Search Less, Find More? Examining Limited Consumer Search with Social Media and Product Search Engines

Completed Research Paper

Anindya Ghose
Stern School of Business, New York University
44 West 4th Street, New York, NY 10012
aghose@stern.nyu.edu

Panagiotis G. Ipeirotis
Stern School of Business, New York University
44 West 4th Street, New York, NY 10012
panos@stern.nyu.edu

Beibei Li
Heinz College, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
beibeili@andrew.cmu.edu

Abstract

With the proliferation of social media, consumers' cognitive costs during information-seeking can become non-trivial during an online shopping session. We propose a dynamic structural model of limited consumer search thatcombines an optimal stopping framework with an individual-level choice model. We estimate the parameters of the model using a dataset of approximately 1 million online search sessions resulting in bookings in 2117 U.S. hotels. The model allows us to estimate the monetary value of the the search costs incurred by users of product search engines in a social media context. On average, searching an extra page on a search engine costs consumers $39.15 and examining an additional offer within the same page has a cost of $6.24, respectively. A good recommendation saves consumers, on average, $9.38, whereas a bad one costs $18.54. Our policy experiment strongly supports this finding by showing that the quality of ranking can have significant impact on consumers' search efforts, and customized ranking recommendations tend to polarize the distribution of consumer search intensity. Our model-fit comparison demonstrates that the dynamic search model provides the highest overall predictive power compared to the baseline static models. Our dynamic model indicates that consumers have lower price sensitivity than a static model would have predicted, implying that consumers pay a lot of attention to non-price factors during an online hotel search.

Keywords: *Consumer Search, Search Cost, Varying Choice Sets, Click-Through, Conversion, Search Engine, Ranking, Econometrics, Dynamic Structural Model, Optimal Stopping*

Introduction

With the growing pervasiveness of social media and Web 2.0 techniques, the volume and complexity of information has become increasingly large. For example, websites such as Amazon.com, TripAdvisor.com or Yelp.com can easily attract hundreds or even thousands of review postings that constantly compete for users' attention. The onslaught of the exploding social media content can lead to a significant information overload for consumers during product search. According to the 2012 Consumer Review Survey (SearchEngineLand 2012), although a large majority of consumers read online reviews before purchase, 68% of consumers read only 2-10 reviews, and only 7% read more than 20 reviews. Excess social content may hinder consumers from efficiently seeking information and making decisions. What is worse, it may discourage consumers from searching and cause unexpected termination of search (e.g., session drop-out). Clearly, with the deluge of structured and unstructured content generated by the online social communities, consumers' cognitive costs in searching and evaluating product information become non-negligible and may potentially aggravate the frictional market.

Based on a recent study, 86% of Internet consumers ranked online search as the most critical step in their buying process (GroupM Search 2011). During the past decade, product search engines have been trying to combine advanced techniques from information retrieval (e.g., Google Product Search) and recommender systems (e.g., Amazon.com) into their ranking design to improve search performance. Recently, product search engines start looking at social media and social networks (e.g., Bing Social Search and TripAdvisor.com) in an effort to improve consumer search experiences with richer "social" information. However, although reducing search cost has been the main focus for search engines and online market designers, little research has been done on quantifying exactly how the evolving social content on product search engines and the various ranking recommendations affect consumers' cognitive costs of searching and evaluating product information. Therefore, one major goal of our study is to examine the role of social media and product search engines in influencing consumer search cost in the online market. In particular, search cost should be not only an inherent attribute of a consumer, but also a consequence of the social context in which the consumer is embedded. By modeling search cost as a random-coefficient function of inherent and social contextual variables, we aim to examine the nature of search cost, which would otherwise have been modeled as a black box.

However, analyzing search cost and its influence on product demand can be challenging. Under economic theory of consumer choice, traditional demand estimation for the online market assumes that consumers search exhaustively with zero costs and that choice sets (consideration sets) are complete and exogenous. However, in reality, consumers are endowed with non-zero search cost and can search only within limits. Therefore, consumers' choice sets are limited. In addition they are formed dynamically, given that they are endogenous to consumers' heterogeneous preferences. A static demand estimation framework that simply takes the consideration sets as being exogenously given (e.g., a static discrete choice model) is not an ideal modeling choice in this scenario. Ideally, during the online search process, even if a consumer does not end up purchasing a product, the decision to search can convey rich information about the consumer's heterogeneous preferences and, therefore, should be incorporated into the demand model. Unfortunately, although there has been extensive theoretical research on the economics of consumer information search since Stigler (1961), due to model complexity and data limitations, empirical work on this issue in the online market is still in its infancy.

Another challenge in estimating product demand with search cost is how to simultaneously identify consumers' heterogeneous preferences and search cost. As pointed out by Sorensen (2001) and Hortacsu and Syverson (2004), explaining search decisions by consumers with heterogeneous preferences imposes an identification problem: A consumer may stop searching either because of a high valuation for the products already found or because of a high search cost. The same observed search outcome can be explained either by the preferences for product characteristics or by the moments of the search cost distribution (Koulayev 2010). Therefore, it is crucial to understand how these two causes can be uniquely recovered and what types of data are needed for the empirical identification. The key identification strategy in our estimation relies on the fact that consumer preferences enter the decision-making processes of both search and purchase of the product, whereas consumer search cost enters only the search decision-making process. Once the consideration set is generated after search, the conditional purchase decision should depend only on the consumer preferences. Our unique dataset containing both consumer search information and purchase information allows us to successfully identify these two effects.

More specifically, in this paper, we relax the "exhaustive search" assumption from the standard demand estimation approaches and examine the limited nature of consumer online product search under the proliferation of social media. To achieve this, we propose a dynamic structural model for sequential search. It combines an optimal

stopping framework with an individual-level random utility choice model, which allows us to jointly estimate consumer heterogeneous preferences and search cost. Our estimation is validated on a unique dataset from the online hotel search industry. We have detailed individual-level search and transaction data from November 2008 through January 2009, containing approximately one million online sessions for 2117 hotels in the United States. We find that a dynamic model with limited consumer search provides a more precise measure of consumer price sensitivity and heterogeneous preferences than does a static model that does not account for the endogenous formation of choice sets.

Our results indicate that too much feedback from online social communities, as well as long sentences, complex words or spelling errors in the social media content, may lead the consumer to terminate the search early. In particular, our findings allow us to quantify the consumers' cognitive costs of seeking and absorbing the structured and unstructured product information available in social media contexts. Furthermore, we are able to quantify the search cost associated with the use of product search engines. On average, the effort of continuing to search an extra page on search engines costs $39.15, while the effort of continuing to search an additional screen position on the same page costs $6.24. Our findings are consistent with previous findings suggesting a non-trivial search cost in online markets. For example, Koulayev (2010) found a search cost of $43.80 per page on a travel search engine. Brynjolfsson et al.(2010) found that the benefits from searching lower screens equal $6.55 for the median consumer. Hann and Terwiesch (2003) quantified rebidding costs to be $4-$7.50 in a reverse auction channel. Hong and Shum (2006) found consumers' median non-sequential search costs to be $1.31-$2.90 for a sample of text books. And de los Santos (2008) found search costs ranging from $0.90 to $1.80 per search in the online book industry. Furthermore, our results suggest that a good ranking recommendation can save consumers, on average, $9.38. A bad ranking recommendation, on the contrary, can lead to an $18.54 loss for consumers. Our findings strongly illustrate the importance of effective ranking design for product search engines.

Our study builds on Weitzman's (1979) optimal sequential search framework. To the best of our knowledge, four existing studies that are closest to our work are Koulayev (2010), Kim et al. (2010), Bronnenberg et al (2012) and Chen and Yao (2012). However, our research differs from these three studies in the following ways: (1) Our model incorporates not only consumers' search behaviors, but also their purchase behaviors, whereas the first two studies considered consumers' search information only as an approximation to their actual purchase decisions. (2) Our observations include the detailed click-throughs from each ranking position on a page, which allows us to precisely model the individual click probability for a product, rather than for a page with a bundle of products (i.e., a page of 15 hotels in Koulayev 2010). (3) Our analysis is conducted at the individual-consumer level as opposed to at the aggregate market level (Kim et al. 2010 and Bronnenberg et al 2012). (4) We consider not only consumers' efforts to refine their searches (e.g., choosing to customize the ranking method), but, moreover, we examine the search costs associated with the refinement tools. We model consumer search refinement and the actual search/click as separate steps. However, Chen and Yao (2012) assume zero costs of the customization efforts and, therefore, treat search refinement as a prerequisite to consumer search. (5) We focus not only on estimating demand, but, more importantly, we are interested in how structured and unstructured information across social media and search engine platforms affect consumer search cost in an online social environment, whereas Koulayev (2010), Kim et al. (2010) and Chen and Yao (2012) focus mainly on demand estimation and consumer welfare analysis from the classic economic and marketing perspectives.

