On the Depth and Dynamics of Online Search Behavior

August 2001

Eric J. Johnson
Department of Marketing
Columbia Business School
Columbia University

Wendy W. Moe
Department of Marketing
McCombs School of Business
University of Texas at Austin

Peter S. Fader
Department of Marketing
The Wharton School
University of Pennsylvania

Steven Bellman
University of Western Australia

Gerald L. Lohse
Accenture

1 We thank Media Metrix, Inc., for providing the data used here, and the supporting firms of the Wharton Forum on Electronic Commerce for their financial support. Please do not quote or cite without permission. Correspondence should be directed to the first author, who can be contacted at 212.854.5068 or ejj3@columbia.edu.
On the Depth and Dynamics of Online Search Behavior

Abstract

This paper examines search across competing electronic commerce sites. By analyzing panel data from over 10,000 Internet households and three commodity-like products, books, CDs and travel, we show that the amount of search is actually quite limited. On average, households that browse a category initially visit only 1.1 book sites, 1.2 CD sites, and 1.8 travel sites. Using probabilistic models, we characterize consumer search behavior at the individual level in terms of (1) depth of search, (2) dynamics of search, and (3) activity of search.

We model an individual's tendency to search as a beta-geometric process, finding that consumers search across very few sites in a given shopping month. We extend the beta-geometric model of search to allow for any time-varying dynamics that may exist causing the consumer to evolve and, perhaps, learn to search over time. We find that for two of the three product categories studied, search propensity does not change from month-to-month. However, in the third product category, we find mild evidence of time-varying dynamics, where search decreases over time from already low levels. Finally, we model the level of a household's shopping activity and integrate it into our model of search. Interestingly, the results suggest that more active online shoppers tend also to search across more sites. This consumer characteristic largely drives the dynamics of search often mistaken as time-varying effects.
On the Depth and Dynamics of Online Search Behavior

1 Introduction

The advent of electronic commerce has engendered a widely held belief: because the Internet lowers search costs, people should search more. There is little doubt that search costs, as measured by time, have decreased: McKinsey Consulting, for example, estimates that finding a high interest rate CD required 25 minutes using the telephone, 10 minutes using the World-Wide Web, and less than a minute with an electronic agent (Butler et al. 1997). Consider the contrast of shopping for a book, knowing its title. Traveling from bookstore to bookstore, even in a shopping mall, might take minutes, but a typical search agent (such as DealTime) will search dozens of stores across multiple countries and provide prices, including shipping and sales tax, within 20 seconds.

In accord with most economic models of search (Bakos 1997), many expect this increased search would lead to lower prices, as consumers explore a larger number of vendors, reducing both the average price paid and the dispersion of prices (Smith et al. 1999). Most search theory (see Diamond 1987 for a review) concentrates on optimal strategies for search, usually in the form of a cutoff stopping price $U(p^*)$, which, under fairly strict assumptions\(^2\), varies as a function of search cost. For example, the infinite horizon stationary case with search costs, $c$, and discount factor, $R <= 1$, yields the following:

\(^2\) In this example, search costs are additive and constant, that all consumers will buy all their units of the good at one store, and for H3, we assume the $f(p)$ is a uniform distribution.
\[ (1 - R)U(p^*) = -c + R \int_0^{p^*} [U(p) - U(p^*)]dF(p) \]

If upon visiting a store one observes a price greater than \( p^* \), optimal behavior calls for further search, while prices below \( p^* \) calls for an end of search and purchase. Thus as \( p^* \) decreases, search increases. This first order condition yields simple comparative statics for optimal search behavior: Search increases and \( p^* \) decreases as search costs \( c \) decrease. Similarly, an increase in the variance of the distribution of prices (holding constant the mean utility from a randomly selected price) makes search more valuable and lowers the cut-off price.

Bakos (1997, Model 1) develops a model that includes the strategic responses of retailers using a spatial unit-circle model and similarly concludes that “Consumers with access to electronic marketplaces, and thus facing lower search costs, become more demanding and are willing to make fewer compromises concerning their ideal product. If the cost of search is low enough, buyers look at all product offerings and purchase the one best serving their needs.” His first model demonstrates that, in commodity markets, reductions of search costs may “destabilize a monopolistic equilibrium and eradicate sellers’ profits.” The picture for differentiated goods is that sellers’ prices decrease smoothly as search costs are reduced. If search costs are low enough, buyers will search all vendors, both finding products that are more closely matched to their needs and, when the number of vendors is large, reducing sellers’ profits to zero. It is important to realize that most economic models of search have two separable predictions for the effect of search cost reductions upon prices: (1) a decrease in observed prices and (2) a decrease in the dispersion of prices (Smith, Bailey and Brynjolfsson, 1999).

