# Modeling Consumer Footprints on Search Engines: An Interplay with Social Media

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[1]

Heinz College, Carnegie Mellon University
beibeili@andrew.cmu.edu

**Acknowledgements:** The authors thank Travelocity for providing data and support to this research. The authors thank the department editor, the anonymous associate editor, three anonymous reviewers for their constructive comments. For valuable feedback and suggestions, the authors also thank Ravi Bapna, Pradeep Chintagunta, Daria Dzyabura, Alok Gutpa, Elisabeth Honka, Sam Hui, Sergei Koulayev, Ramayya Krishnan, Michael D Smith, Rahul Telang, Russell S. Winer, and audience at the 2012 International Conference on Information Systems, 2012 Workshop for Information Systems and Economics, 2012 INFORMS Annual Conference, 2014 Marketing Science Conference, Carnegie Mellon University, Stern School of Business at New York University, and Carlson School of Management at University of Minnesota. An earlier conference version of this paper has won the Best Theme Paper Award at the International Conference on Information Systems in 2012.

[1] Author names are in alphabetical order.

## Abstract

It is now well understood that social media plays an increasingly important role in consumers' decision making. However, an overload of social media content in product search engines can hinder consumers from efficiently seeking information. We propose a structural econometric model to understand consumers' preferences and costs on search engines to improve user experience under unstructured social media. Our model combines an optimal stopping framework with an individual-level random utility choice model and analyzes click behavior in conjunction with purchase choices. Our model takes into accounts three major constraints in a consumer's decision making process: (1) interdependency in decision making for different alternatives; (2) sequential arrival of information revealed by click-throughs; (3) non-negligible search cost. Our approach allows us to jointly estimate consumers' heterogeneous preferences and search costs under the interplay of social media and search engines, and predict search and purchase behavior for each consumer. We validate the model using an individual session-level dataset of approximately 7 million observations resulting in room bookings in 2,117 U.S. hotels. Interestingly, our analysis allows us to quantify the trade-off between consumers' benefits and cognitive costs from using large-scale unstructured social media information during decision making. Our policy experiments show that providing a carefully curated digest of social media content during the earlier stages of consumer search (i.e., on the search results summary page) can lead to a 12.01% increase in the overall search engine revenue.

## 1. Introduction

With the growing pervasiveness of social media, the volume and complexity of information product search engines need to access from their own platforms has been increasing rapidly. For example, websites such as Amazon.com, TripAdvisor.com, or Yelp.com can attract hundreds or even thousands of review postings that compete for users' attention. The onslaught of the exploding social media content can lead to a significant information overload for consumers during product search. Such excess content can hinder consumers from efficiently seeking information and making decisions (e.g., Iyengar and Lepper 2000). What is worse, it may discourage consumers from searching and cause unexpected termination of search (e.g., session drop-out).

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 the user search experience. Recent studies show that product search engines can improve the ranking design and the user search experience based on the prediction of consumer preferences (e.g., Ghose et al. 2012, Do los Santos and Koulayev 2014). Because consumers want the most desirable results early on, search engines can reorder the results by the predicted probabilities of consumer preferences (e.g., clicks or purchases).

Previous studies have examined how to estimate customer preferences based on online purchase information only (e.g., Ghose et al. 2012). However, consumer footprints on search engines provide us with a tremendous amount of information that reveals their preferences, even in the absence of purchases (e.g., Koulayev 2014, Kim et al. 2010, Do los Santos and Koulayev 2014). When this search behavior is combined with purchases, the signals become even more comprehensive and useful. However, although many studies have worked on using either historical click-throughs or conversions separately to estimate consumer preferences, there is little work in jointly analyzing the search and purchase behavior to infer individual consumer preferences and identify the products that satisfy most the user needs.

With the deluge of structured and unstructured social media content, consumers' cognitive costs in searching and evaluating product information become non-negligible. As a result, search costs also play an important role in affecting consumers' choices in product search engines. Therefore, *a major goal of our study is to better understand consumers' online footprints by taking into account consumers' heterogeneous preferences and search costs, using both click and purchase information*. However, this task can be challenging, because the cause of an observed search behavior by a consumer is hard to identify − e.g., The fact that a consumer prefers to click product A over product B may be because of a higher valuation for A, or because the consumer has incurred a lower search cost in searching for A than for B.

More generally, the challenge in predicting consumer choice with search cost is to simultaneously identify consumers' heterogeneous preferences and search costs (Hortacsu and Syverson 2004). A consumer may stop searching either because of a high valuation for the products already found or because of a high search cost. Either the preferences for product characteristics or the moments of the search cost distribution can explain the same observed search outcome (Koulayev 2014). Keeping the above in mind, *another major goal of our study is to identify heterogeneous search costs under the social media context, examine their effect on consumer search behavior, and provide insights to product search engines on better design and management of social media content to improve user experience.* The key identification strategy for consumer search cost in our study relies on the exclusion restriction 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. In addition, we model search cost as a function of an exclusive set of variables. From an empirical identification perspective, we can simply view the search cost variables as additional product characteristics.

In summary, we propose a structural econometric model to understand consumers' preferences and search costs on product search engines to improve user experience under large-scale, unstructured social media. 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 results, we are able to predict the probability that a consumer clicks or purchases a certain product and provide a better understanding of what drives consumer engagement. Our analysis also allows us to quantify the trade-off between consumers' benefits and cognitive costs from using large-scale unstructured social media information during decision making. Our policy experiments offer insights to search engines on what product information they should display during different stages of consumer search (i.e., on the search result summary page vs. product landing page), to improve user experience, click/purchase probabilities as well as search engine revenues.

Our model is validated by a unique dataset from the online hotel search industry. We have detailed individual consumer session-level search and transaction data from November 2008 through January 2009, containing approximately seven million observations resulting in room bookings in 2,117 hotels in the United States on Travelocity.com. Our model provides more precise measures of consumer price elasticity and heterogeneous preferences than does a static Mixed Logit model that does not account for consumer search cost or the sequence of the prior clicks. Our model also provides better predictive performance than does a click model that purely relies on the click information. More specifically, our model demonstrates the best performance in predicting the consumer click and purchase probabilities compared to other benchmark models. We see a 14.92 % and an 18.77% improvement in the out-of-sample prediction using our model compared to the next best performing model, with respect to click-through and conversion probabilities, respectively.

Our policy experiments show that providing additional product information, especially the location-related information, on the travel search engine summary page will lead to a 22.16% increase in the overall search engine revenue. By contrast, although hiding all hotels' price information from the search summary page may lead to higher user "engagement" (when engagement is measured by number of clicks), it can hurt the travel search engine eventually by leading to a 7.08% drop in the overall search engine revenue. On the contrary, providing a carefully curated digest of social media textual content on search results summary page can lead to a 12.01% increase in the overall search engine revenue. This finding suggests that it is important for product search engines to leverage the economic value of large-scale unstructured social media information, while in the meantime reducing the cognitive burden of consumers by automating the extraction of such information and presenting it to the consumers during the earlier stages of the decision making process.

## 2. Prior Literature

Our paper draws from multiple streams of work. We summarize them in this section.

### *2.1 Search Cost and Consumer Information Search*

First, our work builds on the literature on search cost and consumer information search. Recent studies have found that consumers have cognitive limitations, and search costs exist during the information search processes. Disregarding consumers' cognitive limitations and the limited nature of choice sets can lead to biased estimates of demand (e.g., Mehta et al. 2003, Kim et al. 2010, Brynjolfsson et al. 2010).

The existing literature in this field 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 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. 2012, Honka 2014). By contrast, the other view, arising from the job-search literature (e.g., Mortensen 1970), argues 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), in multi-agent scenarios, have laid theoretical foundations for sequential search models. In our paper, we assume 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 2014, Chen and Yao 2016).

With the growing interests and the recent development of information technologies that have made many intensive computation tasks more tractable today, empirical work to date has increased. 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. 2012), or the identification of search cost from switching cost (Honka 2014). Recent empirical work on sequential search examines consumers' limited search and the associated demand, with a focus on the online search market (Koulayev 2014, Kim et al. 2010). Meanwhile, De los Santos et al. (2012) use web browsing and purchasing behavior based on book-price distribution across 14 online bookstores to compare 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. (2012), 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. Note that consistent with prior literature, the search cost in our study is modeled as exogenous to the consumer's search. By modeling consumer search cost as a random-coefficient function of the textual variables that are related to the unstructured social media content, we aim to examine the nature of search cost given the increasing interplay between product search engines and social media.

Finally, another related stream of consumer search literature has analyzed optimal search behavior when consumers are uncertain about the distribution of the product price or utility (e.g., Rothschild 1974, Rosenfield and Shapiro 1981, Bikhchandani and Sharma 1996, Koulayev 2013, De los Santos et al. 2013). For example, the recent work by De los Santos et al. (2013) has relaxed the assumption that consumers "know" the distribution of offerings while deciding on their search strategy, and allows for learning of the utility distribution. More specifically, consumers learn the utility distribution by Bayesian updating their Dirichlet process priors while sampling information about products and retailers. Our study is related to this stream of work in that we also consider the sequential arrival of information during different search stages, which allows for consumer update of the initial belief towards product utility via information search.

### *2.2 Search Engine Ranking and Unser-Generated Content (UGC)*

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 lot of attention. A number 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, Agarwal et al 2011, Yao and Mela 2011). Other studies focus on 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 the easy price search makes demand highly price sensitive for some products. Ghose et al. (2012) propose a utility-gain-based ranking (using data from past purchases, only, and not browsing behavior) that recommends products with the highest expected utility. The lab experiments indicate a strong preference for utility-based ranking compared to existing state-of-the-art alternatives. Ghose et al. (2014) combined a Hierarchical Bayesian model and randomized user experiments to examine the search engine ranking and personalization effects from a causal perspective.

Finally, our work also relates to the stream of research on social media and User-Generated Content (UGC) (e.g., Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Dellarocas et al. 2007, Duan et al. 2008, Forman et al. 2008). Especially, it builds on the recent research from a multidimensional view of the customer reviews (e.g., Archak et al. 2011, Ghose et al 2012, Netzer et al. 2012, Chen et al. 2017). In this paper, we aim to examine the role of social media from multiple dimensions in affecting not only the product utility evaluation but also the search cost of consumers.

