Surviving Social Media Overload: Predicting Consumer Footprints on Product Search Engines

Anindya Ghose  
Stern School of Business, New York University  
aghose@stern.nyu.edu

Panagiotis G. Ipeirotis  
Stern School of Business, New York University  
panos@stern.nyu.edu

Beibei Li  
School of Information Systems and Management  
Heinz College, Carnegie Mellon University  
beibeili@andrew.cmu.edu

Abstract

The overload of social media content today can lead to significant latency in the delivery of results displayed to users on product search engines. We propose a dynamic structural model whose output can facilitate digital content analytics by search engines by helping predict consumers' online search paths. Such predictive prowess can facilitate web caching of the "most likely-to-be-visited" web pages and reduce latency. Our model combines an optimal stopping framework with an individual-level random utility choice model. It allows us to jointly estimate consumers' heterogeneous preferences and search costs in the context of product search engines, and predict a probability-based search path for each consumer. We estimate the parameters of the model using a dataset of approximately 1 million online search sessions resulting in room bookings in 2117 U.S. hotels. We find that search engine ranking can polarize search costs incurred by users. A good ranking saves consumers, on average, $9.38, whereas a bad one costs $18.54. Our model prediction results demonstrate that the proposed dynamic structural model provides the best overall performance in predicting the probabilities of consumers' online search paths compared to several baseline models that do not include a formal structure or dynamics in them.
1. Introduction

With the growing pervasiveness of social media, the volume and complexity of information that needs to be accessed by product search engines from their own platforms has been increasing rapidly. 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. Excess content can hinder consumers from efficiently seeking information and making decisions (e.g., Iyengar and Lepper 2000). More importantly, the onslaught of the exploding social media content may cause significant latency in the delivery of results on product search engines. A degradation in website performance may cause unexpected termination of search and in the long run may even discourage consumers from visiting the site. A study by Jupiter Research shows that online shopper loyalty is highly contingent upon quick web page loading. 33% of consumers shopping via a broadband connection will wait no more than four seconds for a web page to render (Jupiter Research 2006).

Traditional web search engines use web caching and prefetching as one of the most effective techniques to alleviate search engine bottleneck and reduce network traffic. Caching and prefetching is a browser mechanism, which utilizes browser idle time to download or prefetch documents that the user might visit in the near future. A web page provides a set of prefetching hints to the browser, and after the browser is finished loading the page, it begins silently prefetching specified documents and stores them in its cache. When the user visits one of the prefetched documents, it can be served up quickly out of the browser’s cache. During the past few years, a few studies have focused on designing caching frameworks (e.g., Lempel and Moran 2003, Jonassen et al. 2012), using static (i.e., offline) or dynamic (i.e., online) cache strategies. Unfortunately, existing strategies do not work well for commercial product search engines. First, many current caching strategies are based on user search history (e.g., caching the web pages that were requested in the past based on frequency or recency). This approach fails to locate web pages for new customers in an online shopping scenario because of the well-known “cold start” problem. Moreover, even for repeat customers, their shopping goals or preferences may change greatly over time under different shopping contexts. For example, a customer who has searched for Wall Street Inn in New York City for a business trip may be unlikely to search for it again when planning a romantic getaway on Valentine’s Day in New York City.

Second, recently search engines have been trying to improve the search history-based caching by prefetching web pages that they predict are going to be requested shortly (e.g., Lempel and Moran 2003, Jonassen et al. 2012). However, these predictions are based on the document relevance of web pages in response to a search query. Such design violates the goal of commercial product search engines. Instead of providing the most relevant documents, they should seek to display products with the highest value for money to consumers (Ghose et al. 2012). Furthermore, current prefetching and predictive caching
strategies assume that consumers do not have search costs and therefore search exhaustively. As a result, under the criterion of document relevance, caching will be equally exerted for items associated with different search costs (e.g., product listed on Page 1 vs. product listed on Page 10 of the search results). However, prior work on search engine settings has shown (iProspect 2008) that consumers are highly unlikely to reach the 10th page due to nonzero search costs, leading to little benefit from caching.

An alternative approach to increase product search engine performance and user experience is to improve the ranking mechanism (e.g., Ghose et al. 2012). Since consumers want the most desirable results early on, instead of caching, search engines can also reorder the results by their predicted probability of clicks and conversions. However, many times commercial product search engines are committed to presenting a given ranking (e.g., due to commercial agreements) and cannot or do not want to intervene in consumer search. In addition, even when they are able to reorder the items, they still face the same potential problem of latency in the loading of the product’s landing page. Thus, it is important for product search engines to cache the "most likely-to-be-visited" pages beforehand to improve the response rate.

Furthermore, increasingly we know that mobile internet usage has overtaken PC internet usage, especially in the area of travel. Web caching and prefetching are important techniques used to reduce the noticeable response time perceived by users. As modern mobile applications demand performance that is comparable to desktop machines, it is important to consider web caching and prefetching strategies for users accessing product search engines using mobile devices. Data access optimization and caching are key factors that can dramatically improve performance and therefore user experience. Web caching in mobile networks is critical due to the unprecedented cellular traffic growth that far exceeds the deployment of cellular infrastructures. Caching on handsets is particularly important as it eliminates all network-related overheads. The user experience improvement brought by caching is more notable in cellular networks whose latency is usually higher than those in Wi-Fi and wired networks.

Therefore, one major goal of our study is to design a predictive caching strategy for search engines to prefetch consumers' dynamic search paths by taking into accounts consumers' heterogeneous preferences and search costs. To understand the idea better, let's first consider the following example.

**Example 1:**

**Consumer Search Path:** Two different consumers try to book hotels online for their upcoming trip to Miami, FL: 1) John Doe, an IT consultant, is going on a business trip, and 2) Mr. Smith is going on a honeymoon trip with his spouse. They both go to Travelocity.com and start searching. John chooses to sort all hotels by name. After that, he starts by clicking "Airport Inn" and "Best Western," skips "DoubleTree," and clicks "Four Seasons." Finally, he clicks "Hilton Downtown" and decides to stop searching and make a reservation in it. Let the first letter in a hotel name represent that hotel. We use underline to denote the final purchase. Then in this scenario, John’s search path is A → B → F → H. Similarly, suppose Mr. Smith chooses to sort hotels by review rating. He skips "Airport Inn" and "Best Western," clicks on Hilton and Four Seasons. He stops searching...
after clicking "DoubleTree" and decides to make a reservation in it. His search path thus contains three hotels $H \rightarrow F \rightarrow D$. Figure 1 illustrates the online search path for each consumer, together with the major characteristic of each hotel indicated by the corresponding image (i.e., near airport, highway exit, good restaurants, downtown, beach).