Our key contributions can be summarized as follows. First, we quantify the effects of social media and product search engines on consumer search cost. By modeling search cost as a random-coefficient function of inherent and social contextual variables, we are able to unveil the nature of search cost in the online market with growing structured and unstructured business information. Second, we show the advantage of incorporating multiple and large data sources (e.g., online social media content, consumer search, clickstream, and transaction data) to efficiently estimate online product demand and uniquely identify consumer heterogeneous preferences and search cost. Third, we demonstrate the value of using structural econometric methods in analyzing emerging and important IS phenomena. By combining the optimal stopping framework with an individual-level choice model, we are able to more precisely predict consumer click and purchase probabilities on product search engines. Our dynamic model with limited consumer search can indeed "search less, but find more," providing better insights in the online search market than can a static demand model that does not account for the endogenous formation of consumers' choice sets.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 discusses our unique dataset, including the search data, transaction data and the additional social media variables extracted using text mining techniques. We also briefly discuss the preliminary model-free evidence of consumers' limited search

behaviors. In Sections 4, 5, and 6, we provide detailed discussions of our dynamic structural model for consumer sequential search, identification strategies, and empirical results, respectively. Finally, Section 7 concludes with a summary of potential insights and future directions.

Prior Literature

Our paper draws from multiple streams of work. We summarize them as the following.

Bounded Rationality and Satisficing Consumer

First, our work is related to the theory of bounded rationality and consumer satisficing behavior. Classical economic theory postulates that consumers seek to maximize their utility across different decisions. The theory of utility-maximizing choice has been the predominant framework for empirical analyses of consumer choice (e.g., McFadden 1974, Guadagni and Little 1983, Berry et al. 1995, McFadden and Train 2000). However, the assumption that a rational consumer has unlimited cognitive capabilities to acquire full information on the universal choice set has long been challenged as being inapplicable to actual human decision makers (e.g., Simon 1955, Kahneman and Tversky 1979, Johnson et al. 2004). As Simon pointed out, human beings lack the cognitive resources to maximize (Simon 1955). Instead, we make decisions only with attempts to meet an acceptability threshold—namely, following a "satisficing" process that combines "satisfy" with "suffice." Taking into account the cognitive limitations in human decision making, Simon (1955) coined the term "bounded rationality." A satisfying behavior-based model can better explain the observed limited consumer search and choice under incomplete information (e.g., Caplin, Dean and Martin 2011). It has brought renewed attention to the model of economic choice for demand estimation. In particular, recent studies have found that disregarding consumers' cognitive limitations and the limited nature of choice sets can lead to biased estimates of demand (e.g., Chiang et al. 1999, Mehta et al. 2003, Bruno and Vilcassim 2008, Kim et al. 2010, Brynjolfsson, Dick and Smith 2010).

Search Cost and Consumer Information Search

Second, our work builds on the literature on search cost and consumer information search. Since Stigler's seminal 1961 paper, consumer information search has been an important topic in both marketing and economics, trying to explain imperfect competition and information asymmetry in product and labor markets. The existing literature typically holds two different views of the nature of consumer search: non-sequential search and sequential search. The former strand of research follows Stigler's original model, assuming that consumers first sample a fixed number of alternatives and then choose the best from among them (e.g., Burdett and Judd 1983, Roberts and Lattin 1991, Mehta et al. 2003, Moraga-Gonzalez, Sandor and Wildenbeest 2011). In contrast, the other view, arising from the job-search literature (e.g., McCall 1970, Mortensen 1970), argues that the actual consumer search should follow a sequential model in which consumers keep searching until the marginal cost of an extra search exceeds the expected marginal benefit. Weitzman (1979), in single-agent scenarios, and Reinganum (1982, 1983), in multi-agent scenarios, have laid theoretical foundations for sequential search models. Recent theoretical work on modeling sequential search examines consumer search behavior and market structure from the traditional offline market to the online market (e.g., Branco, Sun and Villas-Boas 2012).

Although extensive theoretical research has been done in this field, due to model complexity and data limitations, there has been very little empirical work to date. Hong and Shum (2006) were the first to develop a structural methodology to recover search cost from price data only. Moraga-Gonzalez and Wildenbeest (2008) extend the approach of Hong and Shum to the oligopoly case and provide a maximum likelihood estimate of the search cost distribution. Both papers focus on markets for homogeneous goods, using both sequential and non-sequential search models. Hortacsu and Syverson (2004) extend this methodology to markets with differentiated goods and develop a sequential search model to recover search cost from the utility distribution. More recent empirical studies on non-sequential search tend to focus on the offline market with search frictions to study price dispersion (e.g., Wildenbeest 2011), endogenous choice sets and demand (e.g., Moraga-Gonzalez, Sandor and Wildenbeest 2011), or the identification of search cost from switching cost (Honka 2012). Recent empirical work on sequential search tries to examine consumers' limited search and the associated demand, with an initial focus on the online search market (Koulayev 2010, Kim et al. 2010). Meanwhile, de los Santos, Hortacsu and Wildenbeest (2011) use web browsing and purchasing behavior based on book price distribution across 14 online bookstores to compare to the extent to which consumers are searching under non-sequential and sequential search models.

One common practice in the existing empirical studies on both types of search models is that they typically model search cost as an inherent attribute of the consumer. Two exceptions are Kim et al. (2010), who model search cost as

a function of the product's appearance frequency on Amazon.com, and Moraga-Gonzalez, Sandor and Wildenbeest (2011), who consider explanatory variables such as geographic distance from a consumer's home to different car dealerships. In our paper, we further demonstrate that search cost should not be only an inherent attribute of a consumer, but also should be a consequence of the social context in which the consumer is embedded. By modeling consumer search cost as a random-coefficient function of the inherent and social contextual variables that capture the social environment and the search engine design, we aim to deeply examine the nature of search cost.

Search Engine Ranking

Finally, our work is also related to the literature on search engine ranking. Examining the rank position effect on the click-through rate (CTR) and conversion rate (CR) on search engines has attracted a tremendous amount of attention from the economics, marketing and computer science communities (e.g., Baye et al. 2009, Ellison and Ellison 2009, Richardson et al. 2007, Craswell et al. 2008, Chapelle and Zhang 2009). A large majority of recent studies focus on the context of search engine-based keyword advertising and find significant empirical evidence on the rank order effect (e.g., Rutz and Bucklin 2007, Ghose and Yang 2009, Goldfarb and Tucker 2011, Aggarwal et al 2011, Yao and Mela 2011). Other recent studies focus on the search engine ranking for commercial products. For example, Baye et al. (2009) use a unique dataset on clicks from one of Yahoo's price comparison sites to estimate the search engine ranking effect on clicks received by online retailers. Ellison and Ellison (2009) focus on the competition of retailers ranked on price search engines and find that the easy price search makes demand highly price-sensitive for some products. Ghose, Iperotis and Li (2012) propose a new utility gain-based ranking approach that accounts for consumers' multidimensional preferences and recommends products with the best value.

Data

We obtain our unique dataset from Travelocity.com, a major online travel search agency. The dataset contains detailed information on session-level consumer search, click and purchase events from November 2008 through January 2009, with a total of approximately one million sessions for a random sample of 2117 hotels in the United States. More specifically, a typical online session involves the following events: the initialization of the session; the search query; the results returned from that search query in a particular rank order; whether the consumer has used any special sorting criteria; the clicks on any hotels; the login and actual transactions; and the termination of the session. Notice that we also have detailed information associated with each event for every corresponding hotel, such as the displayed nightly price and hotel online position (i.e., "Page" and "Rank"). Moreover, we have the detailed transaction information from Travelocity.com that links with all the session-level consumer search data, including the final transaction price and the number of room units and nights purchased in each transaction. This linkage allows us to more precisely model consumer preferences from both the search and the purchase processes.

Meanwhile, we collect additional hotel-related information from Travelocity.com, including hotel class, hotel brand, number of amenities, number of rooms, online reviewer rating, number of reviews, and the textual content of reviews. We collect customer reviews on a daily basis up to January 31, 2009 (the last date of transactions in our database). To capture consumers' potential cognitive costs in reading the online reviews, we looked into two sets of review text features that are likely to affect consumers' intellectual efforts in digesting the review content: “readability” (i.e., complexity, syllables and spelling errors) and “subjectivity” (i.e., mean and standard deviation). Both of them have been found to have significant impacts on product sales in the past (e.g., Ghose and Ipeirotis 2010). However, it is not clear how these cognitive variables may affect consumer search cost.

To derive the probability of subjectivity in the review's textual content, we apply the text mining techniques (e.g., Ghose and Ipeirotis 2010). In particular, we train a classifier using as “objective” documents the hotel descriptions of each of the hotels in our dataset. We randomly retrieved 1000 reviews to construct the “subjective” examples in the training set. We conduct the training process by using a 4-gram Dynamic Language Model classifier provided by the LingPipe toolkit[1]. Thus, we are able to acquire a subjectivity confidence score for each sentence in a review, and then derive the mean and variance of this score, which represent the probability of the review being subjective.

In addition, we also have supplemental data on hotel location-related characteristics collected independently. We only briefly discuss them here. We use geo-mapping search tools (in particular the Bing Maps API) and social geo-tags (from geonames.org) to identify the number of external amenities (e.g., shops, bars, etc) in the area around the hotel. We use image classification together with human annotations (from Amazon Mechanical Turk, AMT) to

[1] http://alias-i.com/lingpipe/

examine whether or not there is a nearby beach, lake or downtown area, and whether the hotel is close to a highway or public transportation. We extract these characteristics within an area of 0.25-mile, 0.5 mile, 1-mile, and 2-mile radius. We also collect local crime rate from FBI statistics. For a better understanding of the variables in our setting, we present the definitions and summary statistics of all variables in Table 1.

