Modeling Consumer Footprints on Search Engines:
An Interplay with Social Media
(Online Appendices)

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

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

Online Appendix B.

Optimal Search Framework

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

$$u_i^* = \max{}_{j \in S_i} \{u_{ij}, 0\}. \tag{B1}$$

Note that Equation (B1) is the same as Equation (4) in the paper.

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

$$\Psi(u_i^*, \overline{S}_i) = \max\left(u_i^*, \max_{j \in \overline{S}_i} \left(-c_{ij} + d_i \cdot \left[\underbrace{\Psi(u_i^*, \overline{S}_i - \{j\}) \cdot \int_{-\infty}^{u_i^*} f(u_{ij}) du_{ij}}_{u_{ij} \le u_i^*} + \underbrace{\int_{u_i^*}^{\infty} \Psi(u_{ij}, \overline{S}_i - \{j\}) f(u_{ij}) du_{ij}}_{u_{ij} > u_i^*} \right] \right) \right), \tag{B2}$$

where $F(\bullet)$ is the CDF of u_{ij} and $f(\bullet)$ is the probability density function of u_{ij}. Therefore, at current state $(u_i^*, \overline{S}_i)$, the consumer can either terminate search and collect reward u_i^*, or search any $j \in \overline{S}_i$ to maximize $\Psi(u_i^*, \overline{S}_i)$. Given the short time span in online search, we set the discount rate d_i to 1. Equation (B2) is the principle of optimality for dynamic programming.

As pointed out by Weitzman (1979) and Lippman & McCall (1976) in the classical economic literature of search, the optimal solution to this dynamic programming has a *myopic* solution: Namely, the consumer needs only compare her return from stopping and accepting reward u_i^* with the expected return from exactly one more search.

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

$$B_{ij}(u_i^*) = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) f(u_{ij}) du_{ij}. \tag{B3}$$

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

$$c_{ij} < B_{ij}(u_i^*). \tag{B4}$$

Therefore, the optimal search strategy for a consumer is to continue searching until a value u_i^* is found that violates Equation (B4).

Online Appendix C.
Computation of Reservation Utility

Our model builds on the optimal sequential search framework by Weitzman (1979). Define the *reservation utility* z_{ij} as the utility value that satisfies the following boundary condition, where the search cost equates the expected marginal utility from searching product *j* (same as Equation (6)).

$$c_{ij} = B_{ij}(z_{ij}) = \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij}. \tag{C1}$$

The optimal search strategy for a consumer is to continue searching until she finds a value u_i^* larger than the boundary solution z_{ij}.

The reservation utility z_{ij} can be solved from Equation (C1) given the search cost, $z_{ij} = B_{ij}^{-1}(c_{ij})$. Weitzman (1979) has proved the function $B_{ij}(z_{ij})$ is continuous and monotonic. Therefore, there exists a unique solution z_{ij} to the equation $c_{ij} = B_{ij}(z_{ij})$.

Let $\overline{u_{ij}}$ be the mean and σ_{ij}^2 be the variance of the utility distribution $f(u_{ij})$. Based on our model setting, before the click-through we can write down the mean and the variance of the expected utility as the following:

$$\begin{aligned} \overline{u_{ij}} &= E[u_{ij}] = E(X_j\beta_i - \alpha_i P_j + \widetilde{L_j}\lambda_i + e_{ij}) \\ &= X_j\beta_i - \alpha_i P_j + \widetilde{L_j}\lambda_i + E(e_{ij}), \end{aligned} \tag{C2}$$

and

$$\begin{aligned} \sigma_{ij}^2 &= VAR[u_{ij}] = VAR(X_j\beta_i - \alpha_i P_j + \widetilde{L_j}\lambda + e_{ij}) \\ &= VAR(e_{ij}). \end{aligned} \tag{C3}$$

$\widetilde{L_j}$ is the expected value of the unobserved landing-page characteristics before click. It can be estimated based on the mean of the bootstrapping samples drawn from the empirical distribution of landing-page characteristics for hotel *j* conditional on the observed summary-page characteristics (X_j, P_j). Meanwhile, based on the assumption $e_{ij} \sim Type\ I\ EV(0,1)$, we can derive $E(e_{ij}) = 0.5772$ ("Euler-Mascheroni constant") and $VAR(e_{ij}) = \pi^2/6$. Therefore, $\overline{u_{ij}}$ and σ_{ij}^2 can be derived accordingly.