**Figure 1. A Graphical Example of Consumer Online Search Path**

![Diagram of search paths](image)

**Key Challenge for Predicting Consumer Search Path:** If one can predict the path prior for a given consumer's search, then search engines can prefetch the related web page information (e.g., hotel location, services, photos, social media content, etc.) to minimize their response time. However, making such predictions can be challenging, because at each stage of the search, the cause of an observed decision by a consumer is hard to identify – e.g., The fact that Mr. Smith prefers to click "Four Seasons" over "Airport Inn" may be because of a higher valuation for hotel brand, or because he has incurred a lower search cost in searching for the former than the latter. Similarly, John stops searching and decides to make a booking in the "Hilton Downtown" either because he has a high evaluation towards "Hilton," or because his search cost becomes too high for him to continue searching.

More generally, the challenge in predicting consumer choice with search cost is to simultaneously identify consumers' heterogeneous preferences and search costs. 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). Thus, it is crucial to understand how these two leading factors can be uniquely recovered and what types of data are needed for the empirical identification.
Keeping the above in mind, another major goal of our study is to identify heterogeneous search costs under the social media context and examine its effect on consumer search behavior. The key identification strategy in our estimation relies on the fact that consumer preferences enter in both the search and purchase decision-making processes, 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 and purchase information allows us to successfully identify these two effects.

In summary, we propose a dynamic structural model for predictive digital analytics by search engines. It combines an optimal stopping framework with an individual-level random utility choice model. It allows us to jointly estimate consumers' heterogeneous preferences and search costs. Based on the analytical results, we predict the probability that a consumer clicks or purchases a certain product, and therefore predict a probability-based search path for the consumer.

Our model 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 on Travelocity.com. Our model provides more precise measures of consumer price sensitivity and heterogeneous preferences than does a static model that does not account for consumer search cost. Moreover, our proposed dynamic model demonstrates the best performance in predicting the consumer click and purchase probabilities compared to other existing benchmark models.

Our model builds on Weitzman's (1979) optimal sequential search framework. To the best of our knowledge, four existing studies that use similar methodologies to ours are Koulayev (2010), Kim et al. (2010), Bronnenberg et al (2012) and Chen and Yao (2012). However, our research differs from these studies in the following key ways: (1) Our model incorporates not only consumers' search behaviors, but also their purchases. The first two studies consider consumers' search information only as an approximation of their actual purchase decisions. (2) Our observations include 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, also examine the search costs associated with the refinement tools. We model consumer search refinement and the actual search/click as separate steps. This is different from Chen and Yao (2012) who assume zero costs of refinement and, therefore, treat search refinement as a prerequisite to consumer search. (5) Most importantly, our goal is to use the structural econometric approach as a tool for predictive digital analytics by product search engines to
improve their web caching strategy and user experience. This in turn can reduce search cost and increase market efficiency. A summary of the differences between this paper and the existing studies is in Table 1.

Our key contributions can be summarized as follows. First, we show the advantage of incorporating multiple and large data sources to analyze human decision making under cognitive constraints (i.e., search cost) in response to the emerging interplay of social media and search engines. Moreover, we are able to quantify the interaction effects of social media and search engines on human search cost. Second, we demonstrate the value of using predictive digital analytics by search engines based on structural econometric methods in finding new solutions to important business problems. Our dynamic model for consumer search combines the optimal stopping framework with an individual-level random utility choice model. It allows us to predict consumers’ dynamic search paths to improve the caching strategy for search engines. Our analytical approach can also be generalized to any platform or electronic market with an in-house search engine (e.g., Amazon.com, BestBuy.com, Aple’s iTunes store) given the commonality in the goal of reducing user search cost, reducing website latency and improving market efficiency.

Table 1. Comparison with Recent Literature

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>Amazon, View-Rank, 18 months</td>
<td>Amazon, View-Rank, Sale-Rank, 18 months</td>
<td>Hotels, Click (page), 1 month, (Chicago)</td>
<td>Hotels, Click, Purchase, 15 days</td>
<td>Hotels, Click, Purchase, 215 sessions, 15 days, ~1 M sessions, 2117 hotels</td>
</tr>
<tr>
<td><strong>Level of Analysis</strong></td>
<td>Market</td>
<td>Market</td>
<td>Page</td>
<td>Individual</td>
<td>Individual</td>
</tr>
<tr>
<td><strong>Real Transactions</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Search Refinements</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Costs for Refinements</strong></td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Social Media</strong></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

2. Prior Literature

Our paper draws from multiple streams of work. We summarize them in this section.
2.1 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, 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, we make decisions 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 et al. 1999, Mehta et al. 2003, Kim et al. 2010, Brynjolfsson et al. 2010).

2.2 Search Cost and Consumer Information Search

Second, our work builds on the literature on search cost and consumer information search. The existing literature holds two different views of the nature of consumer search: non-sequential and sequential search. The former strand of research follows Stigler's (1961) original model, assuming that consumers first sample a fixed number of alternatives and then choose the best from among them (e.g., Mehta et al. 2003, Moraga-Gonzalez et al. 2011). In contrast, the other view, arising from the job-search literature (e.g., 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. In our paper, we assume that consumers search sequentially on product search engines. This assumption is consistent with the mainstream research by the web search community (e.g., 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 et al. 2012, Chen & Yao 2012, Bronnenberg et al. 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) examine 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 et al. 2011), or the identification of search cost from switching cost (Honka 2012). Recent empirical work on sequential search examines consumers' limited search and the associated demand, with a focus on the online search market (Koulayev 2010, Kim et al. 2010). Meanwhile, De los Santos et al. (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 et al. (2011), who consider explanatory variables such as geographic distance from a consumer's home to different car dealerships. In our model, search cost is not only an inherent attribute of a consumer, but also a consequence of the social media context in which consumers of today are embedded. By modeling consumer search cost as a random-coefficient function of product-specific and associated social media variables, we aim to examine the nature of search cost given the interplay between product search engines and social media.