Table 1. Definitions and Summary Statistics of Variables

Variable	Definition	Mean	Std. Dev.	Min	Max
PRICE_DISP	Displayed price per room per night	230.98	179.76	16	2849
PRICE_TRANS	Transaction price per room per night	148.08	108.18	52	2252
COMPLEXITY	Average sentence length per review	17.50	3.77	4	44
SYLLABLES	Average # syllables per review	246.81	50.53	76	700
SPELLERR	Average # spelling errors per review	1.17	.33	0	3.86
SUB	Review subjectivity - mean	.91	.03	.05	1
SUBDEV	Review subjectivity - standard deviation	.02	.03	0	.25
CLASS	Hotel class	3.62	.70	1	5
AMENITYCNT	Total # hotel amenities	14.37	6.22	2	23
ROOMS	Total number of hotel rooms	210.12	258.27	12	2900
REVIEWCNT	Total # reviews	13.56	25.60	0	202
RATING	Overall reviewer rating	3.94	.39	1	5
PAGE	Page number of the hotel	20.86	13.44	1	192
RANK	Screen position of the hotel	12.09	4.32	1	25
SPECIALSORT	Dummy for a special sorting method	.10	.30	0	1
BEACH	Beachfront within 0.6 miles	.19	.36	0	1
LAKE	Lake or river within 0.6 miles	.23	.44	0	1
TRANS	Public transportation within 0.6 miles	.31	.45	0	1
HIGHWAY	Highway exits within 0.6 miles	.70	.42	0	1
DOWNTOWN	Downtown area within 0.6 miles	.66	.45	0	1
EXTAMENITY	Number of external amenities within 1 mile,	4.63	7.99	0	27
CRIME	City annual crime rate	194.99	127.22	3	1310
BRAND	Dummies for 9 hotel brands: Accor, Best	--	--	0	1
Total # Sessions:	**969,033**	**Time Period:**	**11/1/2008-1/31/2009**	**Total # Hotels:**	**2117**

Model-Free Evidence of Limited Search by Consumers

Before we propose our model, we seek from the data any direct evidence that supports our assumption of consumers' limited search. First, we plot the distribution of the total number of pages a consumer browses in her search session. Figure 1 illustrates this distribution in detail, with the x axis representing the page counts and the y axis representing the density. We notice that over 25% of consumers browse only one page; over 50% of consumers browse less than three pages; and less than 10% of consumers browse more than 15 pages during their search for hotels. This finding is consistent with prior industry evidence that consumers seldom search more than three pages (e.g., Iprospect. 2008). Second, we further look into the distribution of the average number of click-throughs made per page during each search session. Figure 2 illustrates this distribution, with the x axis representing the click-throughs per page and the y axis representing the density. We find that, on average, consumers click less than one hotel (out of a total of 25 hotels) per page during their search. Moreover, a large majority of consumers click even less than 0.5 hotels per page, on average. This finding seems to imply that consumers' search costs are considerably high and that consumers only selectively devote their efforts to investigating a small subset of choices. These two figures provide us with preliminary evidence that consumers' search costs indeed exist and that consumer search is highly limited. Consumers are not able to obtain complete information on products, which contradicts the assumptions made by the traditional demand estimation approaches.

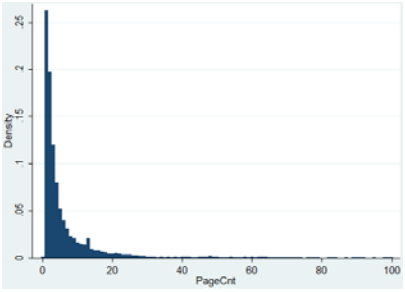

Figure 1. Distribution of # Pages Browsed (Session Level)

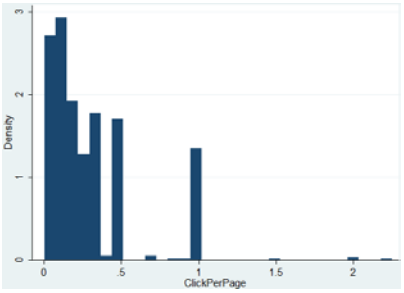

Figure 2. Distribution of # Click-thoughs Per Page (Session Level)

A Dynamic Structural Model of Consumer Sequential Search

In this section, we discuss how we design our dynamic structural model for consumer sequential search based on an optimal stopping framework (Weitzman 1979) and then combine it with an individual-level random utility choice model to jointly estimate consumer heterogeneous preferences and search cost. The key advantage of our proposed model is not only that it captures the dynamics of consumers' search and click-through behaviors, but also that, with the detailed transaction information, it captures consumers' final purchase decisions under limited search. Moreover, by modeling consumer search cost as a random-coefficient function of inherent and social contextual variables, we are able to deeply examine the nature of search cost.

We assume that consumers search sequentially on product search engines. The sequential search assumption is the basis of our model. Although the existing literature holds two different views of the nature of consumer search—non-sequential search and sequential search—we believe that the sequential approach is a closer match for *online* consumer search. This assumption is consistent with the mainstream research by the web search community and major search engine companies (e.g., Richardson et al. 2007, Craswell et al. 2008, Chapelle and Zhang 2009). In addition, many recent studies in economics and marketing have also adopted the sequential search strategy for examining consumer search in an online environment (e.g., Kim et al. 2010, Koulayev 2010, Branco, Sun and Villas-Boas 2010).

Model Setting

(1) Product Utility.

Assume the utility of product *j* for consumer *i* to be a random-coefficient model as follows:

$$u_{ij} = V_{ij} + e_{ij}, \tag{1}$$

where $V_{ij} = V_{ij}^S + V_{ij}^L$ represents the expectation of the overall product utility. It consists of two parts [2]: the expected utility from "summary-page" product characteristics that consumers can directly observe on the search result summary page, V_{ij}^S, and the expected utility from "landing-page" product characteristics that consumers can only observe after clicking the hotel and arriving at the hotel's landing page, V_{ij}^L.

Let X_j be a vector of summary-page characteristics for product *j*. In our study, X_j includes *Hotel Class*, *Hotel Brand, Customer Rating* and *Total Review Count*. Let P_j represent the *Price* for product *j* that is also directly available to consumers on the search result summary page. Thus, we can model the expected summary-page utility as $V_{ij}^S = X_j\beta_i - \alpha_i P_j$, where β_i and α_i are consumer-specific parameters capturing the heterogeneous preferences

[2] We have also tried an alternative model where the overall expected utility contains only V_{ij}^L, meaning that a consumer can only reveal the product utility after the click-through and the choice set contains only products that are clicked. We estimate this alternative model accordingly and find the results are very consistent. Due to space limitation, we do not provide the results in this paper. They are available from the authors upon request.

of consumers. Consistent with the prior literature (e.g., Kim et al. 2010), we assume that $\beta_i \sim N(\overline{\beta}, \sum_\beta)$ where $\overline{\beta}$ is a vector containing the means of the random effects and $\sum_\beta$ is a diagonal matrix containing the variances of the random effects. Moreover, we assume that $\alpha_i \sim N(\overline{\alpha}, \sigma_\alpha^2)$.

Similarly, we can model the expected landing-page utility as $V_{ij}^L = L_j \lambda_i$, where L_j represents a vector of landing-page characteristics for product *j*. In the estimation, L_j includes *Total Amenity Count, Total Number of Rooms, Total Number of External Amenities, Beach, Lake, Downtown, Highway, Public Transportation* and *Crime Rate*. λ_i represents consumer-specific parameter capturing the heterogeneity. Consistent with previous assumptions, it follows a normal distribution $\lambda_i \sim N(\overline{\lambda}, \sum_\lambda)$.

Thus, the overall utility function can be written as

$$u_{ij} = X_j \beta_i - \alpha_i P_j + L_j \lambda_i + e_{ij}. \quad (2)$$

Note that e_{ij} represents the unknown stochastic error during the consumer's decision process. It is assumed to be i.i.d. across consumers and products. For estimation tractability, we assume it to follow a Type I Extreme Value distribution $e_{ij} \sim Type\ I\ EV(0,1)$.

(2) Search Cost.

Meanwhile, consumers have cognitive limitations in searching and evaluating choices in the decision- making process. Consequently, consumers' choice sets are limited and endogenously formed in the search market. According to Simon's theory of bounded rationality (Simon 1955), cognitive cost may occur due to decision makers' limitations of time, knowledge and cognitive capacity. In the online environments, extensive prior literature investigates the factors that influence the complexity and effectiveness of web-based information systems (e.g., Germonprez and Zigurs 2005, Hauser et al. 2009). Theoretical framework is developed to examine three dimensions (Germonprez and Zigurs 2005): *content* (e.g., amount of information (Schneider 1987)), *form* (e.g., user interface, navigation and structure), and *cognition/user perception* (e.g., orientation as website coherence via hypertext links, orientation as cognitive overhead via the amount of information (Thuring et al. 1995), perceived usefulness and perceived ease of use (Davis et al. 1989)).