We rewrite Equation (C1) as follows:

$$\begin{aligned} c_{ij} &= \int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) f(u_{ij}) du_{ij} \\ &= (1 - F(z_{ij}))\left[\int_{z_{ij}}^{\infty} (u_{ij} - z_{ij}) \frac{f(u_{ij})}{1 - F(z_{ij})} du_{ij}\right], \end{aligned} \tag{C4}$$

where $F(z_{ij})$ is the CDF of u_{ij} evaluated at z_{ij}.

We compute the reservation utility using a similar approach as Kim et al. (2010).[1] Let $\eta_{ij} = \frac{z_{ij} - \overline{u_{ij}}}{\sigma_{ij}}$ and $\tau_{ij} = \frac{c_{ij}}{\sigma_{ij}}$, we can rewrite Equation (C4) as follows:

$$\tau_{ij} = \frac{c_{ij}}{\sigma_{ij}} = \frac{(1-F(\eta_{ij}))\left[\overline{u}_{ij} - z_{ij} + \sigma_{ij}\frac{f(\eta_{ij})}{1-F(\eta_{ij})}\right]}{\sigma_{ij}}$$

$$\Leftrightarrow \quad \tau_{ij} = g(\eta_{ij}) = (1-F(\eta_{ij}))\left[\frac{f(\eta_{ij})}{1-F(\eta_{ij})} - \eta_{ij}\right], \tag{C5}$$

If we can solve $\eta_{ij} = g^{-1}(\tau_{ij})$ from the above Equation (C5), then we can solve $z_{ij} = \eta_{ij}\sigma_{ij} + \overline{u_{ij}}$.

Note that Equation (B5) does not involve any model parameters. Therefore, in practice we only need to solve it once and use the results in the model estimation (Kim et al. 2010, Koulayev 2014). For computational tractability, we apply an interpolation approach to solve Equation (C5).

More specifically, we compute the reservation utility z_{ij} in the following four steps:

1) Pre-construct a lookup table for each pair (η_{ij}, τ_{ij}) based on Equation (C5).

2) At any stage in the estimation, use the current values of c_{ij} and σ_{ij} to compute $\tau_{ij} = \frac{c_{ij}}{\sigma_{ij}}$.

3) Based on the value of τ_{ij}, look up the corresponding value η_{ij} in the pre-constructed table.

4) Based on the current values of η_{ij}, σ_{ij} and $\overline{u_{ij}}$, solve the reservation utility $z_{ij} = \eta_{ij}\sigma_{ij} + \overline{u_{ij}}$.

[1] 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. To calculate the function with regard to the inverse Mill's ratio ($f(u_{ij}) / \left[1 - F(z_{ij})\right]$) from Equation (C4) to Equation (C5), we first need to transform the logit error into standard normal disturbances using an inverse standard normal CDF 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)).

Online Appendix D.

More Details on Using the Simulated Approach to Construct the Conditional Purchase Probability

As we discussed in section 4.4, 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:

$$\begin{aligned}\eta_{i,r(j)} &= P(r(j)\ is\ booked\ by\ consumer\ i)\\ &= \Pr\left(u_{i,r(j)} \ge u_{i,r(j')},\ \ \forall r(j)\neq r(j'),\ \ \mathrm{r}(j), r(j')\in S_i\right)\\ &= \Pr\begin{pmatrix} V^S_{i,r(j)} + V^L_{i,r(j)} + e_{i,r(j)} \ge V^S_{i,r(j')} + V^L_{i,r(j')} + e_{i,r(j')},\\ \forall r(j)\neq r(j'),\ \ \mathrm{r}(j), r(j')\in S_i\end{pmatrix},\end{aligned} \tag{D1}$$

where S_i is the click-generated choice set for consumer *i*. Equation (D1) is identical to Equation (8) in the paper.

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 circumstance the conditional choice probability does not have a close-form expression (e.g., Logit form). To address this selection issue, we applied a simulation approach. Similar methods have been adopted by the previous studies (Chen and Yao 2016, Honka 2014, McFadden 1989).