### *2.3 Comparisons with Recent Literature*

Our model builds on Weitzman's (1979) optimal sequential search framework. To the best of our knowledge, five existing studies use similar methodologies to ours: Kim et al. (2010), Koulayev (2014), De los Santos and Koulayev (2014), Kim et al. (2014), and Chen and Yao (2016). However, our research differs from these studies in the following ways: (i) Our model incorporates not only consumers' search behavior, but also their purchases. Kim et al. (2010), De los Santos and Koulayev (2014), and Koulayev (2014) consider only consumers' search information as an approximation of their actual purchase decisions. (ii) Our observations include detailed click-throughs from each ranking position on a page, which allows us to precisely model the individual click probability for each product, rather than for a page with a bundle of products (i.e., a page of 15 hotels as in Koulayev 2014). More broadly speaking, Koulayev (2014) and our paper are complimentary: Koulayev models the costly process of discovering new hotels by flipping pages, but stops short of modeling what happens between click and booking. Our paper focuses on the second stage, starting from the costly click to the final booking. (iii) We conduct our analysis at the individual-consumer level as opposed to at the aggregate market level (Kim et al. 2010, 2014). Such individual-level data allow us to leverage the detailed information of the *sequence of clicks* per session, rather than only the independent click-throughs. (iv) Chen and Yao (2016), De los Santos and Koulayev (2014), and Koulayev (2014) focus on constructing models that examine the joint use of search refinement tools (e.g., sorting) during consumer search. However, search refinement is not our focus in this paper. (v) Kim et al. (2010, 2014) and Chen and Yao (2016) assume a simpler information structure where consumers do not update their information set during search. Whereas, our paper allows for a more realistic information structure by allowing consumers to update their information set before and after click-through. (vi) Most importantly, our paper initiates a special focus on the interplay between consumer search and social media. Our goal is to use the structural econometric approach as a tool for analytics by product search engines to improve the user experience, especially under an overload of the unstructured social media content. We model the trade-off between the *value* and the *cognitive cost* associated with the large-scale unstructured social media information. We aim to examine how search engine policies regarding social media content, such as what information to show on the search summary page versus product landing page, may affect consumer search/purchase behaviors and search engine revenues.

In addition, two recent papers, Ghose et al. (2012) and (2014), have also initialized their focus on the interplay of search engine and social media. This current paper distinguishes from these two studies in the following: (i) Ghose et al. (2012) studied only the consumer purchase decisions, not search/click decisions, whereas this paper jointly studies the click and purchase decisions. (ii) Both Ghose et al. (2012) and (2014) focused on only the "benefit" of social media on consumer evaluation of product quality for the purchase decision, but did not consider the "cognitive cost" associated with processing social media information during consumer search. This is one major unique advantage of this paper. None of the previous work has studied the "cost" of social media content in affecting consumer search and purchase decisions on product search engines. (iii) Both Ghose et al. (2012) and (2014) used aggregated data on click/purchase share at product level, while this paper models consumer decision at individual level. (iv) From a methodology perspective, different from Ghose et al. (2012) and (2014), this paper takes into accounts three unique constraints in the model: (1) interdependency in clicks/purchases among different products; (2) sequential arrival of information revealed by click-throughs; (3) non-negligible search costs.

A summary of the differences between this paper and the existing studies is in Table 1.

**Table 1. Comparison with Recent Literature**

| | **Kim et al. (2010)** | **Ghose et al. (2012)** | **Ghose et al. (2014)** | **Kim et al. (2014)** | **Koulayev (2014)** | **De los Santos & Koulayev (2014)** | **Chen & Yao (2016)** | **This Paper** |
|---|---|---|---|---|---|---|---|---|
| **Data** | Amazon, View-Rank, 18 months | Hotels, Purchase, ~8k observations, 3 months | Hotels, Click, Purchase, ~30k observations 3 months | Amazon, View-Rank, Sale-Rank, 18 months | Hotels, Click, Search Refinement, 1 month, (Chicago) | Hotels, Click, Search Refinement, 1 month, (Chicago) | Hotels, Click, Search Refinement, Purchase, 215 sessions, 15 days | Hotels, Click, Purchase, ~7M observations ~1M sessions, 2117 hotels, 3 months, |
| **Level of Analysis** | Market | Market | Market | Market | Individual Clicks at Page Level | Individual Clicks | Individual Clicks and Purchases | Individual Clicks and Purchases |
| **Real Transactions** | No | *Yes* | *Yes* | No | No | No | *Yes* | *Yes* |
| **Click Sequence** | No | No | No | No | *Yes* | *Yes* | *Yes* | *Yes* |
| **Search Refinements** | No | No | *Yes* | No | *Yes* | *Yes* | *Yes* | No |
| **Interplay with Unstructured Social Media** | No | *Yes* | No | No | No | No | No | *Yes* |
| **Update of Consumer Information Set** | No | No | No | No | No | No | No | *Yes* |
| **Major Objectives** | Consumer Welfare, Market Structure | Design a Novel Consumer Surplus-based Ranking for Search Engine | Causal Effect of Search Engine Ranking and Personalization | Market Structure, Innovation | Price Sensitivity | Design Search Engine Ranking by Maximizing Aggregate CTR | Search Refinement →Welfare Decrease | Interplay between Search & Social Media, Cognitive Cost of Unstructured Social Media |

## 3. Data

Clickstream and Transaction 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 7,059,122 observations from 969,033 individual sessions resulting in room bookings in 2,117 hotels in the United States.[2] A typical online session observed in our dataset 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. We observe the hotel listings displayed to the consumer during the search session (regardless of whether any click occurs). If a click occurs, we observe hotel listings the consumer observed prior to that click. Moreover, we also observe the sequence of the clicks.

Notice we also have detailed information associated with each event for every corresponding hotel, such as nightly room prices 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 the entire session-level consumer search data, including the final transaction price and the number of room units and nights purchased in each transaction. This information allows us to model consumer preferences for both the search and the purchase processes.

Hotel General Information: We collected hotel-related information from Travelocity.com, such as 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).

Hotel Location Information: In addition, we have independently collected supplemental data on hotel location-related characteristics using automatic social geo-mapping techniques together with image data mining. 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) in the area around the hotel. We use image classification methods together with human annotations (from Amazon Mechanical Turk, AMT) to extract whether a beach, lake, or downtown area is nearby, and whether the hotel is close to a highway or public transportation. We extract these characteristics from different zoom levels of the satellite images of a hotel location within a 0.25-, 0.5-, 1-, and 2-mile radius. We also collect local crime rates from FBI statistics.

Hotel Service Quality Information Extracted from Social Media: To fully exploit the information about hotel service quality, we combine text mining and sentiment analysis to examine the natural-language text of the customer reviews. For example, the helpfulness of the hotel staff is a service feature one can assess by reading the consumer opinions. Toward extracting such information, we build on the previous work of Archak

[2] In our dataset, 2,117 hotels had at least one booking during the data collection period. A total of 13,546 hotels had at least one display in consumer search sessions.

et al. (2011) and Ghose et al. (2012). First, we extract the important hotel features. Following the automated approach introduced previously (Archak et al. 2011, Ghose et al. 2012), we use a part-of-speech tagger to identify the frequently mentioned nouns and noun phrases, which we consider candidate hotel features. We then use WordNet (Fellbaum 1998) and a context-sensitive hierarchical agglomerative clustering algorithm (Manning and Schutze 1999) to further cluster the identified nouns and noun phrases into clusters of similar nouns and noun phrases. The resulting set of clusters corresponds to the set of identified product features mentioned in the reviews. For our analysis, we kept the top six most frequently mentioned features, which were *hotel staff*, *food quality*, *bathroom*, *parking facilities*, *bedroom quality* and *check-in/out front desk efficiency*.

For sentiment analysis, we extracted all the evaluation phrases (adjectives and adverbs) that were used to evaluate the individual service features (for example, for the feature "hotel staff," we extracted phrases such as "helpful," "smiling," "rude," "responsive"). The process of extracting user evaluation phrases can be automated. To measure the meaning of these evaluation phrases, we used AMT to exogenously assign explicit polarity semantics to each word. To compute the scores, we used AMT to create our ontology, with the scores for each evaluation phrase. Our process for creating these "external" scores was done using the methodology of Archak et al. (2011). Finally, to handle the negation (e.g., "I didn't think the staff was helpful"), we built a dictionary database to store all the negation words (e.g., "not," "hardly") using an approach similar to NegEx (http://code.google.com/p/negex; accessed Sept 10, 2015).

Consumer Cognitive Cost Indicators Extracted from Social Media: Although the textual content of customer reviews can reveal important information about hotel quality, there is a non-negligible cognitive cost associated with processing such information. To capture consumers' cognitive costs in reading the user-generated reviews, we analyzed two sets of review text features that are likely to affect consumers' intellectual efforts in internalizing review content: "readability" (i.e., textual complexity, syllables, and spelling errors) and "subjectivity" (i.e., mean and standard deviation). Research has shown both of them have had significant impact on consumer online shopping behavior in the past (e.g., Ghose and Ipeirotis 2011). To derive the probability of subjectivity in the review's textual content, we apply text-mining techniques. In particular, we train a classifier using the hotel descriptions of each of the hotels in our dataset as "objective" documents. We randomly retrieved 1,000 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/; accessed Sept 10, 2015). 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 to review textual readability and subjectivity, we also extracted an additional cognitive cost indicator based on the topic complexity of the customer reviews. In particular, built on prior literature (Gong et al. 2016) we analyzed the entropy value for the distribution of topics extracted from all customer reviews for each hotel ("Topic Entropy"). This entropy value measures the diversity of topics covered by the customer reviews for each hotel. Prior literature suggests the diversity in search results affects consumer search behavior (e.g., Weitzman 1979, Dellaert and Haubl 2012). In addition, consumer psychology theories suggest that as the information become noisier, users are more likely to abandon their search (e.g., Jacoby et al. 1974; Dhar and Simonson 2003), because users tend to get overwhelmed and discouraged by the complexity of information, and therefore lose their interest or trust in the search results. Therefore, we derived a Topic Entropy score using probabilistic topic models from machine learning and natural language processing to capture the "noisiness" of information provided by the customer reviews. Topic models are unsupervised algorithms that aim to extract hidden topics from unstructured text data. In particular, we measure the topic complexity of reviews for each product by estimating a topic model using Latent Dirichlet Allocation model (LDA; Blei et al. 2003), and subsequently computing the entropy (i.e. diversity) of the topic distribution of reviews for that product. We provide more technical details on the topic modeling in Online Appendix E.

For a better understanding of the variables, we present the definitions and summary statistics of all variables in Table 2. Note the dataset in this paper use not only the transaction data (i.e., purchases), but the complete session-level data (i.e., both clicks and purchases). The resulting dataset contains approximately seven million observations from one million individual user sessions.