2.3 Search Engine Caching and Ranking

Finally, our work is also related to the literature on search engine caching and ranking. During the past twenty years, many studies have examined the capability of web search engines to estimate and cache in advance search results that are likely to be requested in the future. For example, Lempel and Moran (2003) present a Probability Driven Caching strategy based on a probabilistic model of search engine users. Jonassen et al. (2012) investigate the impact of query result prefetching on the efficiency and effectiveness of web search engines. They propose both office and online caching strategies for selecting and ordering queries whose results are to be prefetched.

Meanwhile, 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, 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., Ghose and Yang 2009, Goldfarb and Tucker 2011, Aggarwal et al. 2011, Yao and Mela 2011). Other 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 et al. (2012) propose a new utility gain-based ranking approach that accounts for consumer multidimensional preferences and recommends products with the highest expected utility.

3. Data

Our dataset comes from Travelocity.com, a leading 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. A typical online session involves the following events: the initialization of the session; the search query; the hotel listings returned from that search query in a particular rank order; whether the consumer has used any special sorting criteria to rerank the hotels; clicks on any hotel listing; the login and actual transactions in a given hotel; and the termination of the session. Notice that we also have detailed information associated with each event for every corresponding hotel, such as the nightly price and the hotel’s position in the set of listings returned by the search engine (i.e., "Page" and "Rank"). We have the detailed transaction-level information from Travelocity.com that is linked to 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 allows us to model consumer preferences for both the search and the purchase processes.

Our data also included additional hotel-related information from Travelocity.com, including hotel class, hotel brand, number of amenities, number of rooms, reviewer rating, number of reviews and the textual content of all the reviews up to January 31, 2009 (the last date of transactions in our database). To capture consumers' cognitive costs in reading reviews, we analyzed two sets of review text features that are likely to affect consumers' intellectual efforts in internalizing 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 had significant impact on product sales in the past (e.g., Ghose and Ipeirotis 2011). To derive the probability of subjectivity in the review's textual content, we apply standard text mining techniques. 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 (http://alias-i.com/lingpipe/). 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.¹

In addition, we also have supplemental data on hotel location-related characteristics collected independently. 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 techniques together with human annotations (from Amazon Mechanical Turk, AMT) to 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 2.² Notice that the data set we use in this paper is significantly different from the one used in Ghose et al. (2012). In this study, we use not only the transaction data (i.e., purchases), but the complete session-level data (i.e., both clicks and purchases). The resulting data set contains approximately seven million observations from one million individual user sessions, compared to a much smaller set of 8099 observations containing only the purchase information in Ghose et al. (2012).

3.1 Model-Free Evidence of Limited Search by Consumers

Before we describe our model, we seek from the data suggestive evidence that could motivate 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 2a 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 2b 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. In fact, a large majority of

¹ For detailed information on the text mining process, readers can refer to Ghose and Ipeirotis 2011.
² Due to space limitation, we only briefly discuss the data collection process here. Since the hotel-level location and service attributes are similar to (Ghose et al. 2012), interested readers can refer to the paper for more details, especially regarding how we use machine learning methods to collect different hotel level features.
consumers click less than 0.5 hotels per page, on average. These two figures provide us with preliminary evidence that consumers’ incur non-trivial search costs in this context and that consumer search is limited.

4. A Dynamic Structural Model of Consumer Sequential Search

In our dataset, we have the complete browsing session and the purchasing decisions that consumers made. Consumers have three options for a hotel during a search session: A) Do not click on the hotel at all; B) Click on the hotel but do not purchase it; C) Click on the hotel and also purchase it. To identify option A from options B and C, we need to model consumers’ click decision making. To identify option B from option C, we need to model consumers' purchase decision making. As a key contribution of this analytical study, we build a holistic model of user behavior that models both the clicking and purchasing behavior. Our model, in summary, works as follows:

1. A consumer session starts with a series of clicks \( r(1), \ldots, r(N) \), where the consumer visits the “details” landing pages of hotels, and estimates the utility that is expected to get from the hotel.
2. The consumer stops exploring new hotels (and hence stops clicking), when the expected marginal benefit of adding an additional hotel in the consideration set is less than the expected cost of searching. We adopt the concept of “reservation utility” from (Weitzman 1979) to define when the consumer stops exploring.
3. Once the consumer stops searching, the consideration set is fixed, and the consumer makes a decision to purchase one of the hotels in the choice set (or skip purchasing anything at all).

More details follow in the subsequent sections.

4.1 Model Setting

(1) Product Utility.

Assume the utility of hotel \( j \) for consumer \( i \) to be a random-coefficient model as follows:

\[
 u_{ij} = V_{ij} + \epsilon_{ij},
\]

(1)

where \( V_{ij} = V_{ij}^S + V_{ij}^L \) represents the expectation of the overall hotel utility. It consists of two parts \(^3\): the expected utility from "summary-page" hotel characteristics that consumers can directly observe on the search summary page, \( V_{ij}^S \), and the additional expected utility from "landing-page" hotel characteristics that consumers can only observe after clicking and arriving at the landing page, \( V_{ij}^L \).

---

\(^3\) 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.
Let $X_j$ be a vector of summary-page characteristics for hotel $j$. Let $P_j$ represent the Price for hotel $j$ that is also directly available to consumers on the search result summary page. Thus, we can model the expected summary-page utility as $V_j^s = X_j \beta_i - \alpha_i P_j$, where $\beta_i$ and $\alpha_i$ are consumer-specific parameters capturing the heterogeneous preferences of consumers. Consistent with the prior literature (e.g., Kim et al. 2010), we assume that $\beta_i \sim N(\bar{\beta}, \Sigma_\beta)$ where $\bar{\beta}$ is a vector containing the means of the random effects and $\Sigma_\beta$ is a diagonal matrix containing the variances of the random effects. Moreover, we assume that $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2)$. Similarly, we can model the expected landing-page utility as $V_j^l = L_j \lambda_i$, where $L_j$ represents a vector of landing-page characteristics for hotel $j$. $\lambda_i$ represents consumer-specific parameter capturing the heterogeneity. Consistent with previous assumptions, it follows a normal distribution $\lambda_i \sim N(\bar{\lambda}, \Sigma_\lambda)$. Thus, the overall utility function can be written as

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

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 hotels. For estimation tractability, we assume it to follow a Type I Extreme Value distribution $e_{ij} \sim \text{Type I EV}(0,1)$.