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 choice set; 2) Composition of the click-generated choice set; 3) Utility optimality of the final choice. More specifically, the simulated purchase probability has to satisfy the following conditions:

1) At any moment during the consumer search process, the utility of the currently being clicked product *j*, $u_{i,r(j)}$, is smaller than the reservation utilities of those products clicked after *j*. This is because the consumer continues to search afterwards. Here, $S_i^{clicked}$ denotes the set of all products that have been clicked by consumer *i*.

$$u_{i,r(j)} < \min(z_{i,r(j')},\ \forall r(j')\in S_i^{clicked} \text{ and } r(j') > r(j)) \tag{D2}$$

2) The utility of the final purchased product, $u_{i,r(j^*)}$, is greater than the reservation utilities of all the remaining unsearched products, $z_{i,r(j')}$. This is because the consumer stops searching afterwards. Here, $S_i^{unclicked}$ represents the set of all products that have not been clicked by consumer *i*.

$$u_{i,r(j^*)} \geq z_{i,r(j')}, \forall r(j') \in S_i^{unclicked} \tag{D3}$$

3) The utility of the final purchased product, $u_{i,r(j^*)}$, is greater than the utility of any other product in the click-generated choice set, $u_{i,r(j')}$. This is the final choice utility optimality condition.

$$u_{i,r(j^*)} \geq u_{i,r(j')}, \forall r(j') \in S_i^{clicked} \text{ and } r(j^*) \neq r(j') \tag{D4}$$

Hence, when simulate the error term in the utility function to construct the conditional purchase probability, we need to draw from the truncated Type I EV distribution by taking into consideration all the three optimality conditions (D2)-(D4) above. As discussed in Chen and Yao (2016), for different products in the click-generated choice set, the error terms are truncated differently. For clicked products that are not purchased, the error term is right truncated. For the purchased product, the error term is left truncated if it is the final click during the search process; if it is not the final search, then it is truncated on both sides.

To construct the conditional purchase probability, an intuitive approach is to draw the error term from the Type I EV distribution based on the three truncation optimality conditions in (D2)-(D4), by counting the frequency that the three optimality conditions are satisfied. More specifically, we adopted a similar approach as used by Chen and Yao (2016). The step-by-step implementation of our simulation method can be summarized as follows:

i) Conditional on a given set of other parameters in our model, for each consumer *i* and each product *j* in the consideration set, draw 200 $e_{i,j}$ from Type I EV distribution, depending on the three truncation optimality conditions;

ii) Count the frequency of the three optimality conditions (D2)-(D4) being satisfied across the 200 random draws of $e_{i,j}$'s;

iii) Iterate 100 times the above two steps, repeatedly making new draws of other parameters during each round of iteration;

iv) Average the simulated frequencies from step ii) across the 100 iterations to calculate the final simulated purchase probability of consumer *i*.

However, this approach can be computationally expensive. As a robustness check, we also tried an alternative method with a kernel-smoothed frequency simulator which was proposed by McFadden (1989) and was suggested by Honka (2014). In this approach, we smoothed the probabilities using a multivariate scaled logistic CDF (Gumbel 1961) with all the scaling factors equal to 15.[2] Notice that due to data limitation, Honka (2014) considers only the composition of the click-generated choice set but not the sequence of clicks as the optimality condition, whereas Chen and Yao (2016) observe the sequence of search process which allows them to consider both the composition and the sequence of the click-generated choice set. In our study, similarly as Chen and Yao (2016) we observe both types of information. Therefore, we are able to account for the additional optimality condition regarding the sequence of consumer clicks compared to Honka (2014).

[2] For more details on the kernel-smoothed frequency simulator, we refer interested readers to Online Appendix B in Honka (2014). The main idea of this simulator is to calculate the smoothed conditional purchase probability using a multivariate scaled logistic CDF (Gumbel 1961). In our estimation, we allow all the scaling factors to be equal to 15 as suggested by Honka (2014). However, we have also tried other values for the scaling factors ranging from 10-30. We found our estimation results stay qualitatively consistent.

Online Appendix E.
Using Topic Modeling to Generate the Topic Entropy Score for Each Hotel

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. This topic entropy 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. The intuition behind topic models is that a topic is a cluster of words that frequently occur together, and that documents, consisting of words, may belong to multiple topics with different probabilities. A probabilistic topic model tries to discover the underlying topic structure in a statistical framework. 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 discuss the details how we use topic modeling to generate the Topic Entropy score for each hotel below.

1. Corpus Construction and Document Pre-processing.

We first construct a corpus of documents that describe the information content conveyed by the hotel reviews. In particular, we collect all the customer reviews for each hotel. Hence, each hotel is associated with a review document, and all documents together construct the overall corpus. After constructing the corpus of review documents, we pre-process the documents following a standard procedure (e.g., Aral et al. 2011, Gong et al. 2016). We first remove annotations and tokenize the sentence into distinct terms. Then we remove stop words using a standard dictionary.