Does decreased search cost on the Internet create lower prices? Researchers have sought to study the effects of lower search costs by comparing price levels on the internet to those found in traditional “bricks-and-mortar” channels, particularly for products that are difficult to
differentiate, such as books, compact disks, etc. Surprisingly, perhaps, the data are mixed. An initial study by Bailey (1998) examined prices for books, CD’s and software and found that prices were actually higher on the Internet in 1996 and 1997 than in “bricks-and-mortar” stores. Brynjolfsson and Smith (2000b), on the other hand, found that prices for CD’s and books were 9 to 16 percent lower on the Internet. However, the differences observed between studies may be due to differences in the samples and methodologies used across the two studies.

Price dispersion presents a clearer picture of search cost effects. Both Bailey (1998) and Brynjolfsson and Smith (2000b) find significant dispersion for CD’s and books. Prices for identical items differed, on average, by 33% for books and 25% for CD’s. However, Brynjolfsson and Smith found that price dispersion on the Internet is markedly reduced when weighted by market share. However, as Brynjolfsson and Smith note, this does not mean that the market has reached a price-based equilibrium since the least expensive retailer did not hold the largest shares. A study by Clemons, Hann, and Hitt (1999) of online travel agents again showed high levels of dispersion. Even when they controlled statistically for the quality of the booked ticket, using the number of connections and the proximity of a flight to the requested time, there was still significant dispersion among prices. The lowest and highest priced tickets varied, on average, by at least 10% (or $50) across online travel agents.

Why do we see high levels of price dispersion online? One possibility is that people are not searching broadly across online stores despite lower search costs. Adamic and Huberman (1999) found that the top 1% of sites on the Web capture 50% of all visits to the Web. This is consistent with the idea that consumers are limiting their search to just a few of the most popular sites. These ideas have an analogue in physical space as well. There is an extensive literature describing search for goods, mostly based on self reports (see Newman 1977 for a review). For
example, the number of stores visited for the purchase of a major appliance is reported to be about 3 (Beatty and Smith 1987). Prior research has often argued that the observed amount of search is surprisingly low (Wilkie and Dickson 1985). However recent research suggests that consumers’ priors may account for some of this departure (Moorthy et al. 1997). The current research allows us to revisit some of these issues both because of the decreased costs of search provided by the internet and the use of observational data about search, which although imperfect may be superior to self reports which have been reported to be poor measures of actual search (Newman 1977).

In this paper, we try to understand these issues by looking directly at search behavior of consumers as they visit online retailers. It is our hope that this analysis is both descriptively interesting and can help explain some of the observed patterns of prices which have puzzled observers of online markets (Brynjolfsson and Smith 2000b). Because electronic markets often allow us to observe search in addition to transaction, we think this is both an important new source of evidence, and perhaps an “early warning system” for increases in price competition. Since search is a primary explanatory variable in understanding observed prices in economic models (Diamond, 1987) we hope to provide some stylized facts that will inform future research. In addition, we can examine how the depth of search systematically differs across consumers and offer models that will allow us to separate the factors that might influence search.

2 Data

Recent developments in data collection allow us to examine consumer search directly, by looking at the shopping patterns of a large panel of Internet users over time. We use data collected by Media Metrix, Inc., a firm that records every URL visited by families that are
members of its panel using a small program, the PCMeter, which runs continuously as a background application on the family’s home computer.

The Media Metrix U.S. panel consists of approximately 10,000 households at any given point in time. We examine the online shopping behavior of each of these households during the 12-month period from July 1997 through June 1998\(^3\). Specifically, we are interested in the number of unique sites searched by each household within a given product category. A key issue when examining search behavior is defining the period of time that constitutes a search session. For example, the industry practice is to define a user session by “closing” the session after 15 minutes of inactivity. However, that is a fairly narrow definition of a search session as it is conceivable that a shopper’s decision period, and therefore search, may span over multiple days, weeks, etc. One may visit a given site today to examine the product offering, then deliberate the purchase offline, and return to shopping online perhaps days later.

An alternative method of defining a search session is to close the session after a purchase is made and assume that all store visits prior to that purchase contributed to that purchase cycle. There are two key problems with this method as well. First, not all purchases are made online. A shopper may search online and make some purchases online and others offline. In that case, search sessions as defined by observed online purchases may in fact include multiple purchase cycles. Second, clickstream data typically includes only the URLs viewed and does not necessarily offer any insight into purchasing activity. Some sites, such as Amazon and CDNOW, may have easily identifiable URLs that indicate a purchase occurred. However, not all sites do, in which case we would miss not only purchases occurring offline but also purchases

\(^3\) We used the information contained in the URL of the site to determine the product category. The end of this time period corresponds to the entry of one of the on-line stores, Amazon, into multiple categories, making determination of the product class ambiguous.
occurring at these sites. Our objective is to define a session broadly enough to encompass a series of Internet user sessions that contribute to the same purchasing cycle but also narrowly enough not to inadvertently merge multiple cycles. As a result, we chose to examine each household’s search behavior at the monthly level, which allows us to avoid making any assumptions (or incorrect inferences) about the link between purchasing and online visit behavior.

