ N(0, \sigma_v^2) \tag{1}$$

The quality signal μ_{ijt} is observed by the consumers after search. However, its two components μ_{ij} and ν_{ijt} are not observed separately. Consumers are allowed to have different, but persistent mean quality μ_{ij} of firm j over time. ν_{ijt} are confounding part of the quality signal that can not be distinguished from this quality mean. Variance in ν_{ijt} may be the outcome of variation in firm's available services, e.g. payment services, the product related services, such as return policy or even the moods at time of browsing on firm's webpages. [8] Hence, it is beneficial for consumers to use the information from the observed μ_{ijt} to learn about μ_{ij}.In the degenerate case where $\sigma_v^2 = 0$, the learning process described above is a one-search learning process since μ_{ij} is learned after one search. In the non-degenerate case, the sequence of μ_{ijt} only provide information about μ_{ij}.

To model the Bayesian learning process, I assume the distribution of mean quality μ_{ij} of firm j across the population is given by:

$$\mu_{ij} \sim N(\mu_j, \sigma_j^2) \tag{2}$$

and where the mean μ_j of μ_{ij} across population is a parameter that can be interpreted as the overall quality of firm j. A higher μ_j indicates that consumers like firm j more. The initial $(t = 0)$ prior on μ_{ij} is assumed to be:

$$\mu_{ij} \sim N(0, \sigma_j^2 + \sigma_k^2) \tag{3}$$

There are two components of the consumer's prior variance on μ_{ij}: σ_k^2 captures the overall uncertainty about firm j's quality μ_j, which is assumed to be identical across firms. σ_j^2 comes from the individual deviation in μ_{ij} around its mean μ_j. As we can see, consumer beliefs are decided by the parameters μ_j,σ_j^2, and σ_k^2 (j=1,$\cdots$,J).

Based on Bayes' rule, consumer i's posterior on firm j's quality after observing a sequence of quality signals $\{\mu_{ij1}, \cdots, \mu_{ijT_{ijt}}\}$, where T_{ijt} is the number of search on firm j consumer i has had up to period t, is:

$$\mu_{ijt} \sim N(m_{ijt}, \Sigma_{ijt}^2) \tag{4}$$

where:

$$\begin{aligned} \Sigma_{ijt}^2 &= \Big((\Sigma_{j0}^2)^{-1} + T_{ijt}(\sigma_\nu^2)^{-1}\Big)^{-1}, \\ m_{ijt} &= \Sigma_{ijt}^2\Big((\Sigma_{j0}^2)^{-1} m_{ij0} + (\sigma_\nu^2)^{-1} \sum_{k=1}^{T_{ijt}} \mu_{ijT}\Big) \end{aligned} \tag{5}$$

These posterior means and variances contains all the consumer's information on μ_{ij}. Hence, the current posterior (m_{ijt}, Σ_{j0}^2) is sufficient to define future posteriors under the i.i.d. assumption of (1).

According to the pattern in my data, I allow for heterogeneity in consumer preferences.

[8] In all the cases, ν_{ijt} is not distinguishable from μ_{ij}. For example, the consumer may read generally positive comments from previous consumers while searching and enjoy the product more. The exact enjoyment from the firm can come from such firm specific characteristics but also simply the good mood.

In period t, consumer i's indirect utility u_{ijt} of purchasing a product at at jth retailer is given by,

$$u_{ijt} = \mu_{ijt} + X_i\beta_j - \alpha P_{ijt} \tag{6}$$

The utility is decided by the quality signal μ_{ijt}, time invariant consumer characteristics X_i, and the price P_{ijt}. The quality signal and price are observed when consumer i searches the firm j. Note these variables vary over both time and consumers since firms change prices over time and consumers make purchases at different time.

In each time period t, consumer i knows her posterior belief of each firm's quality μ_{ijt}. The posterior beliefs are determined by the mean and variance $s_{it} = (m_{i1t}, \cdots, m_{iJt}, \Sigma_{i1t}, \cdots, \Sigma_{iJt})$ of the Gaussian posterior distribution. Following the literature, I also assume a search set specific utility shock ϵ_{ist} for each choice set s at period t, which follows the type I extreme value distribution. Denotes the vector of all these shocks by $\epsilon_{it} = (\epsilon_{it}(1), \cdots, \epsilon_{it}(S))$, where $S = 2^J - 1$ is the total number of available search sets. The information set I_{it} of consumer i at period t consists of (s_{it}, ϵ_{it}).