(2) Search Cost.

We model consumers' search costs to account for different dimensions in their evaluation of hotel-related information, including both the structured product information (e.g., hotel owner-provided descriptions) and the unstructured product information (e.g., social media content generated by the online communities). 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 hotel's online screen position can also have a significant effect on consumer search cost. Let $Q_j$ denote the set of variables that capture the above three dimensions of consumer information search for hotel $j$. We model the search cost of consumer $i$ for hotel $j$ to follow a log-normal distribution as follows $^4$:

$$c_{ij} = \exp(Q_j \gamma_i), \tag{3}$$

where $\gamma_i \sim N(\bar{\gamma}, \Sigma_\gamma)$, $\bar{\gamma}$ is a vector containing the means of the random effects and $\Sigma_\gamma$ is a diagonal matrix containing the variances of the random effects.

$^4$ The log-normal assumption of search cost is consistent with the prior literature (e.g., Kim et al. 2010, Wildenbeest 2011).
4.2 Problem Description and the Optimal Search Framework

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 hotel. At each stage of the search, the consumer has two options: to continue to search for the next alternative or to stop and purchase the current best alternative (including purchasing nothing, i.e., an outside good). Consider that the consumer is forward-looking. This situation 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 when is the optimal point for the consumer to choose the "stop" option.

Our solution to this problem builds on Weitzman's (1979) optimal sequential search framework. The basic idea is the following. Weitzman proposed an optimal stopping rule in which alternatives are ranked in descending order of their reservation utility. The intuition is that this value indicates a "rate of return" from searching each alternative. 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 hotel \( j \) for consumer \( i \), and let \( J \) be the total number of hotels available in the market. Thus, for each consumer \( i \), rank hotels in descending order of their reservation utility \( z_{ij} \). Denote the rank order by \( r_i(1)...r_i(J) \).

\[
\text{rank order by } r_i(1)...r_i(J)\]

\[
z_{i,r_i(1)}, z_{i,r_i(2)}, z_{i,r_i(3)}, \ldots, z_{i,r_i(j)}, \ldots, z_{i,r_i(J)}
\]

(4)

Note that, intuitively, ranking hotels by their reservation utility implies how "desirable" these hotels appear to consumer \( i \). According to Weitzman's "selection rule" (1979), consumer \( i \) searches sequentially
from the hotel with the highest reservation utility, $z_{i,r(i)}$, to the lowest, $z_{i,r(j)}$ in the ranking list. Given the current best utility $u_i^*$, the expected marginal benefits for consumer $i$ from searching $j$ are

$$B_{ij}(u_i^*) = \int_{u_i^*}^{\infty} (u_j - u_i^*)f(u_j)du_j,$$

where $f(\cdot)$ is the probability density function of hotel utility $u_j$. The expected marginal benefits $B_{ij}(u_i^*)$ represent the expectation of the utility for hotel $j$, 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 $u_i^*$. Thus, the reservation utility $z_{ij}$ meets the following boundary condition, where the marginal search cost $c_{ij}$ equals the expected marginal benefits from searching hotel $j$.

$$c_{ij} = B_{ij}(z_{ij}) = \int_{z_{ij}}^{\infty} (u_j - z_{ij})f(u_j)du_j. \tag{6}$$

So when consumer $i$'s current best utility is equal to the reservation utility of hotel $j$, $u_i^* = z_{ij}$, she is indifferent between searching for $j$ or stopping (and accepting $u_i^*$). Consumer $i$ will continue to search for hotel $j$ if her current best utility is lower than the reservation utility of hotel $j$, $u_i^* < z_{ij}$, and she will stop otherwise. More details on the derivation of the optimal search strategy is provided in Appendix.

4.3 Click Probability

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

$$\pi_{i,r(j)} = \Pr[r(j) \text{ is clicked by consumer } i] = \Pr[\max_{m=1}^{j-1}(V_{i,r(m)} + e_{i,r(m)}) < z_{i,r(j)}] = \prod_{m=1}^{j-1} F_e(z_{i,r(j)} - V_{i,r(m)}), \quad j > 1, \tag{7}$$

where $F_e(\cdot)$ is the CDF of $e_j$, which in our case is $e_j \sim \text{TypeI EV}(0,1)$.

4.4 Conditional Purchase Probability

Hotel $r(j)$ 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 hotel that has not been searched for; 2) The utility
of \( r(j) \) is greater than the utility of any other hotel 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 hotel \( r(j) \) for consumer \( i \) as

\[
\eta_{i,r(j)} = \Pr[ r(j) \text{ is purchased by consumer } i ] \\
= \Pr[ (V_{i,r(j)} + e_{i,r(j)}) > z_{i,r(m)}, r(m) \notin S_{i,N_i} ] \times \Pr[ (V_{i,r(j)} + e_{i,r(k)}) > (V_{i,r(k)} + e_{i,r(k)}), r(k) \in S_{i,N_i} ] \\
= \prod_{m=N_i+1}^{i} (1 - F_{r}(z_{i,r(m)} - V_{i,r(j)})) \times \frac{\exp(V_{i,r(j)})}{1 + \sum_{k=1}^{N_i} \exp(V_{i,r(k)})}.
\]

(Note that the mean utility for outside good \( r(0) \) is normalized to zero, \( V_{i,r(0)} = 0 \).) (8)

### 4.5 Joint Probability of Click and Purchase (Probability of Search Path)

Finally, to model the probability of the consumer’s full search path, we need to account for all the previous click and purchase decisions by the consumer. In particular, we examine the joint probability of all the click and purchase events in that consumer’s search session. Define \( \omega_{i,r(j),N_i} \) as the joint probability that consumer \( i \) has clicked \( N_i \) hotels and then purchased hotel \( r(j) \). Thus, we can model this joint probability as the following.