We focused on three categories: compact disks, books, and air travel, choosing them because: (1) they were relatively frequently purchased online during this time period; (2) they tend to be non-differentiated goods (a book purchased from Amazon is the same book purchased from a low cost provider such as Books.com), increasing the probability that these categories will be subject to broader search; (3) they vary in price from relatively inexpensive (books and CD’s) to more expensive (air travel); and (4) they have been used in studies of price and price dispersion by other researchers. Sites were chosen from each category according to lists of leading online retailers from Media Metrix (http://www.mediametrix.com), BizRate (http://www.bizrate.com) and Netscape’s “What’s Related” feature, which uses Alexa’s records of consumers’ actual surfing behavior to identify related sites. While there may be some sites that are excluded from the analysis, they would probably constitute a very small fraction of the activity in each category. Specifically, our dataset covers consumer search activity across 13 book sites, 16 music sites, and 22 travel sites (Table 1), a more inclusive set of sites within each category than that used by either Brynjolfsson and Smith (2000) or Clemons et al (1999). Sites included in each category range in visitor traffic. For example, the smallest CD site in our dataset attracted only 11 unique visitors in the MediaMetrix sample while the largest in the same category attracted over 1800 unique visitors.
3 Analysis, Models, and Results

Using these data, we first examine the number of stores visited in a typical month. Figure 1 shows the average number of websites visited by households each month in which they were actively shopping in the product category. For example, the average number of CD sites searched in a household's first month of shopping is 1.23 and increases to 1.62 in the fourth shopping month for those households that shopped in four (or more) different months in the dataset.

Two patterns are striking in the data. First, the current level of search is low, initially ranging from 1.1 stores for books to 1.8 for travel. In fact, 70% of the CD shoppers, 70% of the book shoppers, and 42% of the travel shoppers were observed as being loyal to just one site throughout the duration of our data. Second, there seems to be an increase in search from month to month. This seems to suggest that Internet search, while currently fairly low, may be increasing over time. One possible explanation for the increasing trend seen in Figure 1 is that consumers may be evolving and searching more as they gain experience, consistent with the idea that time will lead to lower prices and reduced dispersion. We will more closely examine this dynamic of search in the next section.

However, aggregate patterns like those in Figure 1 can be misleading, since they can, in principle, mask different underlying trends that may exist at the individual level. To model these trends, and to provide a more accurate portrait of shopping behavior, we examine search using
models that allow us to decompose this data into three components: (1) Depth of search: the decision of the household to visit more than one store in a given month, (2) Dynamics of search: the evolution of the number of stores visited over time, and (3) Activity of search: the overall amount of category-level shopping activity for each household in a given product class.

### 3.1 Depth of Search

We first examine the depth of household search behavior using a stationary beta-geometric model as a benchmark. Conceptually, this model describes a process where a consumer first visits one site in a given (active) month. At that point, there is a probability, \( p \), that the shopper will stop searching and not visit any more sites and a probability, \( 1-p \), that the shopper will continue searching and visit yet another store. The same probabilistic process continues for each subsequent decision (e.g., should I stop or continue searching). As a result, the probability of household \( i \) searching \( x \) stores in a given month can be captured by a shifted geometric process,

\[
P(X_i = x_i | p_i) = p_i(1-p_i)^{x_i-1}
\]

where the mean number of sites searched in a given month is \( E[x_i] = 1/p_i \). Since households vary in their propensities to visit multiple stores, we allow for heterogeneity by assuming that \( p_i \) will vary across the population according to a highly flexible beta distribution,

\[
f(p_i) = \frac{1}{B(a,b)} p_i^{a-1}(1-p_i)^{b-1}
\]

where the beta function, \( B(a,b) \), is defined as \( \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \). The resulting model for each individual household, \( i \), searches \( \{x_{ij}, x_{i2}, \ldots, x_{ij}\} \) in months \( j=1, 2, \ldots, J_i \) is:
\[ P(X_{il} = x_{il}, X_{i2} = x_{i2}, \ldots, X_{ij} = x_{ij}) = \prod_{0 \leq j < 1}^{J_i} p_i (1 - p_i)^{x_{ij} - 1} \cdot \frac{1}{B(a, b)} p_i^{a-1} (1 - p_i)^{b-1} dp_i \] (3)

where \( J_i \) is the number of months during which household \( i \) is actively shopping. The resulting likelihood function for the \( N \) households' observed monthly search behavior is then:

\[ L = \prod_{i=1}^{N} P(X_{il} = x_{il}, X_{i2} = x_{i2}, \ldots, X_{ij} = x_{ij}) \]

\[ = \prod_{i=1}^{N} \frac{B(a + J_i, b + \sum_{j=1}^{J_i} x_{ij} - J_i)}{B(a, b)} \] (4)

Using the estimated parameters from the beta distribution, the expected stopping probability is calculated as \( E[p] = \frac{a}{a+b} \), and the average number of sites searched in a given (active) month is calculated as \( E[x] = \frac{a+b-1}{a-1} \).