Consumers are assumed to use a non sequential search strategy. It means that consumer i chooses to search a subset $\mathbf{d}_{it}$ of the total J firms not in particular order. I also assume that consumers encounter a fix cost c for searching each firm. In period t, the expected utility of consumer i to search $\mathbf{d}_{it}$ with her information set I_{it} , denoted by $v(\mathbf{d}_{it}, I_{it})$, can be written as,

$$v(\mathbf{d}_{it}, I_{it}) = \mathbb{E}_{\mu_{it}, p_{it}}\Big[\max_{j\in\mathbf{d}_{it}}\{u_{ijt}\}\Big] - c\|\mathbf{d}_{it}\| + \epsilon_{it}(\mathbf{d}_{it}) \tag{7}$$

where $\|\mathbf{d}_{it}\|$ is the number of firms in subset $\mathbf{d}_{it}$ and $\epsilon_{it}(\mathbf{d}_{it})$ is the search set specific utility shock. To suppress the notation, I use μ_{it} and p_{it} to denote the vector of all quality signals and the vector of all prices, respectively. The expectation is over the quality beliefs and prices, as consumer i does not know μ_{iit} and price p_{ijt} when she makes the search decision. Following Mehta, Rajiv and Srinivasan (2003), I assume that prices follow a type I extreme value distribution with firm-specific location parameter γ_j and common scale parameter σ. Under this distribution assumption, $\max_{j\in\mathbf{d}_{it}}\{u_{ijt}\}$ has a closed-form expression as,

$$\max_{j\in\mathbf{d}_{it}}\{u_{ijt}\} = \alpha_i\sigma \log\Big(\sum_{j\in\mathbf{d}_{it}} \exp\big[\frac{\mu_{ijt} + X_i\beta_j - \alpha_i\gamma_j}{\alpha_i\sigma}\big]\Big) \tag{8}$$

Consumer i is assumed to be forward looking. Her optimal search set decision is to choose a sequence $\{\mathbf{d}_{it}(I_{it})\}$ to maximize her expected long-run utility at period t:

$$\max_{\{\mathbf{d}_{it}, \mathbf{d}_{i,t+1}(I_{it+1}), \cdots\}} \mathbb{E}\Big[\sum_{\tau=t}^{\infty} \beta^{\tau-t} v(\mathbf{d}_{i\tau}, I_{i\tau})\Big] \tag{9}$$

where β is the discount factor per period and the per period utility for searching is given by equation (7). Since the transition probability is Markovian as the evolution of state variables s_{it} only depends on the state variable s_{it-1} one period ahead. The optimal decision rule for infinite-horizon decision problem is time invariant. Therefore, I can formulate the sequential maximization problem as a solution to the following Bellman equation without a time index t:

$$V(s,\epsilon) = \max_{\mathbf{d}} \left\{ v(\mathbf{d}, I) + \beta \int_{s'} \int_{\epsilon'} V(s', \epsilon' | x, \epsilon, \mathbf{d}) p(s', \epsilon' | s, \epsilon, \mathbf{d}) \mathrm{d}s' \mathrm{d}\epsilon') \right\} \tag{10}$$

where (x', ϵ') denote the next-period state variables. In the Bellman equation, the state variables (s, ϵ) contains current posterior distribution of quality for each firm and preference shocks for all choice sets. $V(\cdot)$ is the expected discounted sum of utilities. Given the assumption that preference shocks ϵ_{it} are type I extreme value distributed, the Markov transition probability of state variables (s_{it}, ϵ_{it}) satisfy the conditional independence (CI) condition,[9] which can be decomposed as

$$p(s_{it+1}, \epsilon_{it+1} | s_{it}, \epsilon_{it}, \mathbf{d}_{it}) = p_1(\epsilon_{it+1}) p_2(s_{it+1} | s_{it}, \mathbf{d}_{it}) \tag{11}$$