Therefore, we model consumers' search costs to account for these three dimensions in the evaluation of product-related information, including both the *structured product information* (such as seller-provided product descriptions) and the *unstructured product information* (such as social content generated by the online communities). Meanwhile, eye-tracking studies have shown that consumers tend to scan the search results in order (e.g., Aula and Rodden 2009), and visual attention influences consumer choice (Pieters and Warlop 1999). Thus, the product's *online screen position* can also have a significant effect on consumer search cost.

Note that since in our study the design of each product landing page on the search engine is identical, each providing the same user interface, navigation, structure, hypertext links and website coherence, etc. Because our goal is to examine the variation in the search costs, we focus only on the variables that vary along the above three dimensions. More specifically, we focus mainly on the content dimension and examine the amount and complexity of product-related information. We use the *Total Amenity Count* to approximate the structured product information. Regarding the unstructured product information, we use the *Total Review Count*, *Review Readability* (complexity, syllables and spelling errors) and *Review Subjectivity* (mean and standard deviation) for approximation. In addition, we use the *Page Number*, *Rank Order* and *Whether The Search Results Are Specially Sorted* in a particular consumer's search session (i.e., not under the default ranking) to capture the online position effect. Meanwhile, we assume consumer search cost to follow a log-normal distribution. Taking into consideration consumer heterogeneity, we model the search cost of consumer *i* for product *j* to be a random-coefficient function as follows:

$$\begin{aligned} c_{ij} = \exp(&\gamma_{0i} + \gamma_{1i} PAGE_j + \gamma_{2i} RANK_j + \gamma_{3i} SPECIALSORT_{ij} + \gamma_{4i} AMENITYCNT_j + \gamma_{5i} REVIEWCNT_j \\ &+ \gamma_{6i} COMPLEXITY_j + \gamma_{7i} SYLLABLES_j + \gamma_{8i} SPELLERR_j + \gamma_{9i} SUB_j + \gamma_{10i} SUBDEV_j), \end{aligned} \quad (3)$$

where $\gamma_i \sim N(\bar{\gamma}, \Sigma_\gamma)$, $\bar{\gamma}$ is a vector containing the means of the random effects and Σ_γ is a diagonal matrix containing the variances of the random effects. [3] Our final goal is to estimate the parameters of the random coefficients from equations (2) and (3):

$$\{\theta\} = \left\{(\bar{\alpha}, \sigma_\alpha),\ (\bar{\beta}, \Sigma_\beta),\ (\bar{\lambda}, \Sigma_\lambda),\ (\bar{\gamma}, \Sigma_\gamma)\right\}. \tag{4}$$

Problem Description and the Optimal Search Framework (Weitzman 1979)

In general, our consumer search problem can be described as follows. Assume that a consumer searches sequentially (i.e., examines alternatives one by one) to find a product. At each stage of the search, the consumer has two options (actions): to continue to search for the next alternative or to stop and choose the current best alternative. Consider that the consumer is forward-looking. This implies that at any stage during her search, she always tries to choose an action that maximizes her *expected utility from the current stage going forward*—meaning that she tries to maximize the marginal benefits from both the current stage and all potential future stages. Therefore, the key problem here is to determine the consumer's "optimal stopping point."

Our solution to this problem builds on Weitzman's (1979) optimal sequential search framework. Weitzman proposed an optimal stopping rule in which alternatives are ranked in descending order of their *reservation utility*. This value indicates a "rate of return" from searching each alternative (we will formally define it shortly). A consumer searches sequentially according to the ranking list. She stops searching if the utility from the current best alternative exceeds the reservation utility of the next best alternative. Otherwise, she continues to search the next alternative in the ranking and repeats the process until she finds an alternative that meets the stopping criterion.

Reservation utility plays an important role in this model framework. It is defined as the utility value for an alternative at which the consumer would be *indifferent* between searching the alternative at a certain cost or accepting this utility value (and stopping). In other words, the reservation utility is the value that satisfies the boundary condition where the marginal cost of searching an extra alternative equals the expected marginal benefits. If the consumer already has an item of higher utility, she should stop since the expected marginal benefits from search are less than the cost. If the consumer does not have a utility as high as the forthcoming reservation utility in the ranking list, she should continue to search because the expected marginal benefits will exceed the expected cost.

More formally, let u_i^* be the current highest utility searched by consumer *i* so far. Let z_{ij} be the reservation utility of product *j* for consumer *i*, and let *J* be the total number of products available in the market. Thus, for each consumer *i*, rank products in descending order of their reservation utility z_{ij}, denoted by $r_i(1)...r_i(J)$.

$$z_{i,r_i(1)},\ z_{i,r_i(2)},\ z_{i,r_i(3)},\ldots,\ z_{i,r_i(j)},\ \ldots z_{i,r_i(J)} \tag{5}$$

Note that, intuitively, ranking products by their reservation utility implies how "desirable" these products appear to consumer *i*. According to Weitzman's "selection rule" (1979), consumer *i* searches sequentially from the product with the highest reservation utility, $z_{i,r_i(1)}$, to the lowest, $z_{i,r_i(J)}$ in the ranking list.

Given the current best utility u_i^*, the expected marginal benefits for consumer *i* from searching product *j* are

$$B_{ij}(u_i^*) = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*)\, f(u_{ij})\, du_{ij}, \tag{6}$$

where $f(\square)$ is the probability density function of product utility u_{ij}. These expected marginal benefits $B_{ij}(u_i^*)$ represent the expectation of the utility for product *j*, given that it is higher than u_i^*, multiplied by the probability that u_i exceeds u_i^*. As we notice, the benefits of search depend only on the distribution of utility above u_i^*.

We know that the reservation utility z_{ij} meets the following boundary condition, where the marginal search cost c_{ij} equals the expected marginal benefits from searching product *j*.

---

[3] The log-normal assumption of search cost is consistent with the prior literature (e.g., Kim et al. 2010, Wildenbeest 2011). In addition, we were able to theoretically demonstrate that the log-normally distributed search cost and Type I EV distributed product utility together lead to a power-law distributed click probability, which dovetails with what is observed in reality. The proof is available from the authors upon request.

$$c_{ij} = B_{ij}(z_{ij}) = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij}. \tag{7}$$

Therefore, when consumer *i*'s current best utility is equal to the reservation utility of product *j*, $u_i^* = z_{ij}$, she is indifferent between searching for *j* or stopping (and accepting u_i^*). Consumer *i* will continue to search for product *j* if her current best utility is lower than the reservation utility of product *j*, $u_i^* < z_{ij}$, and she will stop otherwise. [4]

Click Probability

We define the click probability in a fashion similar to (Kim et al. 2010). Let $r(j)$ denote the product with the jth highest ranked reservation utility $z_{i,r(j)}$. Let $\pi_{i,r(j)}$ be the probability that consumer *i* will click product $r(j)$. This probability equals the probability that the current highest utility among all the previously "searched" *j-1* products (meaning those products that consumers either click or observe on the search result summary page) is lower than the reservation utility of product $r(j)$. Thus, we model the click probability of product $r(j)$ for consumer *i* as

$$\begin{aligned} \pi_{i,r(j)} &= \Pr\left[r(j) \text{ is clicked by consumer } i \right] \\ &= \Pr\left[\max_{m=1}^{j-1} (V_{i,r(m)} + e_{i,r(m)}) < z_{i,r(j)} \right] = \prod_{m=1}^{j-1} F_e(z_{i,r(j)} - V_{i,r(m)}), \quad j > 1, \end{aligned} \tag{8}$$

where $F_e(\square)$ is the CDF of e_{ij}, which in our case is $e_{ij} \sim TypeI\ EV(0,1)$.

Conditional Purchase Probability

Product $r(j)$ is purchased by consumer *i* if and only if consumer *i* stops searching and chooses $r(j)$ over everything else within the choice set. Thus, the following two conditions must be met:

1) The utility of $r(j)$ is greater than the reservation utility of any other product that has not been searched for;
2) The utility of $r(j)$ is greater than the utility of any other product that has already been searched for.