Results of the beta-geometric model estimated on each of the three product categories are presented in Table 2. Consistent with the low search levels seen in Figure 1, the model indicates that depth of search is low with very high average stopping probabilities \( [p_{CD} = 78.8\%, \ p_{books} = 87.2\%, \ p_{travel} = 59.1\%] \). However, the shapes of the associated beta distributions are not what we might expect: We might suspect that the distribution of consumers' search propensities would be segmented with a large proportion of the population with stopping probabilities equal to 1.0 (i.e., loyals always stopping after visiting their first site) and the remainder of the population with more search, as is the case of a beta distribution with \( a>1 \) and \( b<1 \). However, the model parameters suggest that this is not the case. Figure 2 plots the beta distribution of stopping probabilities for each of the three product categories. The probability density function of the stopping probabilities in the CD market, for example, is tightly distributed around a mode at \( (a-1)/(a+b-2) = 94.3\% \). Surprisingly few households (approximately 2.2\%) have \( p_i > 99\% \).
However, there may be a segment of hard-core loyalists that is being overlooked by imposing a single beta distribution on both searchers and loyalists. This argument appears to be supported by the fact that such high percentages of shoppers search no more than one site over the entire data period. One way to test this argument is to include a component in the beta-geometric model that allows for hard-core loyalists:

\[
P(x \mid p) = \begin{cases} 
\pi + (1 - \pi) \cdot P(x \mid p) \cdot f(p) & \text{when } x = 1 \\
(1 - \pi) \cdot P(x \mid p) \cdot f(p) & \text{when } x > 1
\end{cases}
\]  

(5)

where \( \pi \) is probability that any given household is a hard-core loyalist (i.e., stopping probability, \( p \), equals one). But this model does not indicate that a significant segment of hard-core loyalists exists. In fact, this model, when applied to the books and travel categories, produces results that are no different from the simple two parameter beta-geometric model (i.e., the estimated value of \( \pi \) is equal to zero). For the music category, the inclusion of a loyalist segment provides only a slight improvement in model fit (BIC=−7962.7) with a very small segment of loyal shoppers (\( \pi = 6.8\% \)), far less than the 78.8% seen in the aggregate summary statistic. The implication is that the vast majority of households are not strictly loyal. Instead, they have some propensity to search but rarely exercise it.

The beta-geometric model results largely confirm the low levels of search that we observe in Figure 1 and suggest that these low levels may apply to much of the Internet shopping population. However, Figure 1 also seems to suggest an evolution of shopping behavior over
time, with households learning, perhaps, to increase search. We will test this argument in the
next section.

3.2 Dynamics of Search

One interpretation of the beta distribution in this type of probabilistic mixture model is
that it represents the heterogeneity across the population. However, as we observe behavior for
each household over time, we can use the parameters of the beta distribution, in conjunction with
Bayes Theorem, to calculate a more refined “guess” for each household’s likely stopping
probability and identify where they fall on the population distribution:

\[
f(p_i | x_{i1}, x_{i2}, \ldots, x_{ij}) = \frac{1}{B(a + J_i, b + \sum_j x_{ij} - J_i)} p^{a + J_i - 1} (1 - p)^{b + \sum_j x_{ij} - J_i - 1}
\]

and

\[
E[p_i | x_{i1}, x_{i2}, \ldots, x_{ij}] = \frac{a + J_i}{a + b + \sum_j x_{ij} - J_i}
\]

Therefore, an alternative and equally valid interpretation of the parameters of the beta
distribution is that they represent the uncertainty around our estimate of each individual
household’s stopping probability. And as we observe more activity from each household, the
beta distribution that best describes our estimate for that household narrows around its true latent
probability. Additionally, covariates specific to an individual household may also shift their
latent stopping probability (and therefore shift the beta distribution describing it) by increasing or
decreasing the parameters of the distribution. To examine the dynamics of search, we extend the
beta-geometric model to allow its parameters \(a\) and \(b\) to vary over time for each household.

For example, with experience on Internet, the parameters \(a_{ij}\) and \(b_{ij}\), that determine
household \(i\)'s stopping probability for month \(j\) can vary as a function of the number of past

shopping sessions. Combined with a Bayesian update, the beta distribution that captures an individual household’s search behavior can be described with the following parameters:

\[ a_{ij} = a_0 \exp\{\gamma_a (j - 1)\} + j \]  \hspace{1cm} (7a)

\[ b_{ij} = b_0 \exp\{\gamma_b (j - 1)\} + \sum_j x_{ij} - j \]  \hspace{1cm} (7b)

This approach is often called a beta-logistic model. The relative size of \( \gamma_a \) compared to \( \gamma_b \) drives the nature of change over time. For example, if \( \gamma_a \) is negative and \( \gamma_b \) is positive, \( a \) would decrease and \( b \) would increase over time. As a result, stopping probabilities would decrease over time, increasing the number of stores searched in a month. In contrast, if \( \gamma_a \) is positive and \( \gamma_b \) is negative, the number of stores searched would decrease over time.