The ex-ante value function U(s)=$E_{\epsilon}[V(s,\epsilon)|s]$, where the expectation is over the random multivariate type I extreme value distributed shocks ϵ, is the fixed point of

$$U(s) = \ln \Big(\sum_{\mathbf{d}} \exp(u(\mathbf{d}, s) + \beta E[U(s')|s, \mathbf{d}]) \Big) + \gamma \tag{12}$$

where γ is Euler's constant and $u(\mathbf{d}, s) = v(\mathbf{d}, I) - \epsilon(\mathbf{d})$. The conditional choice probability of search set decision has the following multinomial logit formula:

$$P(\mathbf{d}|s_{it}) = \frac{\exp(u(\mathbf{d}, s_{it}) + \beta E[U(s_{it+1})|s_{it}, \mathbf{d}])}{\sum_{\mathbf{d'}} \exp(u(\mathbf{d'}, s_{it}) + \beta E[U(s_{it+1})|s_{it}, \mathbf{d'}])} \tag{13}$$

where $P(d|s_{it})$ is the probability of choosing a search set decision $\mathbf{d}$ given the state variables s_{it}.

In the second stage, consumer i knows prices and quality signals of the firms in the search set $\mathbf{d}_{it}$. Therefore, consumer i knows her utility u_{ijt} for the product from each searched firm j, and choose the firm j providing the highest utility which is $\arg\max_{j \in \mathbf{d}_{it}} u_{ijt}$. The purchase probability $P_{ij|\mathbf{d}_{it}}$ of choosing firm j is stated as,

$$P_{ij|\mathbf{d}_{it}} = \Pr(u_{ijt} > u_{ij't}, \forall j' \neq j \in \mathbf{d}_{it}) \tag{14}$$

[9] The definition of conditional independence is stated in Rust(1987).

To get the probability $P_{ij\mathbf{d}_{it}t}(s_{it})$ of observing a consumer i with state variable s_{it} selecting a search set $\mathbf{d}_{it}$ and purchasing product j, the probability is the product of the probabilities in equations (13) and (14),

$$P_{ij\mathbf{d}_{it}t}(s_{it}) = P(\mathbf{d}|s_{it})P_{ij|\mathbf{d}_{it}} \tag{15}$$

3.4 Estimation

I now move to the estimation of my model using the data, beginning with a discussion of my empirical specification. I then discuss how my data can identify the parameters of the learning process intuitively since my model is complicated and highly non-linear. At the end of the section, I present estimates of two models. The first model is the full dynamic model in which consumers are forward-looking and maximize their expected discounted sum of utilities. The second model assumes consumers' behavior is myopic, that is consumers learn and update according to the full model but maximize only their current period utilities but not the expected discounted sum of utilities.

I model consumer i's search and purchase decision period as each of her observed transaction through the 2 years of data. Because of the price matching issues discussed in the data section, I restrict the number of firms to be 3. The three firms are Amazon, Columbia-House and the other firms. To avoid the problem of curse of dimensionality in the dynamic estimation, I exclude consumer characteristics in the utility function.

It is important to discuss how the parameters in the model are identified by the data. Identification comes from how consumer's search behavior and purchase behavior change over time, particularly after the information acquisition is completed. If there is no learning, search patterns conditional on covariates such as price and consumer characteristics will be constant over time. With learning, search experiences will change a consumer's search and purchasing patterns. Eventually, consumer learns about each firm's experienced quality.

The variance σ_j^2 of the quality μ_{ij} is identified by the difference of "preinformation" heterogeneity and the "postinfomation" heterogeneity across consumers. The per-period variance σ_v^2 in quality is identified by the number of periods for consumer i to learn μ_{ijt}. The variation of consumer's characteristic identifies β_j. The price distribution $F_j(p)$ is identified from the price distribution of store j. The quality taste parameter α is identified from the variation of the mean qualities and the variation of the price. The search coefficient c is identified from the "postinformation" search set decision.