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

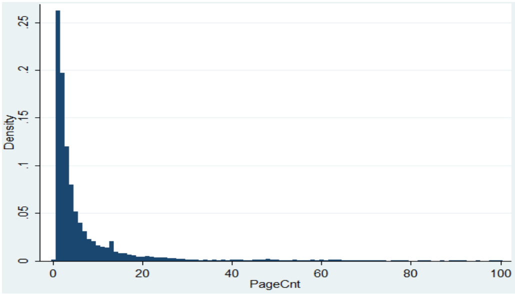
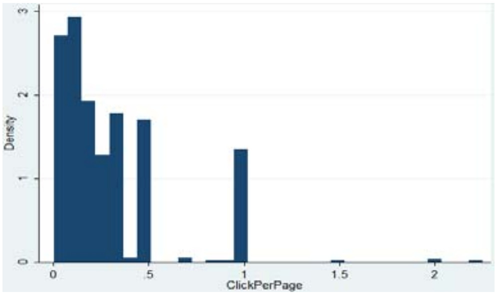


**Figure 1b. Distribution of #Click-thoughs Per Page (Session Level)**



### *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 a search session. Figure 1a 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 1b 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 consumers click less than 0.5 hotels per page, on average. Besides, over 97%

clicks occurred on the first page. These two figures provide us preliminary evidence that consumers incur non-trivial search costs and that consumer search is limited.[3]

## 4. A Structural Model of Consumer Sequential Search

Our dataset contains the complete information on the browsing session (e.g., list of hotels displayed, sequence of clicks) 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; and (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:

**Before Click:**

1. A consumer session starts with consumer browsing hotels on the search results summary page. A consumer can obtain any hotel information provided on the search results summary page (with no clicks needed) at zero cost.
2. Before clicking on a hotel, the consumer does not observe the exact information shown on the "details" landing page for that hotel. Instead, she forms a belief about what information would appear on the landing page, conditional on the information observed in the search results summary page. Because no click is needed to form the belief, we assume the consumer incurs zero cost at this step.
3. Given the observed information on the search results summary page and the conditional belief of the unobserved information on the landing page, the consumer is able to infer the *expected utility* of each hotel before the click-through at zero cost.
4. Meanwhile, before clicking on a hotel, the consumer also forms a belief about what the *expected search cost* would be if she were to click on the hotel (e.g., due to the additional cognitive efforts needed for processing the unstructured information on the landing page), conditional on the information observed from the search results summary page. Again, no click is required to form the belief of search cost, and we assume the consumer incurs zero cost at this step.

**After Click:**

5. The consumer session continues with a series of clicks, where the consumer decides to click on the landing pages of some hotels and to find out the exact utilities from these hotels. The goal of search (i.e., via click-through) is to reveal any uncertainty in the utility (i.e., uncertainty in the landing-page characteristics as well as the unobserved error). The set of clicked hotels and the order of the clicks reveal information about the preferences and search costs of the user.

[3] For some cities, the number of hotels might be small and therefore no additional page is available for searching. We find that 56 out of 4,845 cities (approximately 1.15%) in our data have less than 25 hotels (which means only one page is available for searching). After excluding these small cities, the model-free evidence shows a similar trend that consumer search is highly limited.

6. The consideration set is being generated during the search process. It contains all the hotels the consumer has clicked. After the costly click-through, the consumer knows the *actual* utilities (rather than the expected utility) of the clicked hotels, which form the consideration set.
7. The consumer stops searching new hotels (and hence stops clicking) when the expected marginal benefit of doing so is less than the expected search cost. We adopt the concept of "reservation utility" from Weitzman (1979) to define when the consumer stops searching. The decision of whether to continue searching or to stop relies on the actual utilities of the hotels in the consideration set at that moment [4] and her expected utilities and expected search costs of the upcoming hotels.
8. Once the consumer stops searching, the consideration set is fixed. Based on the final consideration set, the consumer makes a purchase decision (or skips purchasing anything at all).

### *4.1 Model Setting*

#### *(1) Product Utility.*

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

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

where $V_{ij} = V_{ij}^S + V_{ij}^L$ represents the hotel utility from the hotel characteristics displayed on the website. It consists of two conceptual components: (i) a deterministic component: $V_{ij}^S$, the exact utility from "summary-page" hotel characteristics consumers can directly observe on the search summary page, and (ii) a stochastic component: the *additional utility,* $V_{ij}^L$, from "landing-page" hotel characteristics consumers cannot directly observe before the click-through but can observe after the click-through. To evaluate the overall expected utility before the click-through, a consumer *i* forms a belief of the distribution of the unobserved landing-page utility $f(V_{ij}^L)$ based on $V_{ij}^S$. This belief comes from the consumer's knowledge about the utility distribution for hotel *j* conditional on the observed summary-page characteristics for this hotel. The consumer makes the click decision based on the exact value of the summary-page utility $V_{ij}^S$ and the *expected* value of the landing-page utility $E(V_{ij}^L)$. Once the consumer decides to click on the hotel, the click-through will reveal the actual value of the landing-page characteristics, and the consumer updates the expected value $E(V_{ij}^L)$ with the actual value $V_{ij}^L$. Moreover, we let $e_{ij}$ represent the unobserved uncertainty in the consumer's evaluation. The consumer does not know the realization of $e_{ij}$ unless she clicks on hotel *j* and visits its landing page. In particular, we assume $e_{ij}$ to be i.i.d. across consumers and hotels, and to follow a Type I Extreme Value distribution $e_{ij} \sim Type\ I\ EV(0,1)$ [5].

[4] In particular, the decision of whether to continue searching or to stop depends on the actual utility of the hotel with the maximum utility in the consideration set.

[5] Note that different from Kim et al. (2010), who assume standard normal distribution of the error, we allow for logit distribution of the error term in our model, as we assume that the consumers may optimize their utility over unobserved (to the econometrician) variables. In our estimation, we transform the logit error into standard normal disturbances using an inverse standard normal CDF

In summary, our utility setting assumes the consumer does not know the full realization of the utility of hotel *j* before the click-through. However, the consumer knows the distribution of the utility. This assumption is critical and is consistent with Weitzman (1979) and many recent studies that have examined consumers' sequential search behavior in the online search contexts (e.g., Kim et al. 2010, Chen and Yao 2016, Koulayev 2014). Hence the goal of search (i.e., click) is to solve the uncertainty in the consumer's evaluation toward both the landing-page characteristics and the unobserved error to reveal the true utility of a hotel.

More specifically, 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 results summary page. Thus, we can model the summary-page utility as $V_{ij}^S = X_j\beta_i - \alpha_i P_j$ , where $\beta_i$ and $\alpha_i$ are consumer-specific parameters capturing the heterogeneous preferences of consumers. We assume $\beta_i \sim N(\overline{\beta}, \Sigma_\beta)$ , where $\overline{\beta}$ is a vector containing the means of the random effects and $\Sigma_\beta$ is a diagonal matrix containing the variances of the random effects. Similarly, we assume $\alpha_i \sim N(\overline{\alpha}, \sigma_\alpha^2)$ .

Meanwhile, we model the expected value of the pre-click stochastic part of the utility as $E(V_{ij}^L) = \widetilde{L_j}\lambda_i$, where $\widetilde{L_j}$ represents the consumer expectation toward the landing-page characteristics for hotel *j* conditional on the observed summary-page characteristics $(X_j, P_j)$ . Note that $\widetilde{L_j}$ may not equal the actual values of the landing-page characteristics. We use the tilde sign to distinguish $\widetilde{L_j}$ from the realization of its deterministic value, $L_j$. Using a similar approach proposed by Koulayev (2014), we approximate $\widetilde{L_j}$ by taking the mean of the bootstrap samples from the actual information of the landing pages of the hotels that present the same summary-page characteristics. This approach allows consumers to infer knowledge about the utility distribution of a hotel based on the average knowledge from the population with similar experience (i.e., who are also exposed to $(X_j, P_j)$). The consumer estimates the expected utility of the landing page based on $\widetilde{L_j}$. She updates $\widetilde{L_j}$ with the deterministic value $L_j$ only after she chooses to click on hotel *j* and reveals the actual deterministic values of the landing-page characteristics. Let $\lambda_i$ represent consumer-specific parameter to capture the heterogeneity. Consistent with previous assumptions, we assume it follows a normal distribution $\lambda_i \sim N(\overline{\lambda}, \Sigma_\lambda)$.

Therefore, we have the overall utility function as follows. Before the click-through, the expected utility from hotel *j* for consumer *i* is

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

After the click-through, the realization of the actual utility becomes

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

function. This transformation approach was proposed and widely used by previous studies to compute the inverse Mill's ratio for logit distribution (e.g., Lee (1983), Greene (2002)). We provide more details in Online Appendix C. In addition, we have also tried the normal distribution assumption for the error term. We find our final results stay very consistent.

***(2) Search Cost.***

We model a consumer's search cost as a result of the landing-page-evaluation behavior associated with a click (i.e., cognitive cost of processing additional unstructured information). More specifically, let $Q_j$ denote the set of actual cognitive cost variables for evaluating the landing-page unstructured information of hotel *j*. We model the actual search cost of consumer *i* after clicking on hotel *j* to follow a lognormal distribution: [6]

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

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. To model the consumer's cognitive cost of evaluating the unstructured information on the landing page, we consider different dimensions in the cognitive-cost variables $Q_j$, including both the *readability* and the *subjectivity* of the textual content of online reviews.

However, because the landing-page information is not directly observable to the consumer before click, to decide whether to click on a hotel, the consumer needs to form a belief of her expected search cost conditional on the observed summary-page characteristics of that hotel. This means that in our model, $Q_j$ is not directly observable to the consumer before the click-through. Similarly, the consumer forms an expectation based on the observed summary-page characteristics. Let $\widetilde{Q_j}$ capture the consumer's expectation toward the unobserved cognitive-cost variables of hotel *j*. We approximate this expectation value by taking the mean of the bootstrap samples from the actual information of the hotels with the same summary-page characteristics.

Based on the discussion above, we can write the (expected) search cost of consumer *i* for hotel *j* before the click-through as the following:

$$c_{ij} = \exp(\widetilde{Q_j \gamma_i}). \tag{3b}$$

Thus, before the click-through of hotel *j*, a consumer *i* makes the click decision based on the expected search cost for *j*.

Note that a consumer's search cost is a sunk cost. It enters only the consumer's click decision process but not the purchase decision process. Once the consumer forms an evaluation about the expected search cost, she will make a click decision based on this evaluation one time, and will not need it again in the future. Therefore, the realized actual search cost after click-through in Equation (3a) does not enter either the click model or the purchase model in reality. Only the expected search cost before click-through in Equation (3b) will enter the model estimation process (i.e., click model). Hence, we can treat the consumer's expected search cost as a deterministic value in modeling her search (click) decision, which is consistent with Weitzman (1971). For simplicity of notation, we therefore keep the same notation $c_{ij}$ to denote the expected search cost, although the expected search cost in Equation (3b) represents an expectation value (based on $\widetilde{Q_j}$, not $Q_j$).

[6] 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 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: continue to search for the next alternative, or 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 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 optimal point for the consumer to choose the "stop" option.