Let S_{i,N_i} be the search-generated optimal choice set of size N_i for consumer *i*. Thus, we can model the purchase probability of product $r(j)$ for consumer *i* as

$$\begin{aligned} \eta_{i,r(j)} &= \Pr\left[r(j) \text{ is purchased by consumer } i \right] \\ &= \Pr\left[(V_{i,r(j)} + e_{i,r(j)}) > z_{i,r(m)},\ r(m) \notin S_{i,N_i} \right] \times \Pr\left[(V_{i,r(j)} + e_{i,r(j)}) > (V_{i,r(k)} + e_{i,r(k)}),\ r(k) \in S_{i,N_i} \right] \\ &= \prod_{m=N_i+1}^{J} \left(1 - F_e(z_{i,r(m)} - V_{i,r(j)})\right) \times \frac{\exp(V_{i,r(j)})}{1 + \sum_{k=1}^{N_i} \exp(V_{i,r(k)})}. \end{aligned}$$

(Note that the mean utility for outside good $r(0)$ is normalized to zero, $V_{i,r(0)}$=0.) (9)

Joint Probability of Click and Purchase

Finally, to account for the consumer's click and purchase decisions, given the dynamic formation of the choice set, we examine the joint probability of all the click and purchase events in that consumer's online session. More specifically, define $\omega_{i,r(j),N_i}$ as the joint probability that consumer *i* has clicked N_i products and then purchased product $r(j)$. Thus, we can model this joint probability as the following.

$$\omega_{i,r(j),N_i} = \Pr\left[r(1)...r(N_i) \text{ are clicked by consumer } i,\ r(j) \text{ is purchased by consumer } i,\ 0 \le j \le N_i \right]$$

[4] Due to page limitation, we refer interested readers to our online appendix for more details on the derivation of the optimal search strategy at http://pages.stern.nyu.edu/~bli/ICIS2012app.pdf.

$$= \left(\prod_{k=1}^{N_i} \pi_{i,r(k)} \right) \times \eta_{i,r(j)}. \tag{10}$$

Log-Likelihood Function

Based on all of the above, we can derive the overall likelihood function of each consumer searching for and purchasing each product as what we observed from the data in the following way:

$$Likelihood(\theta) = \prod_{i=1}^{I} \prod_{j=0}^{J} \left(\omega_{i,r(j),N_i} \right)^{y_i}, \tag{11}$$

where I is the total number of consumers. $y_i = 1$ if the consumer has clicked and purchased product $r(j)$; $y_i = 0$ otherwise. Correspondingly, the overall log-likelihood function is

$$LL(\theta) = \sum_{i=1}^{I} \sum_{j=0}^{J} \left[y_i \ln(\omega_{i,r(j),N_i}) \right]. \tag{12}$$

Identification

One of the major challenges in the dynamic search demand estimation is how to simultaneously identify consumers' heterogeneous preferences and search cost. As pointed out by Sorensen (2001) and Hortacsu and Syverson (2004), explaining search decisions by consumers with heterogeneous preferences imposes an identification problem. A person may stop searching either because she has a high valuation for the products already found or because she has a high search cost. Therefore, an observed search outcome can be explained either by the preferences for product characteristics or by the moments of the search cost distribution (Koulayev 2010). It is important to understand how these two causes can be uniquely recovered and what type of data are needed for the empirical identification.

In our proposed model, there are four major effects that need to be identified: Consumer Preferences (Mean and Heterogeneity) and Consumer Search Cost (Mean and Heterogeneity). The key identification strategy of our estimation relies on the fact that consumer preferences enter the decision-making processes of both search and purchase, whereas consumer search cost enters only the search decision-making process. Once the consideration set is generated after search, the conditional purchase decision should depend only on the consumer preferences. Our unique dataset containing both consumer search data and purchase data allows us to successfully identify these effects. We provide more detailed discussions below.

(1) Mean Consumer Preferences.

The mean preferences for product characteristics are identified by the correlation between the click and purchase frequencies of products and the frequencies of underlying products' characteristics. We measure the mean effect of a product characteristic by how often the same (or similar) characteristic appears in the products that are clicked or purchased by consumers. This identification is similar to the one in most traditional choice models, except that it takes into consideration not only the observed purchases, but also the clicks, to infer consumer mean preferences.

(2) Heterogeneous Consumer Preferences.

We identify consumer heterogeneous preferences from two perspectives. First, we partially identify them from the search data by the discrepancy between our model's predicted click probabilities, based solely on the mean consumer preferences, and the observed click probabilities. Moreover, since we also observe consumers' final purchases, these purchase data allow us to identify the heterogeneous preferences by the discrepancy between the model's predicted purchase probabilities, based solely on the mean consumer preferences, and the observed purchase probabilities. Notice that the latter source provides us an opportunity to uniquely recover consumer heterogeneous preferences from the heterogeneous search cost because once the consideration set is generated after search, the conditional purchase decision should depend only on consumer preferences.

(3) Mean Consumer Search Cost.

The mean search cost is partially identified by the observed average size of the consumer's search-generated consideration set. Meanwhile, note that we model the search cost as a function of different characteristics (e.g., product online position, the amount and complexity of social media content), which can be viewed simply as additional product characteristics. Thus, similar to the identification of consumer mean preferences, we can identify

the mean search cost coefficients by the correlation between the observed click frequencies and the frequencies of underlying search cost characteristics.

(4) Heterogeneous Consumer Search Cost.

Finally, we identify the heterogeneous search cost through two sources. First, given that consumer heterogeneous preferences are identified through the conditional purchase probabilities, we can then identify the heterogeneous search cost by the joint variation of the consideration set size and the click probabilities. In addition, as Kim et al. (2010) point out, the nonlinear functional form in the reservation utility (i.e., equation (7)) can also help identify consumer preference and search cost parameters. Since the consumer preferences enter the equation in a nonlinear manner (i.e., need to integrate over the utility), whereas the search cost enters the equation in a linear manner, this mathematical nonlinearity helps us separately identify consumer heterogeneous preferences and search cost.

Estimation Results

We iteratively estimate the model using a Maximum Simulated Likelihood (MSL) method. In particular, we apply the Monte Carlo method for numerical simulation, where for each individual observation, we simulate 250 random draws from the joint distribution of the individual heterogeneous parameters $\{\theta\}$ and compute the corresponding individual-level joint probability $\omega_{i,r(j),N_i}$. Then, we construct the objective function—the overall log-likelihood $LL(\theta)$. To maximize this function, we choose to use a non-derivative-based optimization algorithm (i.e., the Nelder-Mead simplex method) for heuristic search[5]. This procedure iteratively searches for the optimal set of parameters $\{\theta^*\}$ until the log-likelihood function $LL(\theta)$ is maximized.

$$\{\theta^*\} = \arg\min_{\{\theta^*\}} \sum_{i=1}^{I} \sum_{j=0}^{J} \left[y_i \ln(\omega_{i,r(j),N_i}) \right]. \tag{13}$$

The main computational complexity of the estimation comes from the calculation of the reservation values. During each iteration of the optimization algorithm, for each observation and each value of the search cost, we need to solve $z_{ij} = B_{ij}^{-1}(c_{ij})$ numerically. To improve the estimation efficiency, we apply an interpolation-based method to compute the reservation values (Kim et al. 2010, Koulayev 2010). The main results are shown in Table 2 column 2.

Discussion

First, we find that the majority of the coefficients are statistically significant at the p ≤ 5% level, including both the mean effects ($\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$, $\overline{\gamma}$) and the heterogeneity (σ_α, Σ_β, Σ_λ, Σ_γ). Consistent with theory, *PRICE* has a negative effect on hotel demand. *CLASS*, *AMENITYCNT*, *ROOMS*, *RATING* and *REVIEWCNT* each has a positive effect on hotel demand. For location-related hotel characteristics, consistent with Ghose et al (2012), we find that *BEACH, TRANS, HIGHWAY, DOWNTOWN* each has a positive effect on hotel demand, whereas *LAKE* and *CRIME* each shows a negative effect. Meanwhile, we find that online screen position has significant effects on consumer search cost. In particular, *PAGE* and *RANK* both can lead to an increase in the search cost.

Interestingly, we find that *SPECIALSORT* has a negative mean effect on consumer search cost, while also showing a large heterogeneity. This result suggests that, on average, when consumers sort the search results by themselves using the ranking recommendation algorithms provided by the product search engines, it helps them to reduce search costs by making the attractive products more visible. However, if the ranking is generally bad, or the top-ranked products are not satisfactory, such sorting action may have an opposite effect and lead to an increase in consumer search cost. This finding highlights the importance of search engine ranking design.

With regard to the cognitive variables that measure the amount and complexity of product information, we find that both the seller-provided structured information and the social media-related unstructured information lead to an increase in consumer search cost. More specifically, *AMENITYCNT* and *REVIEWCNT* both show a positive sign, implying that the more product features or the more feedback from online social communities for a hotel on search

[5] For a robustness check, we also tried the derivative-based optimization algorithms (e.g., the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and the Nested Fixed Point algorithm (NFXP)). We found that different optimization algorithms can recover consistent structural parameters in our case.

engines, the higher cognitive costs it requires for consumers to search and evaluate that hotel. Meanwhile, *COMPLEXITY*, *SYLLABLES* and *SPELLERR* each show a positive sign, suggesting that consumers' abilities in digesting the textual content of social media information is limited. Long sentences, complex words or spelling errors may discourage consumers from continuing to search on product search engines. Moreover, *SUB* and *SUBDEV* show a positive sign, implying that subjective content and an inconsistent, sentiment writing style create a cognitive burden for consumers during product search and may lead to early termination of their search.

To acquire a better intuition of the search cost, we quantitatively derive the dollar value of different search cost variables. This dollar value represents how much a certain variable effect can be translated into price. We find that, on average, the effort of continuing to search an additional page costs $39.15, while the effort of continuing to search an additional screen position on the same page costs $6.24. A good ranking recommendation can, on average, save consumers $9.38. However, a bad ranking recommendation can lead to an $18.54 loss for consumers. Meanwhile, a one-word increase in the average sentence length costs consumers $2.73 to digest the review content on the product search engine. One more syllable or one more spelling error per review can cost consumers $3.77 or $1.60, respectively, during the product search. One more amenity displayed on the product search engine increases search cost by $1.00, and one more customer review increases consumer search cost by $1.17.