The parameters of the type I extreme value price distributions are estimated using prices of the DVDs in the sample that are sold on all three firms (namely Amazon, Columbiahouse and others) in the full data set over the two years. These transactions may be from the consumers out of my sample. The fitted location parameters of price distributions are 11.647 for Amazon, 9.310 for Columbiahouse and 12.320 for other DVD firms and the scale parameter is 5.841.

Unlike Rust(1987), the econometrician does not observe the quality posteriors s_{it} but the corresponding choice set decision $\mathbf{d_it}$ in each period and for each consumer. However, given the structural parameter vector θ and the choice set decision,the econometricians know the distribution of s_{it}. To derive the probability of consumer's observed choice decision in my data, I need to integrate over the unobserved state variables. This leads to the following likelihood function:

$$L(\theta) = \prod_{i=1}^{N} \int \prod_{t=2}^{T_i} \left(\frac{\exp(u(\mathbf{d}_{it}, s_{it}) + \beta E[U(s_{it+1})|s_{it}, \mathbf{d}_{it}])}{\sum_{\mathbf{d}} \exp(u(\mathbf{d}, s_{it}) + \beta E[U(s_{it+1})|s_{it}, \mathbf{d}])}\right) \mathbb{1}(\mathbf{f}_{it}|\mathbf{d}_{it}) p(\mathrm{d}s_{it}|\theta)$$

where $\mathbf{f}_{it}$ is the purchase decision by consumer i in period t. To evaluate the logarithm of the likelihood function $L(\theta)$ for a given θ, one needs to know the value of U(s) which is the fixed point of equation (12). I follow nested fixed-point algorithm proposed by Rust (1987) which consists of an outer loop and an inner loop. I use Newton's method, which is a valid method, to maximize $L(\theta)$ by changing the structural parameters θ in the outloop. For the given θ, I solve for the fixed-point equation (12) to evaluate $L(\theta)$ in the inner loop.

3.4.1 Results

The maximum likelihood estimates of the model are shown in Table 6. The results of the fully dynamic model and also the model with myopic consumers ($\beta = 0$) are in the first column and second column, respectively. In both models, the estimated search cost parameters are highly significant indicates that the existence of search cost in the online movie DVD retail market. The search costs are 1.67\$ and 2.768\$ for full dynamic model and myopic model, respectively, by normalizing the estimates of the search cost parameters by the estimated price scale parameter.[10] The very large significant estimates of σ_ν support that consumers receive very noisy signals so that their learning process is slow. The initial prior variance estimates of σ_k^2 are significantly smaller than the quality variance estimates σ_j^2 support the existence of consumer heterogeneity in the priors of quality beliefs.

Comparing the results of the fully dynamic model with those of the myopic model, the ranking of quality mean of the firms are different. While Amazon has the highest quality mean in the myopic model but ColumbiaHouse has the highest quality mean in the full

[10] Intuitively, the search costs estimates are smaller than the average search cost \$2.67 in the static search model in Chapter 2. As the static search model omits the learning benefit of search behavior so that a larger search cost is needed to rationalize the same search behavior observed in the data in such a static model.

dynamic model. This is likely the result of that Amazon is searched most often in the data. In the myopic model, this fact leads to the quality estimates showing Amazon is the most preferred firm. However, in the fully dynamic model, part of the object of search by consumer is to learn the quality of firm more precisely. Thus, the preference of consumers among firms generated by a true dynamic decision process is likely to be falsely predicted by a myopic model.

In summary, the following are the findings of this research: (1) Search cost is quite sizable even consumers are likely to have less search cost than the traditional movie DVD retail market. (2) Consumers learn from searching at a slow but significant rate in the online movie DVD retail market. (3) Not accounting for consumer learning from searching leads to biased inferences of the preference order of firms.

3.5 Counterfactual Simulations

A significant benefit of adopting a structural modeling approach to understand the search with learning in online DVD retail market is that I am able to estimate the effects of changes on firms' policies (such as pricing strategy) and changes on consumer behavior (such as search costs), on consumer choice and market share.[11] To illustrate the policy implications of this research, I present the following simulations.