\[
\omega_{i,r(j),N_i} = \Pr[ r(1)...r(N_i) \text{ are clicked by consumer } i, r(j) \text{ is purchased by consumer } i, 0 \leq j \leq N_i ] \\
= \left( \prod_{k=1}^{N_i} \pi_{i,r(k)} \right) \times \eta_{i,r(j)}. \quad (9)
\]

### 4.6 Estimation

To model the utility of a hotel, we consider \( X \) to contain all hotel characteristics that are directly available on the search results summary page, including Hotel Class, Hotel Brand, Customer Rating and Total Review Count. We consider \( L \) to contain all additional hotel characteristics that can be revealed only from the hotel landing page, including Amenity Count, Number of Rooms, Number of External Amenities, Beach, Lake, Downtown, Highway, Public Transportation and Crime Rate.

To analyze consumers’ search costs, we consider \( Q \) to contain different factors that capture the structured and unstructured hotel information, as well as the online screen position of a hotel. Note that in our study the design of the landing page for each hotel on Travelocity.com is identical, each providing the same user interface, navigation, structure, hypertext links and website coherence, etc. Since our goal is to examine consumer decisions based on the variation in the search costs, we focus on the variance in the amount and complexity of hotel-related information. We use the Total Amenity Count to approximate the structured hotel information. Regarding the unstructured hotel information, we use the Total Review Count, Review Readability (i.e., complexity, syllables and spelling errors) and Review Subjectivity (i.e.,
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. Taking into consideration consumer heterogeneity, we have the search cost of consumer \( i \) for hotel \( j \) as follows:

\[
c_{ij} = \exp(\gamma_{ui} + \gamma_{j1} \text{PAGE}_{ij} + \gamma_{j2} \text{RANK}_{ij} + \gamma_{j3} \text{SPECIALSORT}_{ij} + \gamma_{j4} \text{AMENITYCNT}_{ij} + \gamma_{j5} \text{REVIEWCNT}_{ij} + \gamma_{j6} \text{COMPLEXITY}_{ij} + \gamma_{j7} \text{SYLLABLES}_{ij} + \gamma_{j8} \text{SPELLERR}_{ij} + \gamma_{j9} \text{SUB}_{ij} + \gamma_{j10} \text{SUBDEV}_{ij}).
\]

(10)

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

\[
\text{Likelihood}(\theta) = \prod_{i=1}^{I} \prod_{j=0}^{J} \left( \omega_{i,r(j),N_i} \right)^{y_{ij}},
\]

where \( \omega_{i,r(j),N_i} \) is the joint probability of consumer click and purchase defined in Equation (9). \( I \) is the total number of consumers and \( J \) is the total number of hotels. \( y_{ij} = 1 \) if the consumer has clicked and purchased hotel \( r(j) \); \( y_{ij} = 0 \) otherwise. Correspondingly, the overall log-likelihood function is

\[
\text{LL}(\theta) = \sum_{i=1}^{I} \sum_{j=0}^{J} y_{ij} \ln(\omega_{i,r(j),N_i}).
\]

(12)

Given the model setting, our final goal is to estimate the parameters of the random coefficients:

\[
\{ \theta \} = \left\{ \{ \alpha, \sigma \}, \{ \beta, \Sigma \}, \{ \gamma, \Sigma \} \right\}.
\]

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 \( \omega_{i,r(j),N_i} \). To maximize the log-likelihood function \( \text{LL}(\theta) \), we choose to use a non-derivative-based optimization algorithm (i.e., the Nelder-Mead simplex method) for heuristic search. This procedure iteratively searches for the optimal set of parameters \( \{ \theta^* \} \) until the log-likelihood function is maximized.

\[
\{ \theta^* \} = \arg \min_{\{ \theta \}} \sum_{i=1}^{I} \sum_{j=0}^{J} y_{ij} \ln(\omega_{i,r(j),N_i}).
\]

(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

---

5 As assumed by Weitzman’s original framework, the search cost of a product is deterministic to a consumer before search. To check the validity of our search cost model setting, we also tried a more conservative definition, using only the variables that are available from the search results summary page (i.e., by excluding variables that are available only from the landing page, such as amenity count or review textual content related variables). We found that the estimation results from the conservative setting are qualitatively very consistent with our main results.

6 As 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 are able to recover consistent structural parameters in our case.
search cost, we need to solve $z_y = B^{-1}_0(c_y)$ numerically. To improve the estimation efficiency, we apply an interpolation-based method to compute the reservation values (Kim et al. 2010, Koulayev 2010).

4.7 Identification

One of the major challenges in this analytical study 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 analytics, 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 identify these effects. We provide more detailed discussions below.

(1) Mean Consumer Preferences.

The mean preferences for hotel characteristics are identified by the correlation between the click and purchase frequencies of hotels and the frequencies of underlying hotels' characteristics. We measure the mean effect of a hotel characteristic by how often the same (or similar) characteristic appears in the hotels 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., hotel online position, the amount and complexity of social media content), which can be viewed simply as additional hotel 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 (6)) 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.

5. Empirical Results

Our main results are shown in Table 3 column 2. First, we find that the majority of the coefficients are statistically significant at the \( p \leq 5\% \) level, including both the mean effects \((\hat{\alpha}, \hat{\beta}, \hat{\lambda}, \hat{\gamma})\) and the heterogeneity \((\sigma_\alpha, \Sigma_\beta, \Sigma_\lambda, \Sigma_\gamma)\). 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 hotel location 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 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 hotel features or the more feedback from online social communities for a hotel on search 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.

For 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. 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 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 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.

Moreover, 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.

5.1 Model Prediction Experiments

Based on the model estimation, our final goal is to predict the probability of the dynamic search path for a consumer. If we can predict this probability, we will be able to infer the likelihood of the consumer's future actions at any stage of the search. However, predicting all possible combinations of search paths
can be computationally expensive. Instead, according to Equation (9), as long as we can predict the individual click probability and purchase probability for each product, we can dynamically derive the overall probability of a particular search path. For better understanding, let's look at another example.