More formally, let $S_i$ be the current search-generated consideration set (i.e., including all hotels consumer *i* has clicked). Let $u_i^*$ denote the current highest value obtained by consumer *i* so far. We define

$$u_i^* = \max_{j \in S_i} \{u_{ij}, 0\} . \tag{4}$$

Note we define $u_i^*$ as the highest value $u_{ij}$ consumer *i* obtains from the hotels in her consideration set. Given the current best value $u_i^*$, the expected marginal benefits for consumer *i* from searching *j* are

$$B_{ij}(u_i^*) = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) f_i(u_{ij}) du_{ij}, \tag{5}$$

where $f_i(\bullet)$ is the probability density function of hotel utility $u_{ij}$ and is individual specific. 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 $u_{ij}$ exceeds $u_i^*$. As we notice, the benefits of search depend only on the distribution of utility above $u_i^*$. Thus, for any hotel *j*, the reservation utility $z_{ij}$ meets the following boundary condition, where the *expected search cost* equals the *expected marginal benefits* from searching the hotel:

$$c_{ij} = B_{ij}(z_{ij}) = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f_i(u_{ij}) du_{ij}. \tag{6}$$

Note that in Equations (4)-(6), because the actual search cost and actual utility for an upcoming unsearched hotel *j* are not observable to consumer *i* before the click-through, her decision of whether to click on hotel *j* is based on her *expected search cost* and *expected utility*. By contrast, $u_i^*$ is derived based on the *actual hotel utilities* because after the click-through the consumer can observe the exact information about each hotel in her consideration set. Thus, when consumer *i*'s current best value 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 value 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 and the reservation utility are provided in Appendices B and C.

### *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 value $u_i^*$ within the consumer's current consideration set is lower than the reservation utility of hotel $r(j)$. Let $S_{i,r(j)}$ be the current consideration set generated prior to hotel $r(j)$. It includes all hotels the consumer has clicked before hotel $r(j)$. For a consumer to click hotel $r(j)$, $z_{i,r(j)}$ has to exceed the maximum value from the clicked sets of hotels. Thus we model the click probability of hotel $r(j)$ for consumer *i* as

$$\begin{aligned} \pi_{i,r(j)} &= \Pr\left[r(j) \textit{ is clicked by consumer } i\right] = \Pr\left[u_i^* < z_{i,r(j)}\right] \\ &= \Pr\left[\max_{m \in S_{i,r(j)}} (V^S_{i,r(m)} + V^L_{i,r(m)} + e_{i,r(m)}) < z_{i,r(j)}\right] \\ &= \prod_{m \in S_{i,r(j)}} F_{e_i}(z_{i,r(j)} - V^S_{i,r(m)} - V^L_{i,r(m)}), \ j > 1, \end{aligned} \tag{7}$$

where $F_{e_i}(\bullet)$ is the CDF of $e_{ij}$, which in our case $e_{ij} \sim TypeI\ EV(0,1)$.[7]

### *4.4 Conditional Purchase Probability*

Conditional on the sequence of clicks consumer *i* has made in the search session, we can derive the conditional probability that she purchases hotel r(*j*) in her consideration set as the following:

$$\begin{aligned} \eta_{i,r(j)} &= \Pr\left(r(j) \textit{ is booked by consumer } i \mid \textit{all clicks by consumer } i\right) \\ &= \Pr\left(u_{i,r(j)} \geq u_{i,r(j')} \mid \textit{all clicks by consumer } i, \ \ \forall r(j) \neq r(j'), \ \ \mathrm{r}(j), r(j') \in S_i\right) \\ &= \Pr\begin{pmatrix} V^S_{i,r(j)} + V^L_{i,r(j)} + e_{i,r(j)} \geq V^S_{i,r(j')} + V^L_{i,r(j')} + e_{i,r(j')} \mid \textit{all clicks by consumer } i, \\ \forall r(j) \neq r(j'), \ \ \mathrm{r}(j), r(j') \in S_i \end{pmatrix}, \end{aligned} \tag{8}$$

where *S_i* is the click-generated consideration set for consumer *i.* Note that because the consideration set *S_i* is selected by consumer *i* based on her search decisions, $e_{ij}$ does not follow a full Type I EV distribution. Instead, it follows a truncated Type I EV distribution based on the optimality conditions used by the consumer. Unfortunately, under such circumstances the conditional choice probability does not have a close-form expression (e.g., Logit form). To address this issue, we applied a simulation approach. Similar methods have been adopted by the previous studies (Chen and Yao 2016, Honka 2014, McFadden 1989). McFadden (1989) proposed a kernel-smoothed frequency simulator to sample the random draws from a truncated Type I EV distribution by smoothing the probabilities using a multivariate scaled logistic CDF (Gumbel 1961). Honka (2014) applied McFadden's approach to sample the error term from a truncated Type I EV distribution by

[7] Note that $u_i^*$ is the maximum utility value from the current consideration set $S_{i,r(j)}$. Hence, the value of $u_i^*$ depends on what products are included in the current stage of the consideration set.

taking into account the composition of the click-generated consideration set and the utility optimality of the final choice to model consumer simultaneous search. Chen and Yao (2016) applied a similar simulation approach to sample the error term from a truncated normal distribution by further accounting for not only the choice set composition and the utility optimality of the final choice, but also the sequence of the click-generated consideration set to model consumer sequential search. Our simulation approach builds on the methods from Chen and Yao (2016) and Honka (2014). It allows us to simulate the error term from a truncated Type I EV distribution by satisfying the follow three optimality conditions: 1) Sequence of the click-generated consideration set; 2) Composition of the click-generated consideration set; 3) Utility optimality of the final choice. We provide the full details on how we use the simulated method to construct the conditional purchase probability in Online Appendix D.

### *4.5 Likelihood Function*

We model the overall likelihood as the product of the probabilities of all the observed consumer clicks and purchases.

$$\begin{aligned} Likelihood &= \prod_i \Pr(CLICK_i,\ PURCHASE_i) \\ &= \prod_i \Pr(CLICK_i)\Pr(PURCHASE_i \mid CLICK_i), \end{aligned} \tag{9}$$

where $PURCHASE_i$ represents the observed purchase event by consumer *i*, and $CLICK_i$ represents the observed sequence of all click events by consumer *i*.

We can then model $\Pr(CLICK_i)$ and $\Pr(PURCHASE_i \mid CLICK_i)$ as follows. First, let *N* be the total number of hotels that consumer *i* has clicked (i.e., size of the consideration set) and *J* be the total number of hotels available in the market. We can model the joint probability of the sequence of click events for consumer *i* as the following:

$$\begin{aligned} \Pr(CLICK_i) &= \Pr\left[click_{i,r(1)},\ \text{then } click_{i,r(2)},\ldots,\ \text{then } click_{i,r(N)},\ \text{then } all\_unclicks_i\right] \\ &= \prod_{r(j)\in S_i^{clicked}}^{N} \Pr(u_{i,n} \le z_{i,r(j)},\ \forall n \in S_i^{clicked\_before\_r(j)}) \prod_{r(m)\in S_i^{unclicked}}^{J-N} \Pr(u_{i,N}^{*} > z_{i,r(m)}) \\ &= \prod_{r(n)\in S_i^{clicked}}^{N} \Pr(u_{i,n-1}^{*} < z_{i,r(n)}) \prod_{r(m)\in S_i^{unclicked}}^{J-N} \left[1-\Pr(u_{i,N}^{*} < z_{i,r(m)})\right] \\ &= \prod_{r(n)\in S_i^{clicked}}^{N} \pi_{i,r(n)} \prod_{r(m)\in S_i^{unclicked}}^{J-N} \left[1-\pi_{i,r(m)}\right], \end{aligned} \tag{10}$$

where $s_i^{clicked}$ represents the set of all hotels that have been clicked by consumer *i*, $s_i^{unclicked}$ represents the set of all hotels that have not been clicked by consumer *i*, and $s_i^{clicked\_before\_r(j)}$ represents the set of hotels that have been clicked by consumer *i* before *r(j)*.

Second, conditional on the sequence of click events, we can derive the conditional probability of the purchase event from Equation (8). Again, *Sᵢ* is the click-generated consideration set for consumer *i*.

$$\Pr(PURCHASE_i \mid CLICK_i) = \Pr\left(u_{i,r(j)} \geq u_{i,r(j')} \mid CLICK_i,\quad \forall r(j) \neq r(j'),\ \ \mathrm{r}(j), r(j') \in S_i\right) \\ = \eta_{i,r(j)} \tag{11}$$

Finally, based on Equations (10) and (11), we can rewrite the likelihood function as follows:

$$Likelihood = \prod_{i} \left\{ \eta_{i,r(j)} \prod_{r(n)\in S_i^{clicked}}^{N} \pi_{i,r(n)} \prod_{r(m)\in S_i^{unclicked}}^{J-N} \left[1-\pi_{i,r(m)}\right] \right\}. \tag{12}$$

With this model setting, we are able to account for the fact that the decision-making processes for the hotels in the same session are not completely independent from each other. Instead, the click and purchase decisions for a hotel depend not only on its own utility, but also on the prior sequence of clicks associated with the consideration set. [8]

### *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 summary page, including *Hotel Class*, *Hotel Brand*, *Customer Rating*, *Total Review Count*, *Page*, and *Rank*. We consider $L$ to contain all additional characteristics that can only be revealed from the hotel-landing page, including *Amenity Count*, *Number of Rooms*, *Number of External Amenities*, the top-6 service characteristics extracted from the social media textual content including *hotel staff*, *food quality*, *bathroom*, *parking facilities*, *bedroom quality* and *check-in/out front desk efficiency*, as well as locational factors such as *Beach*, *Lake*, *Downtown*, *Highway*, *Public Transportation*, and *Crime Rate*.

To analyze consumers' search costs, we consider $Q$ to contain different factors that capture the cognitive cost of the unstructured hotel information on the landing page. In particular, we consider both the *Readability* (i.e., complexity, syllables, and spelling errors) and *Subjectivity* (i.e., mean and standard deviation of the linguistic subjectivity) of the textual content of online reviews.[9]

Note the website also provides a sorting mechanism for consumers to refine their search by sorting the results under criteria other than the default sorting algorithm. Technically, if a consumer chooses to customize the sorting algorithm, her search cost for each hotel in the ranking list may also change, hence becoming endogenous to her own search behavior. However, in reality, we find that with approximately one million online search sessions in our dataset, more than 90% of these sessions do not involve any customized sorting behavior at all. This finding is consistent with previous randomized experimental results (Ghose et al. 2014) that the majority of users tend to stick with the default sorting method during online product search.

[8] Note that based on the model framework, we do not explicitly model the selection rule for the search order, but take it as pre-calculated (i.e., based on the Weitzman optimal search model, this search order is pre-calculated based on the descending order of the reservation utility of each product).

[9] Note that search cost on product search engine might be partly associated with the search engine design. To better account for this factor, in the main analysis we have controlled for the online positions of a hotel (i.e., Page and Rank on the search engine website), by which we aim to control for the search engine design efficiency to a large extent. Under such circumstance, our model estimated search costs indicate that conditional on the same online position, what the cognitive cost of searching a certain product is. In addition, we have conducted additional robustness tests by controlling for the sorting methods in a consumer's search session. This is to control for additional factors introduced by search engine design. We find in all these cases our model estimated results remain highly consistent.