3.5.1 Counterfactuals without search costs

The object of this counterfactual experiment is to show how the model can help to predict the market share changes when the search costs change. Firms such as Amazon and Walmart are known to regularly request their consumers to post product reviews. These reviews may be more informative and ease the search process of new consumers. To accomplish this object, I conduct a policy simulation by estimating changes in market share in an extreme case when the search costs absent in the market.

Under the assumption of no search costs, consumer search set decision will be simply all the three available firms. Then consumers purchase from the firm which gives the highest utility. I use 100 simulated consumers per observation and simulate their utilities and purchase decisions with the estimated parameters in the model. Results of this simulation appear in Table 7 column 3. The key finding is that the market share of Amazon decrease from 35.16% to 24.84%. The result is as expected since Amazon has lower quality than ColumbiaHouse.

[11] Without making strong assumptions on firm side behavior which is unclear in this market, I cannot make any satisfying welfare analysis.

Consumers get quality signals from each firm and they generally get lower quality signals from Amazon and they will purchase from the other firms instead of Amazon when they don't have access to the signals of other firms in the environment with search costs.

3.5.2 Counterfactuals with different pricing strategy

In this set of exercises, I allow firms to change their pricing strategies and use the model parameter estimates to predict the market share changes. These results can help firms to evaluate the profitability of different pricing strategies. I consider two different pricing strategies by the leading firm Amazon while holding the other firms' prices, namely Amazon chooses to match the price of ColumbiaHouse or increases the price of all product by 1$.

Within these two policy changes, I assume that consumers can not realize the change of price distribution since the amount of price change is small. Therefore, in each period, consumers search set does not change and still purchase from the highest utility firm in the search set. I still use 100 simulated consumers per observation and the simulation results are shown in Table 7 column 4 and column 5. The market share changes of Amazon is about 2% in either case. The changes are not large if the margin of movie DVD is 10% based on the norm of retail industry. The reason is that consumers still choose from the same search set comparing to the case without pricing strategy changes. This result suggests that its profitable for firms to increase small amounts of their prices before the consumers noticing that the price change is persistent and potentially change their search behavior.

3.5.3 Counterfactuals with different learning behavior

To understand how important consumer learning is in the online retail market, I assume that consumer can learn the quality of firm by one-search learning process. In the one-search learning process, the quality of a firm is learned after one search experience. This is the degenerate case when $\sigma_v^2 = 0$ and one search experience reveal the exact quality of a firm, not just part of the information about the quality like in the nondegenerate case when $\sigma_v^2 > 0$.

Under the one-search learning assumption, I examine how the search behavior of consumer changes and how the market share changes. Since the learning speed is essentially faster than the learning speed in the nondegenerate case, I expect the search frequency to be lower in the experiment. The simulated result shows 62% of the consumers have a smaller search set and only 2% of the consumers have a larger search set, which confirms the intuition that consumers search less when their search is more informative. The market share of the leading firm Amazon increases from 35.16% to 68.07% while the market share of the smallest firm decreases from 34.28% to 0.34%. This result shows that market becomes more concentrate if consumers have an extremely fast learning process.

3.6 Conclusion

In this paper, I estimate a model with two stages of decisions by consumers: a dynamic discrete choice decision on choosing search sets and a static purchase decision on purchasing from the firms in their search set in each period. I also assume learning-by-searching which means each consumer's search decision depends on her endogenous beliefs evolution on quality. The structural estimation results are interesting. From the result of the estimation of the structural model, I find that the search costs are quite sizable even consumers are likely to have smaller search costs than the traditional movie DVD retail market. Consumers learn from searching at a slow but significant rate in the online movie DVD retail market. Not accounting for consumer learning from searching leads to biased inferences of the preference order of firms which proves the importance of counting the consumer learning behavior of consumer in the consumer search model (for example, the search model from Chapter 2).

The counterfactual experiments show that the consumer learning has huge impact on the market structure. If consumers learn the quality of each retailer in one search, the market will be much more concentrated. As the market share of the leading firm Amazon increases from 35.16% to 68.07%.

This research marks the first attempt to incorporate consumer learning on firms' qualities into a search set decision model. However, my findings are not without limitations. There are several limitations in this study suggesting future research opportunities.