Example 2:
Consider the same scenario described in Example 1. Based on our model estimation, we can infer, for a consumer like John or Mr. Smith, what the probability is for him to click or to purchase "Airport Inn," "Best Western," or "DoubleTree," etc. For instance, suppose we have computed the probabilities for John to click "Airport Inn," "Best Western," "DoubleTree," "Four Searsons" and "Hilton" to be 0.9, 0.7, 0.4, 0.1 and 0.4. We have also derived that his purchase probabilities towards these five hotels are 0.01, 0.02, 0.01, 0.02 and 0.2, and the probability for him to skip purchasing anything at all is 0.1.

Then, we can dynamically predict what John's next move would be at any stage of his search. For instance, suppose he has already clicked "Airport Inn" and "Best Western." In this case, there are five possible options for his next move:

- The probability to continue clicking "Four Seasons" is 0.9*0.7*0.4= 0.252;
- The probability to continue clicking "DoubleTree" is 0.9*0.7*0.1= 0.063;
- The probability to stop searching and purchase "Best Western" is 0.9*0.7*0.02= 0.0126;
- The probability to stop searching and purchase "Airport Inn" is 0.9*0.7*0.01= 0.0063;
- The probability to stop searching and purchase nothing is 0.9*0.7*0.1= 0.063.

Therefore, given the highest predicted probability (0.252), he is more likely to continue his search and click on the "Four Seasons" listed on the search engine screen. Finally, we can derive the probability of any search path prior to an actual search by a user. For example, the probability of John's full search path is 0.9*0.7*0.4*0.4*0.2= 0.02016.

Being able to predict consumers' dynamic search decisions can help search engines improve their web caching strategies and facilitate increased market efficiency. To examine the predictive performance of our model, we conduct a set of model prediction experiments. In particular, we compute the predicted individual click and purchase probabilities for each hotel based on the model-estimated coefficients. For evaluation, we estimate two baseline static demand estimation models. Both of them are widely used for predicting consumer choice probabilities: 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 conduct both in-sample and out-of-sample estimation using our model and the two baseline models. We then compare the predictive performance of both the click and the purchase probabilities of a product. The prediction results for click probability are illustrated in columns 2-4 in Tables 5a and 5b. The prediction results for purchase probability are illustrated in columns 2-4 in Tables 6a and 6b.
Our model prediction results demonstrate that our model outperforms the two static baseline models in both in- and out-of-sample predictive power for both click and purchase predictions. For example, the in-sample results in Table 6a show that with respect to the root mean square error (RMSE), our proposed model can improve the prediction performance of purchase probability 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 predictive power occur with respect to the other two metrics, mean square error (MSE) and mean absolute deviation (MAD), in both in- and out-of-sample analyses. We also find consistent trend in the prediction of click probability. Overall, our dynamic model provides the highest predictive power, followed by the Mixed Logit model with limited choice set. The Mixed Logit model with full choice set provides the lowest predictive power.

Note that since the static models do not consider the search cost, it is likely that the drop in predictive power is caused by the missing variables that used to appear in the search cost from the dynamic model. To examine this potential issue, 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 Tables 5a, 5b, 6a and 6b.

The model prediction experiments indicate that our model is able to better predict consumers’ dynamic search behaviors. Using our analytical approach product search engines are able to more precisely predict consumers’ online moves and prefetch the related web pages to minimize the response time.

5.2 Robustness Checks

To assess the robustness of our analytical model, and meanwhile, to analyze how social media and consumer heterogeneity (e.g., travel purposes, search engine ranking criteria) may affect the search cost and decisions of a consumer, 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 3, 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 elasticity 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 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 3, column 4.

3) **Robustness Test III: 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:

$$
T_i' = [\text{Family}, \text{Business}, \text{Romance}, \text{Tourist}, \text{Kids}, \text{Senior}, \text{Pets}, \text{Disability}]_{i,8}.
$$

(14.1)
We acquire the empirical distribution of \( T_i \) from online consumer reviews and reviewers’ profiles.\(^7\)

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 ascending (CNA) and hotel name ascending (HNA). Let \( S_\text{ij} \) denote the indicator vector of sorting method under which product \( j \) is presented to consumer \( i \) during his/her search:

\[
S_\text{ij} = [DFT_\text{ij}, PRA_\text{ij}, CLD_\text{ij}, CLA_\text{ij}, CNA_\text{ij}, HNA_\text{ij}]_{6 \times 1}.
\]

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

\[
c_\text{ij} = \exp(\gamma_0 + \gamma_1 \text{PAGE}_\text{ij} + \gamma_2 \text{RANK}_\text{ij} + \Gamma \times S_\text{ij} + \gamma_4 \text{AMENITYCNT}_\text{ij} + \gamma_5 \text{REVIEWCNT}_\text{ij} + \gamma_6 \text{COMPLEXITY}_\text{ij} + \gamma_7 \text{SYMBOLS}_\text{ij} + \gamma_8 \text{SPELLERR}_\text{ij} + \gamma_9 \text{SUB}_\text{ij} + \gamma_{10} \text{SUBDEV}_\text{ij}).
\]

where everything else remains the same as that in Equation (10), except that \( \Gamma \) 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 Table 4.

We find that consumers’ travel purposes can explain their heterogeneous search costs under different ranking mechanisms. In general, the default ranking (DFT) reduces 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, we do not find significant interaction effects for consumers who travel with young kids, senior citizens, or families with pets. This finding seems to indicate that the current default ranking captures mainly consumers’ preferences under the most popular travel contexts. The default ranking may not be the most effective when consumers are seeking for certain special amenities during travel search.