### 3.3 Activity of Search

An alternative set of dynamics, that can also explain the patterns shown in Figure 1, offers a very different interpretation of these data. Specifically, the pattern that we observe in Figure 1 could stem from a selection effect. Because we are able to observe more activity from the frequent shoppers in our data period, we may also be observing an increasingly greater proportion of these more active households as we move from left to right in the figure. These relatively active shoppers (i.e., those who visit a particular category more frequently) may be inherently different from those less active and may tend to search across more sites. As a result, the increase we see may not represent more stores being visited by the same individuals but rather a change in the mix of shoppers as we move to a greater number of active months. In short, there may be no household-level dynamics whatsoever, but only an apparent pattern resulting from consumer heterogeneity.
To explore the relationship between the level of shopping activity levels and depth of search, we incorporate a second covariate into the beta-geometric model. In addition to modeling the potential learning effect or time dynamic that may result from repeated searches, we now also incorporate each household’s shopping activity level into the beta parameters:

\[
a_{ij} = a_0 \exp\{\gamma_b (j - 1) + \lambda_a q_i\} + j \quad (8a)
\]

\[
b_{ij} = b_0 \exp\{\gamma_b (j - 1) + \lambda_b q_i\} + \sum_j x_{ij} - j \quad (8b)
\]

where \(q_i\) is a measure of the household's shopping activity level. One measure we could use to characterize a household's activity level is simply the proportion of months for which the household is actively shopping. However, there is a critical drawback of using such a measure -- not all households are present in the panel for equal amounts of time. For example, a household that has been in the panel for only a month and was also active in that month would be represented by this measure as a household that was active 100% of the time (or every month). A measure based on this limited history of behavior is likely to misrepresent that individual’s latent long-term behavior. A better measure would incorporate a shrinkage estimate derived from a separate category-level model of activity. Effectively, we examine the number of months each individual is actively shopping as a beta-binomial process. Each individual household in the panel has a probability, \(q_i\), of visiting a product category in any given month they are in the panel (regardless of which store(s) they choose to visit). Given that household \(i\) was in our dataset (but not necessarily active in a given category) for \(T_i\) months, the probability of shopping in \(J_i\) of those \(T_i\) months is:

\[
P_i(J_i \mid T_i, q_i) = \binom{T_i}{J_i} q_i^{J_i} (1 - q_i)^{T_i - J_i} \quad (9)
\]
Furthermore, $q_i$ is distributed across the population according to a beta distribution with parameters, $k$ and $m$. The resulting beta-binomial model and the associated likelihood function follow:

$$P(J_i) = \binom{T_i}{J_i} \frac{B(k + J_i, m + T_i - J_i)}{B(k, m)}$$  \hfill (10)

$$L = \prod_{i=1}^{N} P(J_i) = \prod_{i=1}^{N} \binom{T_i}{J_i} \frac{B(k + J_i, b + T_i - J_i)}{B(k, m)}$$  \hfill (11)

This beta-binomial model provides a general measure and method of assessing each individual’s shopping activity. Previous studies have shown that the frequency or activity level of shoppers may evolve as they gain experience with the sites (Moe and Fader 2000). However, in this paper, we are concerned only with overall activity levels for the time period in question and its effect of breadth of search, not the specific dynamics driving activity levels themselves. After estimating this model and obtaining the parameter estimates (see Table 3), we can use Bayes theorem to calculate each household's expected activity level in terms of $q_i$. This can be expressed as a combination of the parameter estimates and each individual's observed behavioral history.

$$q_i = \frac{k + J_i}{k + m + T_i}$$  \hfill (12)

Table 4 presents the results of the beta-logistic model incorporating both time-varying as well as activity-dependent dynamics. In all three cases, more active households searched more store sites than the less active households. Evidence of any month-to-month dynamics at the
household level, however, is limited. Table 5 provides fit statistics that illustrate the value of including the time-varying and activity-dependent components in the model. Incorporating a time-varying component improves fit over the static beta-geometric component. However, the time-varying component contributes less to the model fit than the activity-dependent model. More interestingly, when both components are included, the model does not significantly outperform the activity-dependent only model, as indicated by BIC. This implies that any dynamics observed in search behavior is driven primarily by differences in activity levels across households and not any changes in search behavior over time, though this effect is modest in the travel category. Figure 3 plots how the beta distribution governing a household's tendency to search varies across consumers based on their activity levels. For illustration purposes, we plot the beta distributions governing consumer stopping probabilities for "low activity" shoppers versus "high activity" shoppers. The low activity level is represented by the value of $q$ at the lower bound of the interquartile range for each category, while the high activity level is represented by the upper bound. For each of the three product categories, we find that consumers of higher activity levels have lower expected stopping probabilities and hence search across more store sites. Because the time-varying plus activity-dependent model does not fit significantly better than the activity-dependent only model, we conclude that there is no statistical support for time dynamics when searching for books, CDs, or travel. That is, shoppers do not search more as they gain experience with the category. This result contradicts the argument that the trend seen in Figure 1 is a result of learning or experience effects.
In general, there seems to be very little search in the books and CD categories. The travel category experience perhaps a bit more search but still far less than one might expect with the average shopper searching 1.6 sites. The fact that search in the travel category is actually more than the amount of search seen in the books or CD categories is somewhat expected given the high involvement nature of the product category. However, if one examines the types of sites included in the travel category, many serve as search “bots” that scan for prices from multiple airlines. Since some sites provide the shopper with comparison shopping tools within the site itself, it is surprising that individuals search more across travel sites than they do in the books and CD category where the sites represented are pure e-commerce vendors and not search ‘bots. This suggests that search ‘bots are not the reason behind the limited search that we find in these categories.