First, despite the associate computational burdens, it might be interesting to study how the consumers time invariant characteristics interact with their preference on firms. Second, a more flexible search costs structure may provide more information on the interactions between consumer characteristics and search costs if the computational burdens are acceptable. Finally, although I model learning from consumer's search behavior, the presence of alternative sources of learning, such as advertising, off-line word of mouth and even product reviews from other websites, cannot be ruled out. This issue of incomprehensive data sets is mostly generic to all research on consumer learning.

3.7 References

[1] Ackerberg, D. (2003). “Advertising Learning, and Consumer Choice in Experience Good Markets: An Empirical Examination”. *International Economic Review*, 44: 1007-1040.

[2] Bronnenberg, Bart J., Byeong Jun Kim, and Carl F. Mela. "Zooming In on Choice: How Do Consumers Search for Cameras Online?."

[3] Brynjolfsson, E., and Smith,M. (2000). “Frictionless commerce? A comparison of Internet and conventional retailers”. *ÂŤManagement Science*, 46: 563-ÂŰ85.

[4] Burdett, K., and Judd,K. (1983). “Equilibrium Price Dispersion”. *Econometrica*,51: 955-969.

[5] Clay, K., Krishnan R., and Wolff, E. (2001). “ÂŞPrices and price dispersion on the Web: Evidence from the online book industry”. *ÂŤJournal of Industrial Economics*, 49, 521-ÂŰ39.

[6] De los Santos, B. (2008). “Consumer Search on the Internet”. *NET Institute Working Paper*, #08-15.

[7] De los Santos, B., Hortacsu, and A., Wildenbeest, M. (2012). “Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior”. *American Economic Review*, forthcoming.

[8] Erdem,T., and Keane, P. (1996). “Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets”. *Marketing Science*, 15: 1-20.

[9] Goeree, M. (2008). “Limited Information and Advertising in the U.S. Personal Computer Industry”. *Econometrica*, 76:1017-74.

[10] Hong, H., and Shum, M. (2006). “Using Price Distributions to Estimate Search Costs”. *RAND Journal of Economics*, 37: 257-275.

[11] Israel,M. (2005). “Services as Experience Goods: An Empirical Examination of Consumer Learning in Automobile Insurance”. *American Economic Review*, 95: 1444-63.

[12] McCall, J. (1970). “Economics of Information and Job Search”. *Quarterly Journal of Economics*, 84: 113-126.

[13] Mehta, N., Rajiv,S., Srinivasan, K. (2003). “Price Uncertainty and Consumer Search: a Structural Model of Consideration Set Formation”. *Marketing Science*, 22: 58-84.

[14] Nelson.P. (1971). “Information and Consumer Behavior”. *Journal of Political Economy*, 78: 311-29.

[15] Stigler, G. (1961). “The Economics of Information”. *Journal of Political Economy*, 69: 213-225.

Table 3.1: Summary Statistics of Movie DVD Transactions

	Mean	Std.Dev.	10%	median	90%
Transactions of consumer(#)	3.47	4.18	1	2	7
Web retailers purchased from(#)	1.32	0.57	1	1	2
Quantity of each transaction	1.02	0.29	1	1	1
Transaction expenditure (US dollars)	14.51	7.63	5.98	14.48	21.95

Table 3.2: Summary Statistics of Movie DVD Store Visits

	Mean	Std.Dev.
Duration of each website visit (in minutes)		
Visits not within 7 days of transaction	9.3	22.5
Visits within 7 days, exclude transactions	13.2	46.8
Visits with transactions	28.2	40.8
Total duration of each 7-day search period (in minutes)		
On web retailers without transaction	27.2	184.6
On web retailers with transaction	47.6	74.2
Number of visits within each 7-day search period		
Number of visits (repeated sites included)	4.4	6.5
Number of aware stores	8.0	2.1
Number of different web stores visited	2.4	1.5

Table 3.3: Summary Statistics of Movie DVD Price

	Mean	Std.Dev.
Total number of transactions	2624	
Number of different web stores with matched price	4.2	2.4
Number of different web stores visited	2.4	1.5
Number of different visited web stores with matched price	1.4	0.8
Number of transactions with all visited web stores price matched	1652	
Number of transactions not from minimum price web store	234	