Robustness Checks

To assess the robustness of our estimation model and results, we conduct three robustness tests:

1) Robustness Test I: Exclude the social media variables from the search cost specification.

One of the main goals in our paper is to examine how the amount and complexity of product-related social media content affect consumer search cost. So, we are interested in comparing the differences in the search models with and without the set of social media variables. The results of this test are illustrated in Table 2, columns 3. First, we find that the estimated coefficients are qualitatively consistent with the main results. Meanwhile, we notice that the model that does not account for social media cognitive variables presents a significantly higher magnitude in both the mean effect and the heterogeneity from price (1.917 vs. 1.406 and 0.735 vs. 0.427). This result indicates that consumers' cognitive costs to digest social media content during online product search are non-negligible. Failing to account for such costs can lead to an overestimation of price sensitivity in the online search market.

2) Robustness Test II: Use an alternative static model with actual (limited) choice set.

To examine the potential bias from the endogenous and limited nature of search-generated choice sets, we consider one competitive model that is widely used in the static demand estimation: the Mixed Logit model (e.g., McFadden and Train 2000). Moreover, to account for the variation in choice sets, we model the consumer decision process under the actual searched (limited) choice set, rather than under the universal choice set available in the market. Note that the major difference between a static Mixed Logit model with actual choice sets and our proposed model is that our model captures not only the limited nature of the choice sets, but also the dynamic and endogenous formation process of the choice sets. However, a static model takes the choice set as exogenously given.

Interestingly, we notice that using a static model without accounting for consumers' dynamic search behaviors can lead to a significant overestimation of the price coefficient. The interpretation of this finding can be attributed to the nature of the hotel search market. *A model that captures consumers' actual search behaviors finds lower price sensitivity, implying that consumers in the hotel search market tend to highly evaluate the quality of hotels and put weight on non-price factors during search (e.g., class, amenities or reviews).* Our finding on price sensitivity is consistent with prior findings by Koulayev (2010) and Brynjolfsson et al. (2010). Both studies show that when consumers face a highly differentiated market (e.g., product differentiation or retailer differentiation), they are more likely to focus on non-price factors during search. Hence, the estimated price elasticity of demand is lower when incorporating consumers' search behaviors into the model. On the contrary, when a market is less differentiated, consumers become more price-sensitive and tend to focus on price search. Thus, a dynamic model that incorporates consumers' search behaviors may find a higher price elasticity of demand than a static model does (e.g., de los Santos et al. 2011). The results of this robustness test are shown in Table 2, columns 4.

3) Robustness Test III: Examine the interaction effects between consumer travel purposes and sorting methods.

One advantage of this dynamic structural model is that it can account for consumer heterogeneity during the search process. Under the context of hotel search, we are interested in how certain variation in the search cost can be explained by consumers' choices of different sorting methods under heterogeneous travel purposes. To do so, we investigate the interaction effects between consumer travel purposes and sorting criteria on search cost.

First, to capture consumers' heterogeneous travel purposes, we define T_i as an indicator vector with identity components representing the travel purpose:

$$T_i' = [Family_i \ Business_i \ Romance_i \ Tourist_i \ Kids_i \ Senior_i \ Pets_i \ Disability_i]_{1\times 8}. \quad (14.1)$$

We acquire the empirical distribution of T_i from online consumer reviews and reviewers' profiles [6].

Second, to capture the effects from different sorting methods, we break down the scalar dummy variable *SPECIALSORT* into an indicator vector with identity components representing the use of different sorting methods. In particular, we observe six different sorting criteria that consumers use during their searches: default (DFT), price ascending (PRA), class descending (CLD), class ascending (CLA), city name (CNA) and hotel name (HNA). Let S_{ij} denote the indicator vector of sorting method under which product j is presented to consumer i at the moment of his/her search:

$$S_{ij}' = [DFT_{ij} \ PRA_{ij} \ CLD_{ij} \ CLA_{ij} \ CNA_{ij} \ HNA_{ij}]_{1\times 6}. \quad (14.2)$$

Thus, we can extend the basic model of search cost to the following

$$\begin{aligned} c_{ij} = \exp(&\gamma_{0i} + \gamma_{1i} PAGE_j + \gamma_{2i} RANK_j + \Gamma T_i \times S_{ij}' + \gamma_{4i} AMENITYCNT_j + \gamma_{5i} REVIEWCNT_j \\ &+ \gamma_{6i} COMPLEXITY_j + \gamma_{7i} SYLLABLES_j + \gamma_{8i} SPELLERR_j + \gamma_{9i} SUB_j + \gamma_{10i} SUBDEV_j), \end{aligned} \quad (15)$$

where everything else remains the same as that in equation (3), except that Γ is a 8×6 matrix of coefficients that measures how consumers' taste parameters vary with different travel purposes and choices of sorting criteria. The estimation results of interaction effects are illustrated in Tables 3.

We find that consumers' travel purposes can explain their heterogeneous search costs under different ranking mechanisms. In general, the default ranking (DFT) can reduce search costs for different consumers. This reduction appears to be the largest for consumers who plan to travel with their families (i.e., -2.452), followed by business travelers (i.e., -1.757), romance travelers (i.e., -1.289) and tourists (i.e., -0.836). However, no significant interaction effects are found for consumers who travel with young kids, seniors, or pets. This finding seems to indicate that the current default ranking captures mainly consumers' preferences under the most common travel contexts. The default ranking may not be the most effective when consumers are seeking for certain special amenities during travel search.

Meanwhile, the price ascending ranking (PRA) can present significant interaction effects in opposite directions for different consumers. It decreases the search costs for tourists (i.e., -1.869), family travelers (i.e., -1.007) and senior travelers (i.e., -0.537), while it increases the search costs for romance travelers (i.e., 1.203) and business travelers (i.e., 0.989). This finding is consistent with Ghose et al. (2012), indicating that romance and business travelers are less sensitive to price, whereas tourists tend to be the most price-sensitive.

Furthermore, ranking by hotel class does not seem to be effective with regard to reducing search costs. Class descending ranking (CLD) leads to a significant increase in the search costs for business travelers (i.e., 1.073), family travelers (i.e., 0.780) and travelers with young kids (i.e., 0.204). Whereas, class ascending ranking (CLA) leads to a significant increase in the search costs for romance travelers (i.e., 3.030) and family travelers (i.e., 1.291). This finding seems to suggest that starting with similar hotels from either the luxury end or the cheap end may not be informative for consumers during the search. Consumers are willing to explore products with better variety (e.g., Agichtein et al. 2006, Ghose et al. 2012), especially at the beginning of the search.

Interestingly, we find that hotel name ranking (HNA) can save a significant amount of search costs for different travelers. Under this ranking mechanism, the search costs decrease the most for business travelers (i.e., -2.076), followed by senior travelers (i.e., -0.701) and romance travelers (i.e., -0.417). This finding indicates that hotel names (or brands) can significantly reduce consumers' search costs under certain contexts. For example, business travelers may seek directly either cooperative partners or particular hotels that are recommended by the business events. Seniors travelers may prefer special hotel chains and tend to search for them directly. [7]

---

[6] After writing an online review for a hotel, a reviewer is asked to provide additional demographic and trip information—e.g., "What was the main purpose of this trip? (Select one from the eight choices.)" The distribution of T_i is derived based on reviewers' responses to this question. Our robustness test shows that consumers' demographics derived from different online resources stay consistent (Jensen-Shannon divergence D = 0.03).

[7] This finding is consistent with Ghose et al. (2012) where the authors found that senior travelers have a special preference for Best Western hotels.