2. Latent Dirichlet Allocation (LDA).

We use topic models to automatically infer semantic interpretations of keyword meanings. The most widely used topic model is the Latent Dirichlet Allocation model (LDA; Blei et al. 2003), which is a hierarchical Bayesian model that describes a generative process of document creation. Previous research shows that humans tend to agree with the coherence of the topics generated by LDA, which provides strong support for the use of topic models for information retrieval applications.

The goal of LDA is to infer topics as latent variables from the observed distribution of words in each document. In particular, a topic is defined as a multinomial distribution over a vocabulary of words, a document is a collection of words drawn from one or more topics, and a corpus is the set of all documents. Based on the discussion above, we construct a document for each hotel that best reflects the information of the hotel review. We now discuss how we use LDA to infer the topics from the corpus of documents.

Formally, let *T* be the number of topics related to the corpus, let *D* be the number of documents in the corpus, and let *W* be the total number of words in the corpus. We assume that each document in the corpus is generated according to the following process:

Step 1. For each topic t, choose $\emptyset_t = (\emptyset_{t1}, \dots, \emptyset_{tW}) \sim Dirichlet(\varphi)$, where $\emptyset_t$ describes the word distribution of topic *t* over the vocabulary of words.

Step 2. For each document *d*, choose $\theta_d = (\theta_{d1}, \dots, \theta_{dT}) \sim Dirichlet(\omega)$, where θ_{dt} is the probability of topic *t* to which document d belongs.

Step 3. For each word *n* in document *d*, (1) choose a topic $t_{dn} \sim Multinomial(\theta_d)$, and (2) choose a word $w_{dn} \sim Multinomial\left(\emptyset_{t_{dn}}\right)$.

φ and ω are hyper-parameters for the two prior distributions - $Dirichlet(\varphi)$ as the prior distribution of $\emptyset$ (word distribution in a topic) and $Dirichlet(\omega)$ as the prior distribution of θ (topic distribution in a document). We use the values suggested by Steyvers and Griffiths (2007) (φ = 0.01 and ω = $50/\mathrm{T}$).

Based on the generative process described above, we use a Markov chain Monte Carlo (MCMC) algorithm to estimate $\emptyset$ and θ. Specifically, we use a collapsed Gibbs sampler to sequentially sample the topic of each word token in the corpus conditional on the current topic assignments of all other word tokens. We run a collapsed Gibbs sampler using MALLET (McCallum 2002) with 2,000 iterations. For each hotel, we obtain the posterior topic probabilities inferred from its corresponding document of customer reviews. In our study, we estimate the LDA model with a different number of topics, T= 20, 50, and 100.

3. Topic Entropy as a Measure for Keyword Ambiguity.

We propose using Topic Entropy to measure the complexity of hotel reviews. It captures the uncertainty of a document's topic distribution. In information theory, entropy measures the unpredictability of a random variable. In our context, each hotel is associated with its own review topic distribution inferred from the hotel-specific document. Therefore, we treat the topic assignment as a multinomial random variable, and use topic entropy to quantify how "noisy" the customer reviews for a hotel are in terms of underlying topics. The higher the entropy is, the more complex or noisier the reviews for that hotel. In other words, hotels with higher Topic Entropy tend to relate to a broader range of topics (more complex), whereas hotel with lower Topic Entropy tend to relate to fewer dominant topics (less complex).

More formally, let $\widetilde{\theta}_{kt}$ denote the posterior probability that hotel *k* belongs to topic *t*. We therefore define the topic entropy of hotel *k* as follows:

$$TopicEntropy_k = -\sum_{t=1}^{T} \tilde{\theta}_{kt} \log(\tilde{\theta}_{kt}), \tag{E1}$$

where T is the total number of topics.