The price ascending ranking (PRA) decreases the search costs for tourists (i.e., -1.869), family travelers (i.e., -1.007) and senior citizens (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 suggests that romance and business travelers are less price-sensitive, whereas tourists tend to be the most price-sensitive. Ranking by hotel class does not seem to reduce consumers’ search costs. In fact, 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). Meanwhile, 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 suggests that starting with similar hotels (either the luxury ones or the budget ones) may not be informative for

---

\(^7\)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.
consumers during the search. Consumers are willing to explore products with better variety (e.g., Ghose et al. 2012), especially when they face certain constrain and cannot search exhaustively. Interestingly, hotel name ranking (HNA) can significantly reduce search costs for different categories of travelers. Under this ranking mechanism, search costs decrease the most for business travelers (i.e., -2.076), followed by senior citizens (i.e., -0.701) and romance travelers (i.e., -0.417). This finding indicates that hotel brands can significantly reduce consumers' search costs under certain travel contexts. For example, business travelers going to attend a conference may seek particular hotels that are recommended by the conference. Seniors travelers may prefer special hotel chains reputed for being friendly to them and look for them directly.  

Note that the main purpose of the robustness tests is to validate our model and to analyze how the interplay between social media and search engines, as well as consumer heterogeneity (e.g., travel purposes, search engine ranking criteria) may affect the corresponding search cost and the footprints of a consumer online.

6. Conclusions and Future Work

In this paper, we propose a dynamic structural model for predictive digital analytics by product search engines to predict consumers’ online search paths as well as to measure and quantify the search costs incurred by user. Our model combines an optimal stopping framework with an individual-level random utility choice model. It allows us to jointly estimate consumers’ heterogeneous preferences and search costs in a product search engine context where social media is quite pervasive, and to identify the key driver of a consumer’s decision at each stage of the search and purchase process. Our final analytical results can help product search engines predict and cache the “most likely-to-be-visited” web pages beforehand to minimize the response time and improve user experience.

On a broader note, our research makes a number of major contributions. First, we show the advantage of incorporating multiple and large data sources to analyze how humans search, evaluate information, and make decisions under cognitive constraints (i.e., search cost) in response to the emerging interplay between social media and search engines. Moreover, we are able to quantify the interaction effects of social media and search engines on human search cost.

Second, we demonstrate the value of using predictive digital analytics by search engines based on dynamic structural econometric methods in finding new solutions for important business problems. Our dynamic model for consumer search combines the optimal stopping framework with an individual-level random utility choice model. It allows us to harness the advantage of multistage consumer behavioral data

---

8 This finding is consistent with Ghose et al. (2012) where the authors found that senior travelers have a special preference for Best Western hotels.
on search engines to identify the causes of consumer decisions in electronic markets. Furthermore, the final outcome allows us to more precisely predict consumers’ future search paths on product search engines. It also allows us to design effective web caching strategies for search engines to improve the user experience and increase the potential click-through and conversion rates. Moreover, this approach can be generalized to any platform or electronic market with an in-house search engine, given the commonality in the goal of reducing user search cost, reducing website latency and improving market efficiency.

Finally, our proposed model framework 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 studies to build on, with the goal of exploring the tremendous potential of growing big data and sophisticated business analytical tools, to bring about more 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, when the alternatives are sorted on search engines 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 to examine this fact from both theoretical and empirical perspectives. Second, it would be interesting to compare our model with the Dynamic Bayesian Network Model (DBN, Chapelle and Zhang 2009). DBN assumes that a consumer, at first, forms 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 process involves consumer learning for the utilities of the consideration set. It would be interesting to incorporate consumer learning into our model. 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 is helpful to include such information (e.g., age, income) in future.
References

- Ghose, A. and Ipeirotis, P. G. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. IEEE Transactions on Knowledge and Data Engineering (TKDE), 23 (10), 1498-1512.
Table 2. Definitions and Summary Statistics of Variables

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

<table>
<thead>
<tr>
<th>Total # Sessions:</th>
<th>969,033</th>
<th>Total # Hotels:</th>
<th>2117</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Period:</td>
<td>11/1/2008-1/31/2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>Mean Effect (Std. Err)(^M)</td>
<td>Heterogeneity (Std. Err)(^M)</td>
<td>Mean Effect (Std. Err)(^R1)</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>(Preferences)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRICE(^L)</td>
<td>-1.423* (.000)</td>
<td>0.578* (.023)</td>
<td>-1.925* (.001)</td>
</tr>
<tr>
<td>CLASS</td>
<td>1.667* (.002)</td>
<td>1.377* (.087)</td>
<td>1.729* (.003)</td>
</tr>
<tr>
<td>RATING</td>
<td>3.199* (.003)</td>
<td>1.923* (.021)</td>
<td>3.543* (.007)</td>
</tr>
<tr>
<td>AMENITYCNT(^L)</td>
<td>.053* (.006)</td>
<td>.004 (.032)</td>
<td>.076* (.003)</td>
</tr>
<tr>
<td>REVIEWCNT(^L)</td>
<td>1.411* (.003)</td>
<td>1.405* (.090)</td>
<td>1.599* (.006)</td>
</tr>
<tr>
<td>ROOMS(^L)</td>
<td>1.005* (.002)</td>
<td>.056 (.071)</td>
<td>1.336* (.023)</td>
</tr>
<tr>
<td>EXSTAMENITY(^L)</td>
<td>.082* (.001)</td>
<td>.005 (.024)</td>
<td>.064* (.011)</td>
</tr>
<tr>
<td>BEACH</td>
<td>1.001* (.010)</td>
<td>.072* (.012)</td>
<td>1.545* (.012)</td>
</tr>
<tr>
<td>LAKE</td>
<td>-.767* (.089)</td>
<td>1.356* (.059)</td>
<td>-.702* (.065)</td>
</tr>
<tr>
<td>TRANS</td>
<td>1.046* (.003)</td>
<td>.043* (.029)</td>
<td>1.067* (.008)</td>
</tr>
<tr>
<td>HIGHWAY</td>
<td>.602* (.091)</td>
<td>.070* (.005)</td>
<td>.559* (.076)</td>
</tr>
<tr>
<td>DOWNTOWN</td>
<td>.586* (.004)</td>
<td>.116* (.047)</td>
<td>.534* (.017)</td>
</tr>
<tr>
<td>CRIME</td>
<td>-.112* (.001)</td>
<td>.017 (.036)</td>
<td>-.179* (.006)</td>
</tr>
<tr>
<td>BRAND</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Search Cost)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search Base Cost</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PAGE</td>
<td>4.017* (.002)</td>
<td>1.633* (.003)</td>
<td>3.598* (.012)</td>
</tr>
<tr>
<td>RANK</td>
<td>2.178* (.006)</td>
<td>0.340* (.001)</td>
<td>2.241* (.011)</td>
</tr>
<tr>
<td>SPECIALSORT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMENITYCNT(^L)</td>
<td>0.343* (.005)</td>
<td>0.146* (.001)</td>
<td>0.389* (.006)</td>
</tr>
<tr>
<td>REVIEWCNT(^L)</td>
<td>0.500* (.008)</td>
<td>0.211* (.005)</td>
<td>---</td>
</tr>
<tr>
<td>COMPLEXITY</td>
<td>1.349* (.011)</td>
<td>0.142* (.006)</td>
<td>---</td>
</tr>
<tr>
<td>SYLLABLES(^L)</td>
<td>1.668* (.015)</td>
<td>0.378* (.010)</td>
<td>---</td>
</tr>
<tr>
<td>SPELLERR(^L)</td>
<td>0.814* (.005)</td>
<td>0.290* (.008)</td>
<td>---</td>
</tr>
<tr>
<td>SUB</td>
<td>0.205* (.002)</td>
<td>0.079* (.001)</td>
<td>---</td>
</tr>
<tr>
<td>SUBDEV</td>
<td>0.822* (.019)</td>
<td>0.102* (.007)</td>
<td>---</td>
</tr>
</tbody>
</table>