Therefore, for simplicity in this study, we focus on only those search sessions conducted under the default sorting algorithm.[10]

To estimate our model, we derive the overall log-likelihood function as the following:

$$LL(\theta) = \sum_i \left\{ \ln\left(\eta_{i,r(j)}\right) + \sum_{r(n)\in S_i^{clicked}}^{1..N} \ln\left(\pi_{i,r(n)}\right) + \sum_{r(m)\in S_i^{unclicked}}^{N+1..J} \ln\left(1-\pi_{i,r(m)}\right) \right\}, \tag{13}$$

where $\theta$ represents the set of parameters of the random coefficients we aim to estimate:

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

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 click probability $\pi_{i,j}$ and conditional purchase probability $\eta_{i,j}$. To maximize the log-likelihood function $LL(\theta)$, we use a non-derivative-based optimization algorithm (i.e., the Nelder-Mead simplex method) for heuristic search.[11] This procedure iteratively searches for the optimal set of parameters $\{\theta^*\}$ until the log-likelihood function is maximized:

$$\{\theta^*\} = \underset{\{\theta^*\}}{\arg\min}\, LL(\theta). \tag{14}$$

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 2014). We provide more details of this computation procedure in Online Appendix C.

### *4.7 Identification*

One of the major challenges is to simultaneously identify consumers' heterogeneous preferences and search costs. 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, either the preferences for product characteristics or the moments of the search cost distribution can explain an observed search outcome. In our study, we need to identify four major effects: Consumer Preferences (Mean and Heterogeneity) and Consumer Search Cost (Mean and Heterogeneity). The key identification strategy of our estimation relies on the exclusion restriction 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,

[10] In principle, consumers can choose from various search strategies to customize their search results. To study this direction, one needs to separately look into the data under each different strategy. In this paper, our main focus is not on the search refinement strategies. We refer the readers to Chen and Yao (2016) and Koulayev (2014) for a more in depth analysis on that front.

[11] 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.

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.

Moreover, the identification also relies on search-cost "shifters." Recall that we model search cost as a function of a completely different set of variables compared to the consumer-preferences variables. When we choose different sets of covariates for search cost and consumer preferences, the covariates enter search cost function but not utility function serve as the exclusion restrictions for identification. Conditioned on the exclusion restrictions, the utility and the search cost can be separated. From an empirical identification perspective, we can simply view the search-cost variables as additional product characteristics (i.e., similar to Kim et al. (2010)). Thus we can identify the search-cost and consumer-preferences variables simultaneously.[12] We provide more detailed discussions below.

***(1) Mean Effects.***

We identify the mean effects of consumer preferences variables based on the correlation between the observed click/purchase frequencies and the frequencies of underlying preferences variables. In other words, we measure the mean effect of a consumer preference variable by how often the same (or similar) variable appears in the hotels consumers click or purchase. For example, if on average people tend to click (or purchase) low price hotels, we may conclude that people have a high price sensitivity. 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 the mean effect of consumer preferences.

We identify the mean effect of search cost partially based on the observed average size of the consumer's search-generated consideration set. Importantly, note that we model the search cost as a function of completely different variables compared to the consumer-preferences variables, 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 based on the correlation between the observed click frequencies and the frequencies of underlying search cost characteristics.

***(2) Heterogeneous Effects.***

Note that across both purchase data and search data, we have multiple observations per consumer. For a given consumer and her search cost, we observe the deviation of observed purchase and searches from those predicted decisions based on the mean preferences and search cost parameters. The distribution of these deviations across individual consumers allows us to identify the heterogeneity distribution parameters.

More specifically, we identify consumer heterogeneous preferences from two perspectives. First, we partially identify them from the search data based on the distribution of the deviations across individual

[12] One important fact to note is that we also observe rich variation in the characteristics of hotels that enter the consumer's consideration set. In particular, we find that among all the sessions in which consumers incur click-throughs, 8,731 sessions are associated with a size of (click-generated) consideration set that is larger than 5, and 3,506 sessions are associated with a size that is larger than 10. These observations are critical for our model identification.

consumers between our model's predicted click probabilities (based solely on the mean effects) and individual consumers' observed click probabilities. Second, since we also observe individual consumers' final purchases, the purchase data allow us to identify the heterogeneous preferences based on the distribution of the deviations across individual consumers between the model's predicted purchase probabilities (based solely on the mean effects) and individual consumers' observed purchase probabilities.

We identify the heterogeneous search cost through two sources. First, our identification relies on the exclusion restriction that search cost variables do not enter purchase decision processes. After identifying the consumer heterogeneous preferences 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 particular, at each point during a consumer's search, based on mean parameters, her reservation utility, and the products already searched in the consideration set, we can predict the mean probability of her stopping the search. The deviation of her search activities from the predicted values give us the information of one's heterogeneity in search cost. The distribution of these deviations across individual consumers identifies the search cost heterogeneity distribution parameters. Second, the nonlinear functional form in the reservation utility (i.e., Equation (6)) can also help identify consumer preference and search cost parameters (Kim et al. 2010). 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 also helps us separately identify consumer heterogeneous preferences and search cost.

## 5. Empirical Results

### *5.1 Main Results*

Our main results are shown in Table 3. First, we find 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 parameters ($\sigma_\alpha$, $\Sigma_\beta$, $\Sigma_\lambda$, $\Sigma_\gamma$), (for price, summary hotel characteristics, landing page hotel characteristics, and cost of absorbing social media content, respectively). Consistent with theory, *PRICE* has a negative effect on hotel demand. *CLASS*, *AMENITYCNT*, *ROOMS*, *RATING*, and *REVIEWCNT* each have a positive effect on hotel demand. For hotel-location characteristics, we find that *BEACH*, *TRANS*, *HIGHWAY*, and *DOWNTOWN* each has a positive effect on hotel demand, whereas *LAKE* and *CRIME* each shows a negative effect. Consistent with prior literature, online position has a significant effect on consumer click and demand (e.g., Yao and Mela 2011, Ghose and Yang 2009). In particular, *PAGE* and *RANK* each leads to a decrease in the hotel demand. Moreover, we find that three service variables that are extracted from social media textual content demonstrate significant effect on hotel demand. In particular, food quality presents the highest positive impact, followed by hotel staff and parking.

On the other hand, we find the additional unstructured information from the landing page indeed leads to an increase in consumer search cost. In particular, the readability-related review features such as

*COMPLEXITY*, *SYLLABLES*, and *SPELLERR* each have a positive sign, suggesting that long and complex sentences, words with many syllables, or spelling errors in user reviews discourage consumers from continuing to search on product search engines. Moreover, *SUB* has a positive sign, implying that highly subjective and opinionated content that lacks objective information creates a cognitive burden for consumers during hotel search and may lead to early termination of their search. Finally, *SUBDEV* also has a positive sign, which suggests that a mixture of both objective and subjective messages is likely to lead to higher cognitive costs. In other words, *SUBDEV* represents the standard deviation of the subjectivity value and it captures the level of heterogeneity in the type of information provided in the reviews. The higher the heterogeneity, the higher the cognitive cost associated with processing such information (i.e., when a review is a mix of both subjective and objective messages, it adds to the cognitive costs because readers might have to incur additional effort when switching between different types of information).

To get a handle on the actual magnitude of the search cost, we quantitatively derive the dollar value of different search cost variables. This 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 hotel costs \$6.18. The search costs differ across hotels from \$3.43 to \$7.75. Our findings are consistent with previous findings suggesting a non-trivial search cost in online markets. For example, Koulayev (2014) found the page-level median search costs rise from \$4 per first search to \$16 per fifth 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 \$4.00-\$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. In addition, de los Santos (2008) found search costs ranging from \$0.90 to \$1.80 per search in the online book industry. Meanwhile, a one-word increase in the average sentence length increases consumer search cost by \$0.44. One more syllable or one more spelling error per review can cost consumers \$0.56 or \$0.28, respectively, during the product search.

Importantly, our empirical analysis allows us to quantify the trade-off between consumers' benefits and costs toward leveraging social media information for decision making. Our results indicate that more social media information (especially textual content) may not always improve consumer decision making. Certain service and quality related information extracted from review textual content can indeed facilitate consumer decision making and impact product demand. However, due to the size and the unstructured nature of such information, it also brings in non-negligible cognitive costs to the consumers. Our study aims to explore a more effective and scalable way of managing social media information, which can help search engines extract and provide useful information to consumers without introducing high cognitive costs. Moreover, our model and policy experiments (in Section 6) allow us to evaluate the associate economic outcome on consumers as well as on product search engine revenues.

To further analyze the robustness of our model performance, and how social media and consumer heterogeneity (e.g., travel purposes) may affect the search cost and decisions of a consumer, we conduct three robustness tests by (1) excluding the social media variables from the main model, (2) including additional Topic Entropy variable into the main model, and (3) adding interaction effects between consumer travel purposes and summary-page variables. We find the estimated coefficients are qualitatively consistent with the main results. Interestingly, we notice the model that does not account for social media textual variables presents significantly higher price elasticity. This result indicates that the unstructured social media information plays an important role in consumer decision making, and that consumers' cognitive costs to digest such information are non-negligible. Without accounting for such unstructured information during consumer search can lead to an overestimation of price elasticity. We provide more details on the robustness tests in Online Appendix F.

### *5.2 Model Comparisons*

Furthermore, to understand how the type and scale of data or modeling mechanisms may affect the performance of our analysis, we conducted model comparison analyses with a set of alternative benchmark models using different data sets or modeling mechanisms.

#### *5.2.1 Alternative Models*

In particular, we considered four alternative benchmark models: (1) Alternative Model I: Use the purchase data only (Mixed Logit Model), (2) Alternative Model II: Use the purchase data only (Mixed Logit Model + Additional Search Cost Variables), (3) Alternative Model III: Use the click data only (Click Model)[13], and (4) Alternative Model IV: Use both the click and the purchase data (Joint Probabilistic Model of Click and Purchase + Additional Search Cost Variables, But No Click Sequence Information).[14] Due to space limitation, we provide the details on the alternative model mechanisms in Online Appendix G.

Overall, we find the estimation results are qualitatively consistent with our main findings. Interestingly, we find that using a static model without accounting for consumers' search behavior can lead to an overestimation of the price elasticity. 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 elasticity, implying 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 elasticity is consistent with prior findings by Koulayev (2014) 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 focus more on price search. Thus a search model that incorporates consumers' search behaviors may find a higher price elasticity of demand than a static model (e.g., de los Santos et al. 2012).

[13] In particular, we include only the click-sequence-related information in the likelihood function using the click data only. We estimate this click model using a similar simulated maximum likelihood approach based on only the click probability.

[14] The major difference between this join probabilistic model and our main search model is that instead of capturing the sequence of clicks and allowing clicks to be interdependent, the join model assumes each click decision to be independent. Correspondingly, it models the click decisions independently as following a discrete choice process (e.g., Logit model).