Table 2. Estimation Results - Main Results, Robustness Tests (I) & (II) Results

Variable	Mean Effect (Std. Err)^{M}	Heterogeneity (Std. Err)^{M}	Mean Effect (Std. Err) ^{R1}	Heterogeneity (Std. Err) ^{R1}	Mean Effect (Std. Err)^{R2}	Heterogeneity (Std. Err)^{R2}
(Preferences)	$\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$	σ_{α} , Σ_{β} , Σ_{λ}	$\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$	σ_{α} , Σ_{β} , Σ_{λ}	$\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$	σ_{α} , Σ_{β} , Σ_{λ}
PRICE$^{(L)}$	-1.423* (.000)	0.578* (.023)	-1.925*(.001)	0.740* (.001)	-2.531* (.021)	1.137* (.019)
CLASS	1.667* (.002)	1.377* (.087)	1.729* (.003)	1.702* (.004)	2.023* (.062)	2.010* (.015)
RATING	3.199* (.003)	1.923* (.021)	3.543* (.007)	1.188* (.005)	3.776* (.038)	1.344* (.032)
AMENITYCNT$^{(L)}$	.053* (.006)	.004(.032)	.076* (.003)	.007(.040)	.115* (.023)	.019(.102)
REVIEWCNT$^{(L)}$	1.411* (.003)	1.405* (.090)	1.599* (.006)	1.211* (.004)	1.878* (.031)	0.624* (.021)
ROOMS$^{(L)}$	1.005* (.002)	.056(.071)	1.336* (.023)	.049(.056)	1.602* (.106)	.077(.110)
EXTAMENITY$^{L)}$	.082* (.001)	.005(.024)	.064* (.011)	.014(.033)	.089* (.035)	.058(.097)
BEACH	1.001* (.010)	.072* (.012)	1.545* (.012)	.081* (.022)	1.892* (.001)	.063* (.018)
LAKE	-.767* (.089)	1.356* (.059)	-.702* (.065)	1.203* (.044)	-1.005* (.047)	1.986* (.263)
TRANS	1.046* (.003)	.043* (.029)	1.067* (.008)	.068(.067)	1.288* (.142)	.089(.211)
HIGHWAY	.602* (.091)	.070* (.005)	.559* (.076)	.043* (.013)	.304* (.060)	.066(.053)
DOWNTOWN	.586* (.004)	.116* (.047)	.534* (.017)	.123* (.052)	.707* (.196)	.283* (.075)
CRIME	-.112* (.001)	.017(.036)	-.179* (.006)	.010(.049)	-.181* (.083)	0.037(.102)
BRAND	Yes					
(Search Cost)	$\overline{\gamma}$	Σ_{γ}	$\overline{\gamma}$	Σ_{γ}	$\overline{\gamma}$	Σ_{γ}
Search Base Cost	-2.287* (.003)	1.463* (.004)	-2.531* (.005)	1.620* (.015)	----	----
PAGE	4.017* (.002)	1.633* (.003)	3.598* (.012)	1.147* (.002)	----	----
RANK	2.178* (.006)	0.340* (.001)	2.241* (.011)	0.276* (.001)	----	----
SPECIALSORT	-2.582* (.011)	5.835* (.023)	2.103* (.014)	4.669* (.024)	----	----
AMENITYCNT$^{(L)}$	0.343* (.005)	0.146* (.001)	0.389* (.006)	0.158* (.001)	----	----
REVIEWCNT$^{(L)}$	0.500* (.008)	0.211* (.005)	----	----	----	----
COMPLEXITY	1.349* (.011)	0.142* (.006)	----	----	----	----
SYLLABLES$^{(L)}$	1.668* (.015)	0.378* (.010)	----	----	----	----
SPELLERR$^{(L)}$	0.814* (.005)	0.290* (.008)	----	----	----	----
SUB	0.205* (.002)	0.079* (.001)	----	----	----	----
SUBDEV	0.822* (.019)	0.102* (.007)	----	----	----	----
Maximum LL	477587.023619		477342.002341		125786.702515	

(L) Logarithm of the variable. * Statistically significant at 5% level. *M: Main estimation results.*
R1: Robustness Test I (Exclude Social Media Variables). *R2: Robustness Test II (Mixed Logit with Actual Limited Choice Set).*

Model-Fit Comparison

To evaluate the fit of the proposed model, we estimate two baseline static demand estimation models: the Mixed Logit model with full choice set and the Mixed Logit model with actual (limited) choice set. We randomly partition our dataset into two subsets: one with 70% of the total observations as the estimation sample and the other with 30% of the total observations as the holdout sample. To minimize any potential bias from the partition process, we perform a 10-fold cross validation. We compute the predicted purchase probability for each product based on the model-estimated coefficients. We predict for both in-sample and out-of-sample estimation using our proposed dynamic search model and the two baseline models. The results are illustrated in columns 2-4 in Table 4.

Our model-fit estimation results demonstrate that the dynamic search model outperforms the two static baseline models in both in- and out-of-sample predictive power. In particular, our in-sample results in Table 4 show that with

respect to the root mean square error (RMSE), our proposed dynamic search model can improve the model fit by 34.89% compared to the Mixed Logit model with full choice set, and can improve the model fit by 17.30% compared to the Mixed Logit model with limited choice set. Similar trends in improvement in the model fit occur with respect to the other two metrics, mean square error (MSE) and mean absolute deviation (MAD), in both in-sample and out-of-sample analyses. Table 4.5b illustrates the out-of-sample model comparison results. Overall, the dynamic search model provides the best model fit, followed by the Mixed Logit model with limited choice set. The Mixed Logit model with full choice set provides the lowest model fit.

Note that since the static models do not consider the search cost, it is likely that the drop in model fit is caused by the missing variables that used to appear in the search cost from the dynamic model. To exclude this potential alternative explanation, we consider two additional static models by incorporating all the search cost variables into the previous two Mixed Logit models. We find that although the model fit increases for each static model, the overall performance remains the highest from the dynamic model. The corresponding results are illustrated in columns 5-6 in Table 4. Our model comparison results indicate that both the limited nature and the dynamic formation of the search choice set have significant impacts on modeling consumers' online search behaviors and final purchase decisions.

Table 3: Robustness Test (III) Results - Interaction Effects Between Travel Purpose and Sorting Criterion

	DFT	*PRA*	*CLD*	*CLA*	*CNA*	*HNA*
Family	**-2.452* (.079)**	**-1.007* (.391)**	**0.780* (.152)**	**1.291* (.171)**	–	-0.145 (.462)
Business	**-1.757* (.186)**	**.989* (.241)**	**1.073* (.227)**	–	–	**-2.076* (.108)**
Romance	**-1.289* (.211)**	**1.203* (.052)**	-0.323 (.389)	**3.030* (.782)**	–	**-0.417* (.068)**
Tourist	**-0.836* (.233)**	**-1.869* (.543)**	1.690 (1.746)	–	–	-0.674 (1.375)
Kids	0.535 (.662)	0.763(1.041)	**0.204* (.538)**	–	–	-0.422 (.706)
Senior	1.065 (1.753)	**-0.537* (.138)**	1.021 (1.249)	–	–	**-0.701* (.043)**
Pets	0.302 (.998)	0.799 (1.015)	-0.693 (.828)	–	–	–
Disability	–	–	–	–	–	–

* Statistically significant at 5% level.

Note: Some interaction effects are dropped in the estimation due to practical reasons (e.g., collinearity or very low significance).

Table 4: Model Fit Comparison Results

	Dynamic Search Model	**Mixed Logit Model with Full Choice Set**	**Mixed Logit Model with Limited Choice Set**	**Mixed Logit (Full Choice Set) +Additional Search Cost Variables**	**Mixed Logit (Limited Choice Set) +Additional Search Cost Variables**
			(In-sample)		
RMSE	0.0502	0.0771	0.0607	0.0723	0.0589
MSE	0.0025	0.0059	0.0037	0.0052	0.0035
MAD	0.0178	0.0242	0.0215	0.0229	0.0197
			(**Out-of-sample)**		
RMSE	0.1002	0.1767	0.1446	0.1602	0.1285
MSE	0.0100	0.0312	0.0209	0.0257	0.0165
MAD	0.0383	0.0658	0.0505	0.0582	0.0473

Conclusions and Future Work

In this paper, we propose a dynamic structural model for sequential search to examine the limited and endogenous nature of consumer online product search. We combine an optimal stopping framework with an individual-level random utility choice model, allowing us to jointly estimate consumer heterogeneous preferences and search cost.

Our estimation is validated on a unique dataset from the online hotel search industry. We have detailed individual-level search and transaction data from November 2008 through January 2009 containing approximately one million online sessions for 2117 hotels in the United States. We find that a dynamic model with limited consumer search provides a more precise measure of consumer price sensitivity and heterogeneous preferences than does a static demand model that does not account for search cost. Our final results indicate that too much feedback from the online social communities, along with long sentences, complex words or spelling errors in the social media content, may lead consumers to terminate their search early. Our study allows us to monetize the cognitive costs of consumers on seeking and absorbing the structured and unstructured product information under social media contexts. It also allows us to quantify the search cost associated with the use of product search engines.

Our research makes a number of major contributions. First, we quantify the effects of social media and product search engines on consumer search cost. By modeling search cost as a random-coefficient function of inherent and social contextual variables, we are able to unveil the nature of search cost under social environments. Second, we show the advantage of incorporating multiple and large data sources to uniquely identify consumer heterogeneous preferences and search cost. Third, we demonstrate the value of using structural econometric methods in analyzing emerging and important IS phenomena. Our dynamic model for consumer search combines the optimal stopping framework with an individual-level random utility choice model. It allows us to better capture the demand pattern under consumers' imperfect information in an online search market.

More broadly, our research sheds lights on how humans search, evaluate information, and make decisions in response to the emerging social contexts and digital interactions on search engines. In identifying the "cost of social media" to implement efficient and innovative ways of supporting decision making under social contexts, we provide important empirical evidence for future studies to build on. Meanwhile, our study provides critical insights on incorporating social costs into electronic marketplace design. With a deeper understanding of human cognitive limitations, we are able to more carefully design the product search mechanism in a way that can lead to non-trivial reduction in user search cost. For example, product search engines may not want to provide an exhaustive set of product features or customer reviews. Instead, they may focus on only the most unique features of each product and also may provide periodic digest of reviews. Meanwhile, since search cost can vary significantly under different ranking scenarios, it is crucial for electronic market designers to establish effective ranking recommendation mechanisms to facilitate economic exchange. Finally, our research demonstrates the potential of incorporating multiple and diverse large data sources into advanced economic models to more precisely study individual and organizational decision making. Our proposed approach, combining the optimal stopping framework with the discrete choice model, can be generalized to many other single-agent dynamic decision-making situations, as well (e.g., whether and when a company should adopt a new technology). Our empirical analysis aims to provide a rigorous basis for future researchers and businesses, with the goal of bringing about more ideas and inspirations for organizational IT strategy and managerial decision making.