4 Discussion

4.1 Summary

In summary, these three categories show fairly low levels of search overall. While more active shoppers tend to visit more sites in any given month, there is no evidence that experience increases the number of sites visited. We might expect the greatest returns to search for travel services, both because prices can change over time and because this is a more expensive purchase. However, we find that experience leads to a slight decrease in the number of visited sites.

Like Adamic and Huberman (1999) our data suggest that people visit few stores online despite the fact that consumer are “just a mouse click away” from other stores. Browsing
behavior varied by product category and level of activity but showed no increase with experience.

In understanding the significance of our results, it is important to realize that we examined product classes and time periods which are very similar to those used by researchers who have examined changes in price and price dispersion on the web. The attraction of books, compact disks, and travel for examining these questions is their apparent commodity-like status, which should provoke higher levels of search. And our data have been collected during time frames that substantially overlap time periods used in prior studies of prices for CD’s and books (Brynjolfsson and Smith 2000b) and airline travel (Clemons et al. 1999), minimizing the potential for historical differences and suggesting that the lack of search is one reason we see price dispersion.

It is also interesting to note that our results stand in sharp contrast to self-reports based on survey data. Self-reported data are subject to the fallibility of people's memories, idiosyncratic scale use, and even deliberate alteration through social desirability biases and have been found to have very little correlation with actual search. However, industry analysis based on self-report suggests higher levels of search (a reported 2.4 visits per purchase) than we observe (McQuivey 1999). We believe additional research is needed to compare self-reported behavior to actual Web usage, however, since the type of panel data examined here reflects actual behavior, we suspect it may serve as a baseline for such efforts. One example of such effort are results (Brown and Goolsbee, 2000) that suggest that recent declines in life insurance prices are the result of increased use of the Internet shopping sites for insurance. An interesting next step would be to examine actual (as opposed to self-reports) of usage to see if observed search is correlated with lower prices paid.
4.2 The role of shopbots

While the dataset used here contains a rich and realistic portrayal of online shopping, it is an early snapshot of such behavior. While it could be argued that this will change as the online market matures, our analysis of time-dynamics does not suggest this was happening in the time frame observed. All three classes showed no increase in the number of sites visited. One major trend which may well modify these results is the widespread adoption of price search agents, or price robots, termed ‘bots for short. We examined our data for the use of such ‘bots, but found that their use at this time was at very low levels. For example, the most popular was Acses, a price search agent for books, which had a total of 17 uses in the Media Metrix panel. As consumers may become more sophisticated in their search over time and as search agent technology evolves, we may see automated search further lowering search costs and minimizing prices and price dispersion. However recent research indicates that even those who use shopping robots seem to display loyalty to sites previously used. Brynjolfsson and Smith (2000a) examined users of EvenBetter.com, a popular search engine for books (as well as CDs and videos, which they did not examine). It is interesting to note that even though this subset of internet consumers were highly price sensitive and tended to patronize those stores with cheaper prices, over 51% of these customers did not choose the retailer with the cheapest price. Even with this self-selected group of consumers, brand has a significant advantage. Amazon, for example, commands a price premium of $1.85 over unbranded retailers in a logistic analysis of purchases by this segment.
4.3 Caveats and limitations

In this study, the analysis of depth of search and activity level of search was conducted at the household level, rather than at the level of the individual consumer. By aggregating consumers to the household level, we overstate both the activity level and depth of search. Households with multiple members shopping in different months would have an activity level that is higher than that of any one individual in the household. Similarly, households with multiple members that are each loyal to a different site will demonstrate search propensities at the aggregate household levels. Therefore, even though we observe low levels of search at the household level, search at the individual consumer level would be even lower.

A potential disadvantage of the type of clickstream data used here is that we identify all browsing activities, not just search behavior associated with purchases. While such browsing may be an important part of consumer search, we are unable to partition our observations into those associated with purchase-oriented behavior and those that may be associated with other activities (e.g., using Amazon.com to find biographical information about a certain author). While this may generally lead to an overestimation of the amount of search in our dataset, it is possible that purchase related search might feature more extensive search than other “look up” activities.

Finally, the time period studied in this paper represents a fairly early stage in electronic commerce. As such, the subset of the consumer population shopping online may be very different from the consumer population as a whole. Studies conducted around the same period of time have found that online consumers tend to be more time constrained than the average consumer (Bellman, Lohse and Johnson 1999). As such, this subset of the consumer population that shops online may consist largely of those consumers who are unable to spare the time to
search across multiple sites. Projecting the search behavior of this subset of shoppers to the consumer population as a whole may lead to biased conclusions.