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

29
Table 4: Robustness Test (III) Results
- Interaction Effects Between Travel Purpose and Sorting Criterion on Search Cost

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

* 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).

Figure 2a. Distribution of # Pages Browsed (Session Level)

Figure 2b. Distribution of # Click-thoughts Per Page (Session Level)
<table>
<thead>
<tr>
<th>Table 5a: In-sample Model Prediction Results (Click Probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Search Model</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MSE</td>
</tr>
<tr>
<td>MAD</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Table 5b: Out-of-sample Model Prediction Results (Click Probability)</td>
</tr>
<tr>
<td>Dynamic Search Model</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MSE</td>
</tr>
<tr>
<td>MAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6a: In-sample Model Prediction Results (Purchase Probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Search Model</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MSE</td>
</tr>
<tr>
<td>MAD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6b: Out-of-sample Model Prediction Results (Purchase Probability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Search Model</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>RMSE</td>
</tr>
<tr>
<td>MSE</td>
</tr>
<tr>
<td>MAD</td>
</tr>
</tbody>
</table>

31
Appendix. Optimal Search Framework

Our model builds on the optimal sequential search framework by Weitzman (1979). Consider that consumers are forward-looking and trying to maximize the expected present value of utility over a planning horizon (e.g., Erdem and Keane 1996). The expected present value in our setting can be computed as follows. First, we partition the set of available alternatives into $S_i \cup \overline{S}_i$, with $S_i$ containing all the ones that have been searched and $\overline{S}_i$ containing all the non-searched ones. Let $u_i^*$ be the highest utility searched so far, thus we have

$$u_i^* = \max_{j \in S_i} \{u_j, 0\}. \quad (A1)$$

The state of the system at any time during the search is given by $(u_i^*, \overline{S}_i)$. Define $\Psi(u_i^*, \overline{S}_i)$ as the expected present discounted value of following an optimal search policy, from the current state $(u_i^*, \overline{S}_i)$ going forward. Therefore, for each $u_i^*$ and $\overline{S}_i$, the state valuation function $\Psi(u_i^*, \overline{S}_i)$ must satisfy the Bellman equation (Weitzman 1979):

$$\Psi(u_i^*, \overline{S}_i) = \max \left\{ u_i^*, \max_{j \in \overline{S}_i} \left[ -c_j + d_j \cdot \left( \Psi(u_i^*, \overline{S}_i - \{j\}) - \int_{u_j}^{u_i^*} f(u_j)du_j + \int_{u_j}^{\infty} \Psi(u_j, \overline{S}_i - \{j\})f(u_j)du_j \right) \right] \right\}, \quad (A2)$$

where $F(\cdot)$ is the CDF of $u_j$ and $f(\cdot)$ is the probability density function of $u_j$. Therefore, at current state $(u_i^*, \overline{S}_i)$, the consumer can either terminate search and collect reward $u_i^*$, or search any $j \in \overline{S}_i$ to maximize $\Psi(u_i^*, \overline{S}_i)$. Given the short time span in online search, we set the discount rate $d_i$ to 1. Equation (A2) is the principle of optimality for dynamic programming. As pointed by Weitzman (1979), the solution to this optimization framework is to continue searching until a utility $u_i^*$ is found that is larger than some limit called “reservation utility”, $z_j$. Each consumer $i$ has a reservation utility $z_j$ for each product $j$ that – if she already found a product with that utility ($u_i^*_{[0\ldots j-1]} = z_j$) – leaves her indifferent between keep searching and not searching product $j$.

More formally, let the expected marginal utility for consumer $i$ from the search of product $j$ be

$$B_j(u_i^*) = \int_{u_i^*}^{\infty} (u_j - u_i^*)f(u_j)du_j. \quad (A3)$$

Thus, consumer $i$ will continue to search if there exists at least one $j$ such that the expected marginal benefit from searching product $j$ exceeds its corresponding search cost

$$c_j < B_j(u_i^*). \quad (A4)$$
Define the *reservation utility* $z_{ij}$ as the utility value that satisfies the following boundary condition (same as Equation (7) in the paper), where the search cost equates the expected marginal utility from searching product $j$.

$$c_{ij} = B_j(z_{ij}) = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij}. \quad (A5)$$

Therefore, the optimal search strategy for a consumer is to continue searching until a utility $u^*_i$ is found larger than the boundary solution $z_{ij}$. Note that the reservation utility $z_{ij} = B^{-1}_j(c_{ij})$ has been proved to uniquely exist given the monotonicity of $B_j$ (Weitzman 1979). We can numerically compute $z_{ij}$ based on the algorithm suggested by Kim et al. (2010). Due to space restriction, we do not describe the algorithm in details. We refer the interested readers to the Appendix in Kim et al. (2010).