Our work has several limitations, some of which can serve as fruitful areas for future research. First, our model assumes that the consumer knows the general distribution of utilities of alternatives, and each alternative follows the same distribution—there is no prior information to say that one might be expected to be superior to the other. However, on an online search engine, when the alternatives are sorted under certain criteria, they are presented in order of their predicted attractiveness to a consumer. Such recommendations can alter the distribution of the expected utilities of alternatives and may induce a shift in consumers' decision making (Dellaert and Häubl 2012). It would be interesting for future research to examine this fact from both a theoretical and an empirical perspective, and to seek further evidence and explanations that build on Dellaert and Häubl's experimental findings. Second, our model assumes that each search occurs in one step and reveals the exact utility of one alternative. However, this may not be true on a product search engine. A consumer, at first, may form some preliminary belief of the product utility (i.e., "perceived utility") based on the summary information provided on the search result page. Then, she updates her belief of utility (i.e., "actual utility") after clicking on the link to the product and examining the product's landing page. This two-step decision process involves consumer learning for the utilities of the consideration set. It would be interesting for future research to build on the Dynamic Bayesian Network Model (Chapelle and Zhang 2009) and capture the consumer learning process. Finally, due to the data limitation, we do not have the consumer-level demographic information. Since the search cost is likely to relate to the opportunity cost of consumer time, it would be helpful if future work could also incorporate such information (e.g., consumer age, income) into the model.

References

Agarwal, A., K. Hosanagar, M. Smith. 2011. Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets. *Journal of Marketing Research*, 48(6).

Aula, A. and K. Rodden. 2009. Eye-tracking studies: More than meets the eye. http://googleblog.blogspot.com/2009/02/eye-tracking-studies-more-than-meets.html

Baye, M.R., Gatti, J.R.J., Kattuman, P. and Morgan, J. 2009. Clicks, Discontinuities, and Firm Demand Online. *Journal of Economics & Management Strategy*. 18(4), 935-975.

Bellman, R. 1952. On the Theory of Dynamic Programming, *Proceedings of the National Academy of Sciences*.

Berry, S., Levinsohn, J., & Pakes, A. 1995. Automobile prices in market equilibrium. *Econometrica*, 63, 841–890.

Bing Social Search. 2011. Bing Social Search Team make Search 'Less Lonely'. http://www.microsoft.com/en-us/news/features/2011/nov11/11-02bingsocialsearch.aspx.

Branco, Fernando, Monic Sun, and J. Miguel Villas-Boas. 2012. Optimal Search for Product Information, *Management Science*, forthcoming

Bruno, Hernan and Naufel Vilcassim. 2008. Structural demand estimation with varying product availability. *Marketing Science*, 27 (6).

Brynjolfsson, Erik, Astrid Dick and Michael Smith. 2010. A nearly perfect market? Differentiation vs. price in consumer choice., *Quantitative Marketing and Economics*, vol.8, no.1.

Burdett, Kenneth and Kenneth L. Judd. 1983. Equilibrium Price Dispersion, *Econometrica* 51, 955-69, 1983.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. Search and Satisficing. *American Economic Review*, 101(7): 2899–2922.

Chapelle, O. and Zhang, Y. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of WWW 2009*, Madrid, Spain.

Chiang, J., S. Chib, C. Narasimhan. 1999. Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. *J. Econometrics* 89(1–2) 223–248.

Craswell, N. Zoeter, O., Taylor, M. and Ramsey, B. 2008. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of WSDM' 2008*. Palo Alto, CA.

de los Santos, Babur. 2008, Consumer search on the internet, *PhD dissertation*, Chicago University.

de los Santos, Babur, Ali Hortacsu and Matthijs Wildenbeest. 2011. Testing models of consumer search using data on web browsing and purchasing behavior. *Working paper*.

Dellaert, Benedict G.C., Gerald Häubl. 2012. Searching in Choice Mode: Consumer Decision Processes in Product Search with Recommendations. *Journal of Marketing Research.* Vol. 49, No. 2, pp. 277-288.

Ellison, G. and Ellison, S.F. 2009. Search, Obfuscation, and Price Elasticities on the Internet. *Econometrica*, 77, 427-452.

Erdem, T. and Keane, M.P. 1996. Decision-Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets. *Marketing Science*. vol. 15 no.11-20.

Ghose, A. and Ipeirotis, P. G. 2010. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE TKDE*.

Ghose, A., Iperotis, P. and Li, B. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*. Forthcoming.

Ghose A, and Yang, S. 2009. An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. *Management Science*. 55(10), pp. 1605-1622.

Goldfarb, A. and Tucker, C.. 2011. Search Engine Advertising: Channel Substitution When Pricing Ads to Context. *Management Science*, 57:458-470.

GroupM Search. 2011. The Virtuous Circle: The Role of Search and Social Media in the Purchase Pathway. http://groupmsearch.com/research.

Guadagni, P. M. and Little, J. D. C. 1983. A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science*, Vol. 2, No. 3, pp. 203-238.

Hann, I., & Terwiesch, C. 2003. Measuring the frictional cost of online transactions: The case of a name-your-own-price channel. *Management Science*, 49, 1563–1579.

Hong, Han and Matthew Shum. 2006. Can search cost rationalize equilibrium price dispersion in online markets? *Rand Journal of Economics*, 37 (2): 258.276.

Honka, Elisabeth. 2012. Quantifying search and switching costs in the U.S. auto insurance industry. *Working paper*.

Hortacsu, Ali and Chad Syverson. 2004. Product Di¤erentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds, *Quarterly Journal of Economics*, 119: 403.456 (May 2004).

Iprospect. 2008. iProspect Blended Search Results Study. http://www.herramientas-seo.com/pdf/estudio-buscadores-iprospect.pdf.

Johnson, Eric, Wendy W. Moe, Peter S. Fader, Steven Bellman, and Gerald L. Lohse. 2004. On the Depth and Dynamics of On-line Search Behavior, *Management Science*, 50 (3): 299.308.

Kahneman, D. and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263-292.

Kim, Jun, Paulo Albuquerque, and Bart J. Bronnenberg. 2010. Online Demand under Limited Consumer Search., *Marketing Science*, 29(6), pp. 1001-1023.

Koulayev, Sergei. 2010. Estimating Demand in Online Search Markets, with Application to Hotel Bookings. *Working Paper*.

McCall, J. J. 1970. Economics of information and job search, Quarterly *Journal of Economics* 84, 113-26.

McFadden, Daniel. 1974. Conditional Logit Analysis of Qualitative Choice Behavior, in Zarembka, Paul, ed., *FRONTIERS IN ECONOMETRICS*, Academic Press: New York, 105-142.

McFadden, D. and K. Train. 2000. Mixed MNL Models of Discrete Response. *Journal of Applied Econometrics*. 15, 447-470.

Mehta, Nitin, Surendra Rajiv, and Kannan Srinivasan. 2003. Price uncertainty and consumer search: a structural model of consideration set formation, *Marketing Science*, 22(1).

Moraga-Gonzalez, J. L. and Wildenbeest, M. R. 2008. Maximum likelihood estimation of search costs. *European Economic Review*, 52, 820-48.

Moraga-Gonzalez, J.L., Sandor, Z. and Wildenbeest, M.R. 2011. Consumer Search and Prices in the Automobile Market. *Working Paper*.

Mortensen, D.T. 1970. Job search, the duration of unemployment and the Phillips curve, *American Economic Review*, 847-62.

Pieters, R. and Warlop, L. 1999. Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, 16:1-16.

Reinganum, J. F. 1982. Strategic search theory. *International Economic Review*. 23(1) 1–15.

Reinganum, J. F. 1983. Nash equilibrium search for the best alternative. *J. Econom. Theory* 30(1) 139–152.

Richardson, M., E. Dominowska, and R. Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, 521–530. ACM.

Roberts, John H. and James M. Lattin. 1991. Development and Testing of a Model of Consideration Set Composition. *Journal of Marketing Research*, Vol. 28, No. 4, pp. 429-440

Rutz, O. and R.E. Bucklin. 2007. A Model of Individual Keyword Performance in Paid Search Advertising. *Working paper*, Yale University, New Haven, CT.

SearchEngineLand. 2012. Study: 72% Of Consumers Trust Online Reviews As Much As Personal Recommendations. http://searchengineland.com/study-72-of-consumers-trust-online-reviews-as-much-as-personal-recommendations-114152.

Simon, H. A. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99-118.

Sorensen, A. T. 2001. Price dispersion and heterogeneous consumer search for retail prescription drugs. *NBER working paper* 8548.

Stigler, G.J. 1961. The Economics of Information. *The Journal of Political Economy*, Volume 69, Issue 3 (Jun., 1961), 213-225.

Weitzman, M. L. 1979. Optimal search for the best alternative. *Econometrica* 47(3) 641–654.

Wildenbeest, M.R. 2011. An Empirical Model of Search with Vertically Differentiated Products. *Forthcoming in RAND Journal of Economics.*

Yao, S., C. F. Mela. 2011. A Dynamic Model of Sponsored Search Advertising. *Marketing Science*, 30(3), pp. 447-468.