Furthermore, from Alternative Models (I) – (III), we find that using only the click data or only the purchase data are likely to overestimate the price elasticity, and therefore it is important to consider both click and purchase decisions when modeling consumer preferences. However, interestingly, from Alternative Model (IV), we find although incorporating both click and purchase decisions information can improve the model estimation, the joint probabilistic model without considering the click sequence information can still lead to an overestimation of price elasticity. This result indicates that not only the final click or purchase decisions matter, but also the sequential click path is critical in revealing consumer preferences. Failing to capture consumers' search paths can lead to an overestimation of price elasticity in the online search market. For more details on the alternative model results, we illustrate them in Tables G1 and G2 in Online Appendix G.

### *5.2.2 Model Prediction Experiments*

Based on the model-estimated coefficients, our final goal is to predict the click and purchase probabilities for a hotel by an individual consumer. The prediction of the two individual probabilities can be achieved by substituting the model-estimated coefficients into the Equations (7) and (8). To obtain individual-level consumer heterogeneity, we apply the Monte Carlo simulation method. In particular, we use the same random draws we simulated previously (i.e., in Sec. 4.6) from the joint distribution of the individual heterogeneous parameters. Based on the steps above, we are able to compute the corresponding individual click and purchase probabilities for each hotel for an individual consumer.

To examine the predictive performance of our model, we conduct a set of model-prediction experiments. We first compute the predicted individual click and purchase probabilities for each hotel as described above. Then we compare the predicted individual click and purchase probabilities with the observed click and purchase probabilities (i.e., observed search and choice shares for the hotels). We calculate the prediction error for each hotel at individual-session level for both click and purchase probabilities. Then we compute the root mean square error (RMSE) and mean absolute deviation (MAD). We consider all the four alternative models discussed above as our baseline models. Furthermore, we are interested in examining how the use of unstructured data (social media textual variables) may affect the model's predictive power. Therefore, we consider a fifth baseline model: main model without the social media textual variables (*Robustness Test 1*) for both click- and purchase-probability predictions.

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 hotel. The prediction results are illustrated in Tables 5a and 5b (click probability) and Tables 6a and 6b (purchase probability) in Appendix A. Our model-prediction results demonstrate our model has the overall highest predictive power. Our model outperforms the baseline models in both in- and out-of-sample predictive power for both click and purchase predictions. Similar trends in improvement in the predictive power occur with respect to RMSE and MAD.

For example, with regard to the click-probability prediction, the out-of-sample results in Table 5b show that with respect to the RMSE, our proposed model can improve the prediction performance by 14.92% compared to the search model without the social media textual variables. It can improve the prediction performance by 33.20% compared to the click model with only click data. We find a similar trend with regard to the purchase-probability prediction. For example, the out-of-sample results in Table 6b demonstrate that with respect to the RMSE, our main model can improve the prediction performance by 18.77% compared to the next best model—search model without the social media textual variables. It can improve the prediction performance by 37.36% compared to the Mixed Logit model with a limited consideration set, and by 32.81% compared to the Mixed Logit model with a limited consideration set plus the additional search cost variables.

Notice that the model-prediction experiments indicate our model is better able to predict the individual click and purchase probabilities for each hotel than the click model and the static Mixed Logit model. Even after considering various extensions of the Mixed Logit models accounting for the limited consideration set and the additional search cost variables, or considering other alternative behavioral models, our search model still provides the best predictive performance. The potential reasons are the following. Our proposed search model is a holistic model that captures both the click and the purchase decision making processes for a consumer. Therefore, our model is able to account for the following three unique features of consumer search: (1) *Interdependency in decision making*. The search model predicts that a click decision depends on the ordered list of previously clicked products, and consequently, a purchase decision depends on the previous click-generated consideration set; however, static models assume independent decision making during consumer evaluation. (2) *Information arrives sequentially*. Our model assumes that detailed product landing-page attributes can only become available to a consumer after she clicks on the product; however, static models tend to ignore the fact that information arrives sequentially and assume both landing-page and summary-page attributes are available to the consumer at the beginning. (3) *Non-negligible search cost*. The search model predicts that a click decision depends on the (expected) search cost associated with this click, and the formation of the final consideration set depends on the search cost towards each product; however, the static model ignores such opportunity cost.

In summary, three major indications from our model comparison experiments are: (i) Both the click and the purchase data reveal significant information about consumer preferences and search cost, and both are critical to improve the model predictive power. (ii) The *sequence* of the clicks reveals significant information about consumer preferences and search cost. Our main search model incorporates not only the click decisions but also the sequential order of these clicks. However, the static Mixed Logit models ignore the sequence of the clicks and simply take the final consideration set as exogenously given. (iii) Unstructured social media data play an important role in consumer decision making. Incorporating such information into the model can lead to a significant improvement in the model's predictive power.

## 6. Policy Experiment

Based on our model estimation results, we conduct counterfactual analyses under various policy experiments to explore the what-if type of questions. More specifically, considering the amount and type of different information, we are interested in what information product search engines should present during different stages of consumer search (i.e., on the search results summary page vs. product landing page).

### *6.1 Information Shown on the Search Summary Page vs. Landing Page*

As we notice in our dataset, most online products contain a large number of characteristics. However, due to the limitation in screen space, search engines are unable to show all product information on the search summary page. Instead, search engines choose to highlight a snapshot of some product information on the summary page, while leaving the majority of information to the landing page. The information selected for the search summary page for a product becomes critical because it can influence both consumers' perceptions of the utility of the product and their expectations regarding the search costs associated with further evaluation of the product (i.e., via click-through). [15]

To explore what information should be shown on the search summary page, we conduct a policy experiment using our model. In particular, we assume search engines show different sets of hotel characteristics on the summary page— meaning these chosen characteristics are directly observable to consumers before the click-through. We re-estimate consumers' conditional belief regarding the unobserved characteristics using bootstrap samples, and then compute the individual session-hotel-level predicted click and purchase probabilities based on the parameter estimates from the original model estimation. We compute the overall click and purchase probabilities for a hotel by taking an average of the click and purchase probabilities across all sessions for that hotel. Finally, we sum over all hotels based on the prices and the predicted purchase probabilities to compute the predicted search engine revenue.

By doing so, we aim to examine the following question: *Holding consumers' preferences for product characteristics and search cost variables consistent, how would consumers' click and purchase behavior change if the search engine websites were to provide different sets of information on the search results summary page?* Moreover, we are interested in exploring a better strategy for search engines to design the search results summary page such that it improves the overall click/purchase probabilities and the search engine revenue.

More specifically, we focus on six alternative sets of product information that may be potentially useful to show on the search results summary page: (1) Existing summary-page characteristics (i.e., price, hotel class, hotel brand, customer rating, review count, page, rank); (2) Existing summary-page characteristics plus additional location-related characteristics (i.e., # of external amenities, beach, lake, downtown, highway, public

[15] Note that we do not focus on the supply side model in the paper, and we make the implicit assumption that their information providing practices are exogenous and not necessarily optimal. We believe that this is a reasonable assumption for our model, but potentially more research in that direction could shed more light in the information provision decisions of the hotels and examine whether there is any strategic rationale for their actions in that front.

transportation, crime rate); (3) Existing summary-page characteristics plus additional service-related information (i.e., amenity count); (4) Existing summary-page characteristics plus additional review-text-related information (i.e., textual review features); (5) Existing summary-page characteristics plus additional review-topic-related information (i.e., Topic Entropy sore derived from the entropy measurement); (6) Existing summary-page characteristics minus the product price information.

**Figure 2a. Predicted Click Probabilities**
**with Different Information Provided on Search Summary Page**

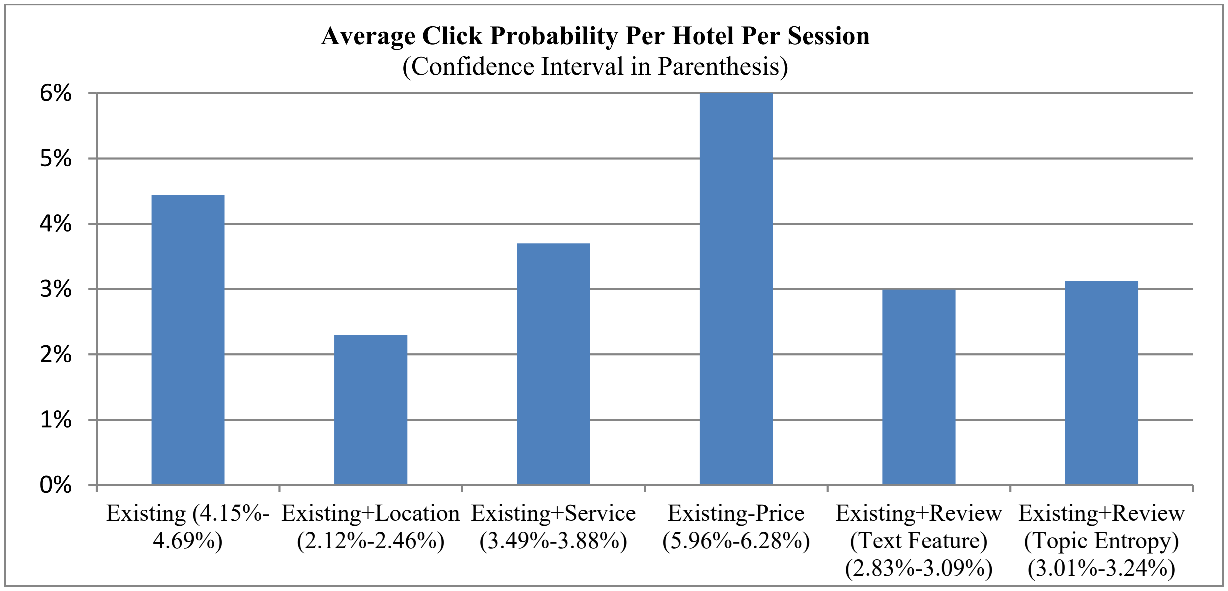
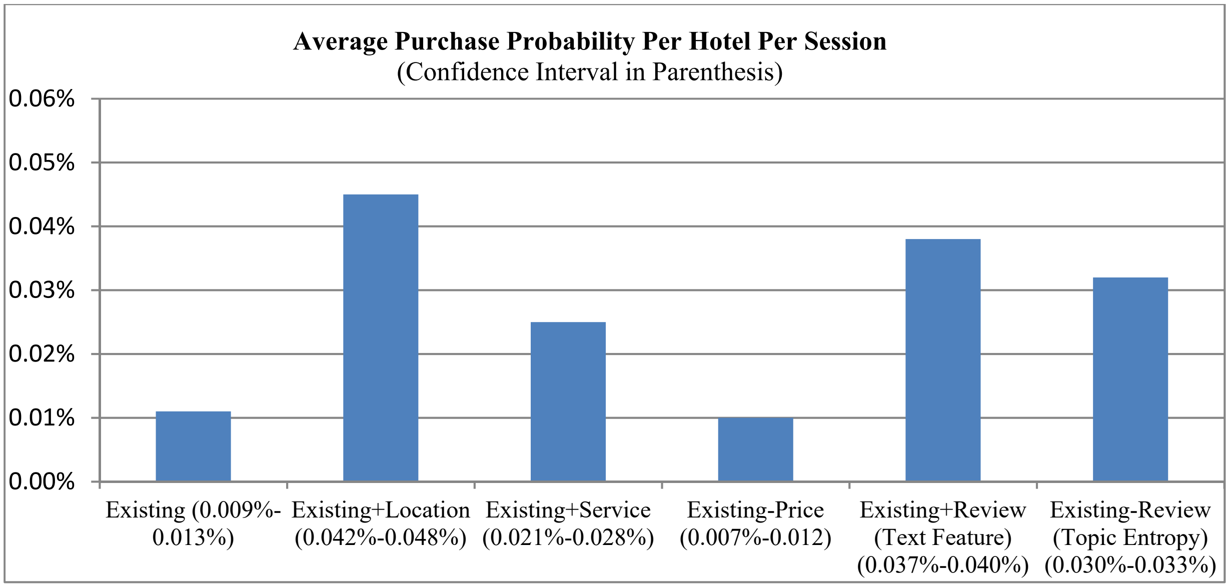