4.4 Why is search so limited?

One possible explanation for why we see such little search is that the current market for these goods is efficient. However, the fact that there is still significant price dispersion in these Internet markets suggests that this is not the case. Brynjolfsson and Smith (2000b) show significant price dispersion on the Internet and examine a number of possible explanations. One is that the product offerings may actually be heterogeneous and not commodities because of value added services of various kinds. They dismiss this critique by examining the offerings of online book and CD vendors and argue that the offers do not differ significantly, and more importantly, do not differ as a function of price. Similarly, Clemons et al (1999) find that dispersion exists even when they control statistically for the quality of the airline tickets in their study. Thus, the dispersion found in these studies suggest that these markets have not yet produced the kind of highly efficient markets thought to be a function of electronic commerce.

Another possibility is that normative models of search are not complete. While clearly any modifications of standard search models exceed the scope of this paper, it is worth noting that the heart of most search models is the tradeoff between the cost of search, usually measured by time, and the benefit of that search to the consumer. In other domains, involving the allocation of time have benefited from richer descriptive frameworks that incorporate the effects of context and mental accounts. Two examples that seem relevant are descriptive theories of time-money tradeoffs (Loewenstein and Prelec 1992; Loewenstein and Thaler 1997) and the observation that out-of-pocket costs are overweighted relative to opportunity costs (Thaler 1999).
A third possibility is the realization that search costs are not constant over time and that they change as consumers gain experience shopping with a particular online store. For example, by visiting a site, one learns its navigational scheme, which reduces the cost of using that site in the future. Similarly, the site can make changes, through customization, user-based recommendations, and memorization of names, addresses and payment details, which lower that cost of that site relative to others. This idea that search costs are dynamic is analogous to the concept of lock-in, which has been discussed in markets for technology goods, and is a topic we are exploring elsewhere (Johnson, Bellman and Lohse, 2001).

4.5 Conclusion

We see this research as in initial demonstration that data from the Web are changing the role of search from an unobservable explanatory variable in the analysis of markets to one that can and should be observed and included in empirical analysis of market behavior. Unlike search in other off-line environments, the kind of data used here can track search at a fine level. In fact, much of the problem in the analysis of the data will be developing theories and methods which provide efficient methods for analysis, given the large amount of data provided by shoppers on the internet. While the challenges are significant, we believe the analysis of search data will be great in expanding our understanding of the role of search in explaining marketplace behavior, and the analysis offered here represents a first step.
4.5.1 References.


Table 1. E-Commerce Sites Included in Books, CDs, and Travel Categories

<table>
<thead>
<tr>
<th>Books</th>
<th>CDs</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acses</td>
<td>CDNOW</td>
<td>Alaska_Airlines</td>
</tr>
<tr>
<td>AltBookStore</td>
<td>MusicBlvd</td>
<td>American_Airlines</td>
</tr>
<tr>
<td>Amazon</td>
<td>BestBuy</td>
<td>Best_Fares</td>
</tr>
<tr>
<td>Barnes_and_Noble</td>
<td>CDConnect</td>
<td>Cheap_Tickets</td>
</tr>
<tr>
<td>Books.com</td>
<td>CDEurope</td>
<td>City.Net</td>
</tr>
<tr>
<td>Books-a-Million</td>
<td>CDUniverse</td>
<td>Continental</td>
</tr>
<tr>
<td>BooksNow</td>
<td>CDUSA</td>
<td>Delta-Air</td>
</tr>
<tr>
<td>Book_Zone</td>
<td>CDWorld</td>
<td>Expedia</td>
</tr>
<tr>
<td>Borders.com</td>
<td>Emusic</td>
<td>European_Travel_Network</td>
</tr>
<tr>
<td>Kingbooks</td>
<td>MassMusic</td>
<td>ITN</td>
</tr>
<tr>
<td>Superlibrary</td>
<td>Newbury</td>
<td>Lowest_Fare</td>
</tr>
<tr>
<td>Powells</td>
<td>Tower</td>
<td>NWA</td>
</tr>
<tr>
<td>Wordsworth!com</td>
<td>Tunes</td>
<td>Preview_Travel</td>
</tr>
<tr>
<td></td>
<td>Ktel</td>
<td>Priceline_Travel</td>
</tr>
<tr>
<td></td>
<td>Music Spot</td>
<td>Southwest_Airlines</td>
</tr>
<tr>
<td></td>
<td>Music Central</td>
<td>The_Trip</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Travel_Web</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Travel_Zoo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Travelocity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TWA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>United_Airlines</td>
</tr>
<tr>
<td></td>
<td></td>
<td>US_Airways</td>
</tr>
</tbody>
</table>
### Table 2. Beta-Geometric Model Results