Table 3.4: Number of Newly Visited Stores and Past Purchase Experience

	(A)	(B)	(C)
Number of Newly Visited Stores	Coef.(Std.Err.)	Coef.(Std.Err.)	Coef.(Std.Err.)
Experience of last purchase			
Bought from the store of last purchase	-0.685 (0.318)***	-0.442 (0.050)***	
Visited the store of last purchase			-0.385 (0.041)***
Demographics			
Broadband connection	0.217 (0.036)***		0.225 (0.037)***
Household Size	0.041 (0.103)***		0.05 (0.011)***
Fixed Effect	No	Yes	No
Constant	1.146 (0.052)***	1.267 (0.039)***	0.938 (0.060)***
Number of observation	6,472	6,472	6,472
R^2	0.075	0.068	0.0216

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table 3.5: Number of Retailers Dropped from Choice Set and Past Purchase Experience

	(A)	(B)	(C)
Number of Retailers Dropped from Choice Set	Coef.(Std.Err.)	Coef.(Std.Err.)	Coef.(Std.Err.)
Experience of last purchase			
Bought from the store of last purchase	-0.579 (0.030)***	-0.379 (0.047)***	
Visited the store of last purchase			-1.081 (0.036)***
Demographics			
Broadband connection	0.144 (0.034)***		0.154 (0.032)***
Household Size	0.023 (0.010)**		0.029 (0.009)***
Fixed Effect	No	Yes	No
Constant	1.104 (0.049)***	1.143 (0.039)***	1.581 (0.052)***
Number of observation	6,472	6,472	6,472
R^2	0.060	0.056	0.125

$*$ significant at 10%; $**$ significant at 5%; $***$ significant at 1%.

Table 3.6: Dynamic Estimates

Parameter	Dynamic Model	Myopic Model
α-Price	0.164	0.164
	(0.017)	(0.003)
c-Search Cost	0.286	0.474
	(0.034)	(0.018)
σ_{ν}^2-Variance of Signals	18.667	8.631
	(0.315)	(0.036)
σ_k^2-Initial Prior Variance	0.0091	0.0074
	(0.055)	(0.045)
δ_1-Quality Mean of Amazon	2.292	5.226
	(0.096)	(0.033)
δ_2-Quality Mean of CH	3.447	3.041
	(0.014)	(0.062)
δ_3-Quality Mean of Others	2.699	2.881
	(0.093)	(0.029)
σ_1^2-Quality Variance of Amazon	0.040	0.161
	(0.012)	(0.093)
σ_2^2-Quality Variance of CH	0.598	0.700
	(0.034)	(0.035)
σ_3^2-Quality Variance of Others	0.041	0.193
	(0.028)	(0.018)

· Standard Errors in Parentheses.

Table 3.7: Movie DVD Market Share Simulations

Market Share	Data	Model Prediction	Simulation 1	Simulation 2.1	Simulation 2.2	Simulation 3
			no search cost	matched price	lower price on Amazon	one-search learning
Amazon	37.97%	35.16%	24.84%	33.82%	33.45%	68.07%
ColumbiaHouse	31.65%	30.56%	57.06%	31.31%	30.73%	31.59%
Other firms	30.38%	34.28%	18.10%	34.87%	35.82%	0.34%

Vita

Jicheng Liu

Personal

Chinese Citizenship
Date of Birth: May 17, 1985

Education

Ph.D. in Economics, The Pennsylvania State University, 2009-2015
M.A in Economics, Shanghai University of Finance and Economics, China, 2007-2009
B.S. in Mathematics, Wuhan University, China, 2003-2007
B.A. in Economics, Wuhan University, China, 2003-2007

Fields of Interest

Primary: Industrial Organization
Secondary: Applied Microeconomics

Research and Teaching Experience

Research Assistant for Professor Joris Pinkse, Summer 2012
Teaching Assistant, Undergraduate, Intermediate Econometrics, 2011-2014
Teaching Assistant, Undergraduate, Intermediate Microeconomics,2009-2010