**Figure 2b. Predicted Purchase Probabilities**
**with Different Information provided on Search Summary Page**



We compute the average predicted click and purchase probabilities per hotel per session under each of the above six assumptions. We provide our results in Figures 2a and 2b. Our findings demonstrate that the type of information search engines choose to show on the summary page has a statistically significant effect on consumers' click and purchase probabilities. In particular, we find providing additional location-, service-, or review-related information for products on the search results summary page will lead to a significant decrease in click probability. This finding is intuitive. Because providing more information on the summary page will reduce the variance of the product utility (i.e., reducing uncertainty in consumer expectation) before click, it lowers the reservation utility of the product (hence making the product less attractive for a consumer to click).

However, interestingly, we find that providing additional product information, especially the location-related information, on the travel search engine summary page will lead to a significant increase in the purchase probability. A potential reason for this finding is that providing additional product information on the search summary page can reduce the potential error in consumers' expectation towards product utility and search costs before click. As a consequence, consumers are more likely to click on the best set of products that will provide them the highest utility. Hence, the maximum utility discovered from this click-generated consideration set is more likely to exceed the utility of the outside good. As a result, consumers are less likely to miss a good-value deal (i.e., leave without purchase).

Meanwhile, we find that although excluding price information from the search summary page can lead to a significantly higher click probability, it does not seem to increase the purchase probability at the end. This finding indicates that strategically hiding price information (i.e., price obfuscation) from the search summary page can make further searching (i.e., clicking) for products on a search engine more attractive. However, this strategy may not increase the overall purchase probability.

Finally, we compute the overall search engine revenue based on the hotel prices and the predicted purchase probabilities. Our results show that the location-related information is the most influential, compared to the service- and review-related information, when the travel search engine presents this information on the search summary page. It can lead to a 22.16% increase in the overall search engine revenue. Providing service-related information, such as the total number of hotel amenities, on the search summary page can lead to a 3.22% increase in the overall search engine revenue. By contrast, strategically hiding price information from the search summary page can hurt the search engine revenue, leading to a 7.08% drop in the overall revenue. We provide more details on the corresponding results in Table 4. [16]

Interestingly, providing a carefully curated digest of social media textual content on product summary page (e.g., top-6 most frequently mentioned product features extracted from the customer reviews, customers' attitudes towards these popular features, readability of the review's textual content) can lead to a 12.01% increase in the overall search engine revenue. Meanwhile, providing an overall "Topic Entropy" score of the

[16] We conducted additional analysis to examine the statistical significance in the difference across the simulated revenues in the policy experiments. In particular, given each different set of information on the search summary page, we replicated our simulation experiments for 200 times (i.e., via bootstrapping) to acquire the confidence interval of the corresponding simulated platform revenue. We found the predicted revenues under different scenarios are statistically different from the existing case (i.e., confidence intervals do not overlap).

review content (i.e., derived from topic models to measure the complexity of the review topic content) can lead to an 8.23% increase in the overall search engine revenue. These findings suggest that it is important for product search engines to leverage the economic value of large-scale unstructured social media information, while at the same time reducing the cognitive burden of consumers by automating the extraction of such information and providing it to consumers during the earlier stages of decision making.

**Table 4. Predicted Overall Search Engine Revenue with Different Information on Search Summary Page**

| | Overall Search Engine Revenue | 95% Confidence Interval * |
|---|---|---|
| *Existing* | $452,781 | $445,263 – $458,260 |
| *Existing + Location Information* | $553,136 | $538,026 – $561,989 |
| *Existing + Service Information* | $467,369 | $460,031 – $474,112 |
| *Existing – Price Information* | $420,132 | $411,585 – $429,203 |
| *Existing + Review Information (Text Features)* | $507,160 | $500,327 – $514,278 |
| *Existing + Review Information (Topic Entropy)* | $490,063 | $481,314 – $498,157 |

* Confidence Interval is calculated based on bootstrapping the policy simulation experiments for 200 times.

Furthermore, to examine where the revenue increase came from, we conducted an additional analysis on the breakdown of the revenue in the simulation. Interestingly, we found that the revenue increase came from both existing consumers and expansion of market coverage. In addition, we also found that the revenue increase occured for both existing hotels and new hotels. This finding provides further supports that with carefully designed information on search summary page, search engine can improve the market coverage of consumers as well as the diversity of products consumed, which can lead to a potential increase in consumer surplus. For more details, we provide the complete revenue breakdown analysis in Online Appendix H.

In sum, our policy experiment offers critical insights on the potential of analyzing large historical user behavioral data for search engines to improve the landing-page design strategy for better user experience and higher overall business revenues.

## 7. Managerial Implications and Conclusion

In this paper, we propose a structural econometric model for product search engines to understand consumers' search and purchase behavior as well as to quantify the search costs incurred by consumers. 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 unstructured social media information 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 results suggest that both the historical clicking decisions and the purchase decisions reveal significant information of consumer preferences and search costs. Moreover, the paths of searches (i.e., sequence of clicks) also reveal significant information of consumer preferences and search costs. Our analyses can help search engines predict consumer online footprints and design the search result summary page to improve user experience and search engine revenues.

On a broader note, our research makes two key contributions. *First*, we show the advantage of incorporating multiple and large-scale data sources to analyze how humans search, evaluate information, and make decisions under cognitive constraints in response to the emerging interplay between social media and search engines. Moreover, we are able to quantify the effects of unstructured social media content on user search cost. Our empirical analysis aims to provide an approach on which future studies can build, with the goal of exploring the potential of "Big Data" and sophisticated customer analytics tools for managerial decision-making. *Second*, we demonstrate the value of using digital analytics by search engines based on structural econometric methods in finding solutions for important business problems. Our structural model of 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 on search engines to identify the drivers of consumer decisions in electronic markets. It enables the prediction of consumers' future search behavior on search engines. Moreover, it offers insights to search engines on the design of the search results summary page (i.e., what information to show on the summary page vs. the landing page) to improve the user experience and the search engine revenues. Importantly, this approach can be generalized to any electronic market with an in-house search engine (e.g., Amazon.com), especially in a mobile search environment (e.g., Apple's iTunes or App store), given the commonality in the goal of improving user experience.

Our work has several limitations, some of which can serve as fruitful areas for future research. First, our model assumes the consumer knows the general distribution of utilities of alternatives, and each alternative follows the same distribution. However, when the alternatives are sorted on search engines under certain criteria including the default method, 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). Examining this fact from an empirical perspective would be interesting. Second, testing other alternative consumer behavioral models would be interesting. For example, instead of searching sequentially, consumers may search in a non-sequential fashion by first choosing a fixed size of a consideration set (e.g., Honka 2014). Comparing the differences in the corresponding model prediction of consumer search strategy would be interesting. Third, in this study, we assume each online consumer session to be an independent search process. Due to the data limitation, we cannot identify the possibility that a consumer may leave a session without booking but come back at a later time to resume the search. In this case, we treat these searches as two separate results in our estimation. Distinguishing such repeated searchers and more precisely estimating the search costs would be an interesting avenue for future research. Meanwhile, due to the data limitation, we do not have the consumer-level demographic information. Because the search cost is likely to relate to the opportunity cost of time, including such information (e.g., age, income) in future would be useful. Finally, it would be very interesting for future research to consider the supply side (e.g., how the hotels/advertisers may respond to the search engine's policy change) in addition to the demand side to examine the effects of policy change on search engines.

## 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).
- Archak, N., A. Ghose, P. G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Sci*. 57(8) 1485–1509.
- 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.
- Bikhchandani, S. and S. Sharma. 1996. Optimal Search with Learning, *Journal of Economic Dynamics and Control,* 20.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation, *Journal of Machine Learning Research* (3).
- Brynjolfsson, E., A. Dick and M. Smith. 2010. A nearly perfect market? Differentiation vs. price in consumer choice. *Quantitative Marketing and Economics*, vol.8, no.1.
- Chapelle, O., Zhang, Y. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. *Proceedings of WWW 2009*.
- Chen, P., Y. Hong and Y. Liu. 2017. The Value of Multi-dimensional Rating Systems: Evidence from a Natural Experiment and Randomized Experiments. *Management Science*, forthcoming.
- Chen, Y. and S. Yao. 2016. Sequential Search with Refinement: Model and Application with Click-stream Data. *Forthcoming in Management Science.*
- Chevalier, J. A., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3).
- De los Santos, B. 2008, Consumer search on the internet, *PhD dissertation*, Chicago University.
- De los Santos, B., A. Hortacsu, and M. Wildenbeest. 2012. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, 102(6), 2955-80.
- De los Santos, B., A. Hortacsu, and M. Wildenbeest. 2013. Search with Learning. *Working Paper.*
- De los Santos, B and Koulayev, S. 2014. Optimizaing Click-Through in Online Rankings for Partially Anonymous Consumers. *Working Paper.*
- Dellaert, B. G.C., G. 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.
- Dellarocas, C., N. Awad, M. Zhang. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4) 23–45.
- Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems,* 45(4) 1007–1016.
- Ellison, G. and Ellison, S.F. 2009. Search, Obfuscation, and Price Elasticities on the Internet. *Econometrica*.
- 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.
- Fellbaum, C. 1998. Wordnet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
- Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3) 291–313.
- 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*, 23 (10), 1498-1512.
- Ghose, A., Ipeirotis, P. and Li, B. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*. 31(3), 493-520.
- Ghose, A., Ipeirotis, P. and Li, B. 2014. Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue. *Management Science*. 60(7).
- 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.
- Godes, D., D. Mayzlin. 2004. Using online conversations to study word-of-mouth communication. *Mkt. Sci.* 23(4).
- Goldfarb, A. and Tucker, C.. 2011. Search Engine Advertising: Channel Substitution When Pricing Ads to Context. *Management Science*, 57:458-470.