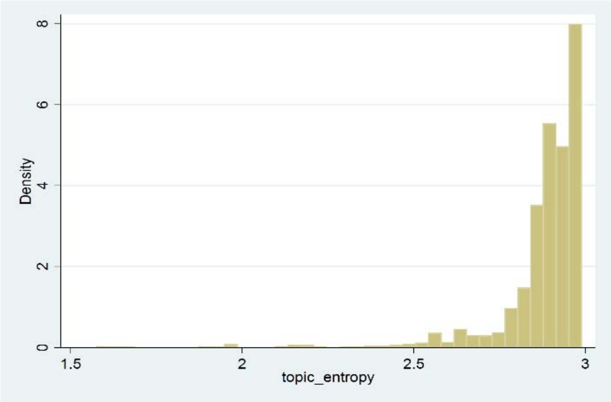
We present the summary statistics for the estimated Topic Entropy in Table E1. As we can see, the maximum entropy value depends on the number of topics chosen. Simple calculation also shows that with T topics, entropy ranges from 0 to *ln(T)*.[3] The high correlations among entropy values derived based on a different number of topics also suggest entropy seems to be fairly robust to the number of topics specified in the LDA model.

Table E1. Summary Statistics of Topic Entropy

	Mean	Std. Dev.	Min.	Max.	Correlation	
					20 Topics	50 Topics
20 Topics	2.78	0.13	1.58	2.99		
50 Topics	3.03	0.17	1.62	3.91	0.89	
100 Topics	3.16	0.20	1.68	4.60	0.88	0.93

In our model estimation (i.e., Robustness Test 2) and policy experiment, we used the Topic Entropy values derived based on 20 topics. We illustrate the distribution of the Topic Entropy in Figure E1 based on 20 topics (T=20). We also tried using 50 topics and 100 topics and the results are qualitatively consistent.

Figure E1. Distribution of Topic Entropy (*T*=20)



References:

- Steyvers, M., and Gri_ths, T. 2007. Probabilistic Topic Models. Handbook of Latent Semantic Analysis (427:7), pp. 424-440.
- McCallum, A. K. 2002. MALLET: A Machine Learning for Language Toolkit.

[3] The number of topics T is pre-specified before estimating the LDA model. Entropy for hotel *k* is the smallest when there exists $t \in \{1, \dots, T\}$ such that $\tilde{\theta}_{kt} = 1$; Entropy is the largest when for all $t \in \{1, \dots, T\}$, $\tilde{\theta}_{kt} = 1/T$.

Online Appendix F.
Robustness Tests (1) – (3)

To understand the robustness of the model estimation, and to analyze how unstructured social media and consumer heterogeneity (e.g., travel purposes) may affect the search cost and decisions of a consumer, we conduct three sets of robustness tests:

***Robustness Test 1:* Exclude the unstructured social media data (i.e., no social media textural variables in utility or search-cost specifications).**

One of the main goals in our study is to examine how social media textual content affects consumer utility and search cost. Therefore, we are interested in comparing the differences in the search models with and without the set of social media textural variables. The results of this test are illustrated in Table F1, column 2 ("R1"). We find the estimated coefficients are qualitatively consistent with the main results. After computing the price elasticity, we notice the model that does not account for social media textual variables presents significantly higher price elasticity (2.128 vs. 1.619, $p<0.05$). This result indicates that the unstructured social media textual 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 product search can lead to an overestimation of price elasticity.

***Robustness Test 2:* Include additional unstructured social media variable (i.e., "Topic Entropy" measurement derived from the online review textual content using topic modeling).**

To further understand the role of unstructured social media content during consumer search, in addition to the readability and subjectivity of review textual content, we also extracted an additional cognitive cost indicator based on the topic complexity of the customer reviews, "Topic Entropy." In Robustness Test 2, we included this new social media variable into the search cost model and re-estimated our model. The results of this test are illustrated in Table F1, column 3 ("R2"). Overall, we find the estimated coefficients are qualitatively consistent with the main results. Moreover, we find the increase of Topic Entropy in the customer reviews for a hotel can lead to a significant increase in the search cost for that hotel. Intuitively, this result indicates that as the topics discussed in customer reviews become noisier, it can significantly increase the cognitive costs for consumers who are reading through them. Therefore, it might be more efficient for product search engines to provide a careful digest of the topics extracted from all the textual reviews, or to provide a guidance on the expected topics to be discussed from the reviewers (e.g., room service, friendliness of the staff, parking facilities, etc.).