<table>
<thead>
<tr>
<th></th>
<th>Books</th>
<th>CDs</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>14.05</td>
<td>5.81</td>
<td>5.17</td>
</tr>
<tr>
<td>$B$</td>
<td>1.92</td>
<td>1.29</td>
<td>2.89</td>
</tr>
<tr>
<td>LL</td>
<td>-5197.4</td>
<td>-3982.4</td>
<td>-1848606</td>
</tr>
<tr>
<td>BIC</td>
<td>10412.2</td>
<td>7948.4</td>
<td>36991.3</td>
</tr>
<tr>
<td>$E[p]$</td>
<td>0.87</td>
<td>0.79</td>
<td>0.59</td>
</tr>
<tr>
<td>$E[x]$</td>
<td>1.15</td>
<td>1.27</td>
<td>1.69</td>
</tr>
<tr>
<td>modal $p$</td>
<td>0.93</td>
<td>0.94</td>
<td>0.69</td>
</tr>
</tbody>
</table>

### Table 3. Results of Beta-Binomial Model of Activity Levels

<table>
<thead>
<tr>
<th></th>
<th>Books</th>
<th>CDs</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>0.48</td>
<td>0.35</td>
<td>0.43</td>
</tr>
<tr>
<td>$m$</td>
<td>4.73</td>
<td>6.60</td>
<td>3.12</td>
</tr>
<tr>
<td>LL</td>
<td>-35544.0</td>
<td>-21806.0</td>
<td>-42979.5</td>
</tr>
<tr>
<td>$E[q]$</td>
<td>0.73</td>
<td>0.13</td>
<td>0.12</td>
</tr>
</tbody>
</table>

### Table 4. Time-Varying and Activity-Dependent Model Results

<table>
<thead>
<tr>
<th></th>
<th>Books</th>
<th>CDs</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>79.60</td>
<td>9.09</td>
<td>7.41</td>
</tr>
<tr>
<td>$B$</td>
<td>4.85</td>
<td>0.74</td>
<td>3.13</td>
</tr>
<tr>
<td>$\gamma_a$</td>
<td>-0.32</td>
<td>-0.11</td>
<td>0.03</td>
</tr>
<tr>
<td>$\gamma_b$</td>
<td>-0.29</td>
<td>-0.22</td>
<td>0.07</td>
</tr>
<tr>
<td>$\lambda_a$</td>
<td>-1.93</td>
<td>0.46</td>
<td>-0.71</td>
</tr>
<tr>
<td>$\lambda_b$</td>
<td>1.29</td>
<td>6.70</td>
<td>-0.12</td>
</tr>
<tr>
<td>LL</td>
<td>-5067.6</td>
<td>-3844.4</td>
<td>-18402.8</td>
</tr>
<tr>
<td>BIC</td>
<td>10187.5</td>
<td>7738.0</td>
<td>36859.9</td>
</tr>
</tbody>
</table>
Table 5. Model Fit Statistics

<table>
<thead>
<tr>
<th></th>
<th>Books</th>
<th></th>
<th>CDs</th>
<th></th>
<th>Travel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>BIC</td>
<td>LL</td>
<td>BIC</td>
<td>LL</td>
<td>BIC</td>
</tr>
<tr>
<td>Stationary</td>
<td>-5,197.4</td>
<td>10,412.2</td>
<td>-3,982.4</td>
<td>7,948.4</td>
<td>-18,486.6</td>
<td>36,991.3</td>
</tr>
<tr>
<td>Time-Varying</td>
<td>-5,142.6</td>
<td>10,320.0</td>
<td>-3,961.3</td>
<td>7,955.5</td>
<td>-18,469.0</td>
<td>36,983.6</td>
</tr>
<tr>
<td>Activity-Dependent</td>
<td>-5,071.4</td>
<td>10,177.7</td>
<td>-3,848.7</td>
<td>7,730.2</td>
<td>-18,406.9</td>
<td>18,406.9</td>
</tr>
<tr>
<td>Activity-Dependent plus Time-Varying</td>
<td>-5,067.6</td>
<td>10,187.5</td>
<td>-3,844.4</td>
<td>7,738.1</td>
<td>-18,402.8</td>
<td>36,859.9</td>
</tr>
</tbody>
</table>
Figure 1. Average Number of Online Stores Visited During Each Observed Shopping Month

![Bar chart showing the average number of online stores visited during each month for CDs, Books, and Travel. The chart indicates the loyalty levels: CDs with 70% loyal to one site, Books with 70% loyal to one site, and Travel with 42% loyal to one site.]

Figure 2. Beta-Distributions of Stopping Probabilities

![Diagram showing beta-distributions for stopping probabilities across different shopping categories: CDs, Books, and Travel. The x-axis represents the stopping probability, and the y-axis represents the density function f(p).]

CDs (70% loyal to one site) | Books (70% loyal to one site) | Travel (42% loyal to one site)
Figure 3. Activity-Dependent Effects on Search

In each of the three categories, the beta distribution governing stopping probabilities shift depending on the activity level of the shopper such that more active shoppers have lower expected shopping probabilities leading to less search across sites.

** Low activity is defined by the lower bound of the interquartile range of activity level (q) for shoppers in each category: q=0.11 for books, q=0.08 for CDs, and q=0.16 for travel. High activity is defined by the upper bound: q=0.23 for books, q=0.15 for CDs, and q=0.31 for travel.