- Gong, J., V. Abhishek, and B. Li. 2016. Examining the Impact of Contextual Ambiguity on Search Advertising Keyword Performance: A Topic Model Approach. *Working Paper*. 2016.
- Greene, W. H. 2002. *Limdep Manual*. Version 8.0. Econometric Software, Inc.
- 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, H. and M. Shum. 2006. Can search cost rationalize equilibrium price dispersion in online markets? *Rand Journal of Economics,* 37 (2): 258.276.
- Honka, E.. 2014. Quantifying search and switching costs in the U.S. auto insurance industry. *Rand Journal of Economics*.
- Hortacsu, A. and C. Syverson. 2004. Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds*, Quarterly Journal of Economics,* 119: 403.456.
- Iprospect. 2008. iProspect Blended Search Results Study. http://www.herramientas-seo.com/pdf/estudio-buscadores-iprospect.pdf.
- JupiterResearch. 2006. Retail Web Site Performance. http://www.akamai.com/4seconds.
- Kim, J., P. Albuquerque, B. Bronnenberg. 2010. Online Demand under Limited Consumer Search*, Marketing Science,* 29(6).
- Kim, J., P. Albuquerque, B. Bronnenberg 2014. The Effects of Product Innovation on Online Search and Choice. *Working Paper*.
- Koulayev, S. 2013. Search with Dirichlet Priors: Estimation and Implications for Consumer Demand, *Journal of Business & Economic Statistics* 31, pp. 226–239.
- Koulayev, S. 2014. Estimating Demand in Online Search Markets, with Application to Hotel Bookings. *RAND J. of Economics*.
- Lee, Lung-Fei. 1983. Generalized Econometric Models with Selectivity. *Econometrica* 51(2): 507-12.
- Manning, C., H. Schutze. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge.
- MarketingLand. 2013. Top Retail Websites Not Getting Faster: Average Web Page Load Time Is 7.25 Seconds. http://marketingland.com/retail-website-load-times-continue-to-decline-with-a-22-decrease-during-the-last-year-37604
- McFadden, D. 1989. A Method of Simulated Moment for Estimation of Discrete Response Models without Numerical Integration. Econometrica, Vol. 57, pp. 905-1026.
- McFadden, D. and K. Train. 2000. Mixed MNL Models of Discrete Response. *Journal of Applied Econometrics*.
- Mehta, N., S. Rajiv, and K. Srinivasan. 2003. Price uncertainty and consumer search: a structural model of consideration set formation, *Marketing Science,* 22(1).
- Moraga-Gonzalez, J. L., Wildenbeest, M. R.2008. Maximum likelihood estimation of search costs. *European Economic Review*, 52.
- Moraga-Gonzalez, J.L., Sandor, Z. and Wildenbeest, M.R. 2012. 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*.
- Netzer, O., R. Feldman, J. Goldenberg, M. Fresko. 2012. Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3) 521–543.
- Reinganum, J. F. 1982. Strategic search theory. *International Economic Review*. 23(1) 1–15.
- Rosenfield, D. B. and R. D. Shapiro. 1981. Optimal Adaptive Price Search, *Journal of Economic Theory*.
- Rothschild, M. 1974. Searching for the Lowest Price when the Distribution of Prices is Unknown, *J. of Political Economy* 82.
- Stigler, G.J. 1961. The Economics of Information. *The Journal of Political Economy*, 69(3), 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. *RAND J. of Economics*, 42(4).
- Yao, S., C. F. Mela. 2011. A Dynamic Model of Sponsored Search Advertising. *Marketing Science*, 30(3).

## Table 2. Definitions and Summary Statistics of Variables

| Variable | Definition | Mean | Std. | 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 |
| *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 |
| *BRAND* | Dummies for 9 hotel brands: Accor, Best western, Cendant, Choice, Hilton, Hyatt, Intercontinental, Marriott, and Starwood | -- | -- | 0 | 1 |
| *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, i.e., restaurants, shopping malls, or bars | 4.63 | 7.99 | 0 | 27 |
| *CRIME* | City annual crime rate | 194.99 | 127.22 | 3 | 1310 |
| **Social Media Variables (Cognitive Cost)** | | | | | |
| *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 |
| *TOPICENTROPY* | Entropy score to measure topic complexity | 2.88 | .13 | 1.58 | 2.99 |
| **Social Media Variables (Hotel Quality)** | | | | | |
| *REVIEWCNT* | Total # reviews | 13.56 | 25.60 | 0 | 202 |
| *RATING* | Overall reviewer rating | 3.94 | .39 | 1 | 5 |
| *STAFF* | Sentiment score for helpfulness of staff | .35 | .62 | -3 | 3 |
| *FOOD* | Sentiment score for food quality | .69 | .66 | -3 | 3 |
| *BATHROOM* | Sentiment score for bathroom quality | .42 | .74 | -3 | 3 |
| *PARKING* | Sentiment score for parking facilities | .16 | .58 | -3 | 3 |
| *BEDROOM* | Sentiment score for bedroom quality | .49 | .86 | -3 | 3 |
| *FRONTDESK* | Sentiment score for check-in/out front desk efficiency | .54 | .55 | -3 | 3 |
| **Model Computed Search Cost (in US Dollar $)** | | | | | |
| $c_j$ | Search Cost for a hotel *j* derived from the model estimation | 6.18 | .38 | 3.43 | 7.75 |
| **Total # Sessions:** | **969,033** | **Total # Hotels:** | | | **2117** |
| **Total # Observations:** | **7,059,122** | | | | |
| | **Time Period:** 11/1/2008-1/31/2009 | | | | |

## Table 3. Estimation Results - Main Model

| Variable | Mean Effect (Std. Err)$^{M}$ | Heterogeneity (Std. Err)$^{M}$ | Variable | Mean Effect (Std. Err)$^{M}$ | Heterogeneity (Std. Err)$^{M}$ |
|---|---|---|---|---|---|
| (Preferences) | $\overline{\alpha}$ , $\overline{\beta}$ , $\overline{\lambda}$ | $\sigma_{\alpha}$ , $\Sigma_{\beta}$ , $\Sigma_{\lambda}$ | (Preferences) | $\overline{\alpha}$ , $\overline{\beta}$ , $\overline{\lambda}$ | $\sigma_{\alpha}$ , $\Sigma_{\beta}$ , $\Sigma_{\lambda}$ |
| *PRICE*$^{(L)}$ | -1.252* (.022) | .417* (.074) | *DOWNTOWN* | 1.198* (.061) | .471* (.093) |
| *PAGE* | -.239* (.003) | .080 (.133) | *CRIME* | -.173* (.043) | .015 (.034) |
| *RANK* | -.314* (.008) | .132* (.067) | *RATING* | 2.661* (.015) | 1.308* (.091) |
| *CLASS* | 1.516* (.023) | .935* (.181) | *REVIEWCNT*$^{(L)}$ | 1.230* (.107) | .369* (.069) |
| *AMENITYCNT*$^{(L)}$ | .146* (.034) | .066 (.070) | *STAFF* | .139* (.027) | .034 (.088) |
| *ROOMS*$^{(L)}$ | .394* (.024) | .195 (.287) | *FOOD* | .225* (.038) | .136* (.002) |
| *EXTAMENITY*$^{L)}$ | .165* (.036) | .041 (.046) | *BATHROOM* | .290 (.271) | .060 (.103) |
| *BEACH* | 1.539* (.028) | .561* (.099) | *PARKING* | .097* (.008) | .075* (.011) |
| *LAKE* | -.663* (.116) | 1.560* (.389) | *BEDROOM* | -.175 (.232) | .253 (.269) |
| *TRANS* | 1.336* (.140) | .192* (.064) | *FRONTDESK* | .065 (.103) | .021 (.076) |
| *HIGHWAY* | .447* (.093) | .068 (.061) | | | |
| *BRAND* | *Yes* | | | | |
| (Search Cost) | $\overline{\gamma}$ | $\Sigma_{\gamma}$ | (Search Cost) | $\overline{\gamma}$ | $\Sigma_{\gamma}$ |
| *Search Base Cost (Constant)* | -7.511*(.089) | .971* (.176) | *SPELLERR*$^{(L)}$ | .329* (.082) | .033 (.101) |
| *COMPLEXITY* | .541* (.094) | .398* (.115) | *SUB* | .196* (.045) | .057 (.229) |
| *SYLLABLES*$^{(L)}$ | .678* (.115) | .721* (.106) | *SUBDEV* | .342* (.056) | .119 (.273) |
| *Maximum LL* | -405,418 | | | | |
| *Price Elasticity* | -1.619 | | | | |

*(L)* Logarithm of the variable. * Statistically significant at 5% *M: Main Model.*

## Appendix A. Model Comparisons

### Table 5a: In-sample Model Prediction Results (Click Probability)

| | Main Model | Main Model w/o Social Media Textual Variables | Click Model (with Only Click Data) | Joint Model of Click and Purchase (No Click Sequence) |
|---|---|---|---|---|
| **RMSE** | 0.0514 | 0.0588 | 0.0627 | 0.0613 |
| **MAD** | 0.0197 | 0.0221 | 0.0278 | 0.0262 |

### Table 5b: Out-of-sample Model Prediction Results (Click Probability)

| | Main Model | Main Model w/o Social Media Textual Variables | Click Model (with Only Click Data) | Joint Model of Click and Purchase (No Click Sequence) |
|---|---|---|---|---|
| **RMSE** | 0.1163 | 0.1367 | 0.1741 | 0.1541 |
| **MAD** | 0.0526 | 0.0614 | 0.0712 | 0.0658 |

### Table 6a: In-sample Model Prediction Results (Purchase Probability)

| | Main Model | Main Model w/o Social Media Textual Variables | Mixed Logit Model | | Joint Model of Click and Purchase (No Click Sequence) |
|---|---|---|---|---|---|
| | | | (Limited Consideration Set) | (Limited Consideration Set +Additional Search Cost Variables) | |
| **RMSE** | 0.0833 | 0.0912 | 0.1107 | 0.1074 | 0.0942 |
| **MAD** | 0.0274 | 0.0292 | 0.0392 | 0.0359 | 0.0312 |

### Table 6b: Out-of-sample Model Prediction Results (Purchase Probability)

| | Main Model | Main Model w/o Social Media Textual Variables | Mixed Logit Model | | Joint Model of Click and Purchase (No Click Sequence) |
|---|---|---|---|---|---|
| | | | (Limited Consideration Set) | (Limited Consideration Set +Additional Search Cost Variables) | |
| **RMSE** | 0.1251 | 0.1540 | 0.1997 | 0.1862 | 0.1662 |
| **MAD** | 0.0670 | 0.0729 | 0.0951 | 0.0845 | 0.0819 |