Table F1. Estimation Results - Robustness Tests (1) & (2)

Variable	Mean Effect (Std. Err) ^{R1}	Heterogeneity (Std. Err) ^{R1}	Mean Effect (Std. Err) ^{R2}	Heterogeneity (Std. Err) ^{R2}
(Preferences)	$\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$	$\sigma_\alpha, \Sigma_\beta, \Sigma_\lambda$	$\overline{\alpha}$, $\overline{\beta}$, $\overline{\lambda}$	$\sigma_\alpha, \Sigma_\beta, \Sigma_\lambda$
PRICE$^{(L)}$	-1.706* (.029)	.490* (.079)	-1.249* (.023)	.423* (.071)
PAGE	-.187* (.003)	.099 (.158)	-.237* (.003)	.082 (.130)
RANK	-.250* (.008)	.189* (.057)	-.317* (.007)	.135* (.066)
CLASS	1.614* (.039)	.806* (.112)	1.511* (.021)	.934* (.180)
AMENITYCNT$^{(L)}$	.156* (.034)	.039 (.080)	.142* (.032)	.065 (.072)
ROOMS$^{(L)}$	.343* (.031)	.238 (.351)	.397* (.022)	.193 (.281)
EXTAMENITY$^{L)}$	.172* (.041)	.039* (.012)	.166* (.035)	.042 (.044)
BEACH	1.890* (.020)	.503* (.095)	1.541* (.028)	.560* (.097)
LAKE	-.784* (.118)	1.075* (.303)	-.667* (.115)	1.555* (.383)
TRANS	1.243* (.171)	.160 (.133)	1.339* (.141)	.197* (.065)
HIGHWAY	.399* (.112)	.055 (.042)	.443* (.091)	.071 (.063)
DOWNTOWN	.962* (.062)	.206 (.074)	1.195* (.063)	.475* (.094)
CRIME	-.159* (.033)	.020 (.053)	-.171* (.045)	.018 (.031)
RATING	2.898* (.020)	.983* (.082)	2.660* (.017)	1.309* (.089)
REVIEWCNT$^{(L)}$	1.653* (.129)	.437* (.081)	1.228* (.106)	.366* (.062)
STAFF	----	----	.135* (.028)	.035 (.082)
FOOD	----	----	.223* (.034)	.138* (.002)
BATHROOM	----	----	.296 (.270)	.067 (.101)
PARKING	----	----	.097* (.005)	.079* (.014)
BEDROOM	----	----	-.179 (.236)	.251 (.271)
FRONTDESK	----	----	.066 (.108)	.021 (.070)
BRAND	Yes		Yes	
(Search Cost)	$\overline{\gamma}$	$\sum_\gamma$	$\overline{\gamma}$	$\sum_\gamma$
Search Base Cost (Constant)	-4.849* (.081)	1.303* (.124)	-7.976*(.095)	.904* (.168)
TOPICENTROPY	----	----	.298* (.036)	.334* (.097)
COMPLEXITY	----	----	.512* (.088)	.373* (.102)
SYLLABLES$^{(L)}$	----	----	.633* (.121)	.705* (.099)
SPELLERR$^{(L)}$	----	----	.286* (.085)	.039 (.111)
SUB	----	----	.187* (.042)	.063 (.232)
SUBDEV	----	----	.302* (.054)	.123 (.267)
Maximum LL	- 338,301		-409,742	
Price Elasticity	-2.128		-1.615	

(L) Logarithm of the variable. * Statistically significant at 5% level.

R1: Robustness Test 1 (Main Model Without Social Media Cognitive Cost Variables).

R2: Robustness Test 2 (Main Model With Additional Topic Entropy Variable to Measure Topic Complexity)

Robustness Test 3: Interaction effects between travel purposes and consumer preferences/search cost variables.

To account for consumer heterogeneity during the search process, we focus on how consumers' heterogeneous travel purposes explain certain variation in search cost and in consumer preferences. To do so, we investigate the interaction effects between consumer travel purposes and consumer preferences/search cost variables. In particular, the variables on which we focus are the summary-page variables. To capture consumers' heterogeneous travel purposes, we define T_i as an indicator vector with identity components representing the travel purpose:

$$T_i' = [Family_i \;\; Business_i \;\; Romance_i \;\; Tourist_i \;\; Kids_i \;\; Senior_i \;\; Pets_i \;\; Disability_i]_{1\times 8}. \qquad \text{(F1)}$$

We acquire the empirical distribution of T_i from online consumer reviews and reviewers' profiles.[4]

Then we interact the summary-page variables with T_i in our search model. We estimate this model using a simulated maximum likelihood approach. We find that consumers' travel purposes can explain their heterogeneity toward specific search-cost and preferences variables. Interestingly, we find interesting interaction patterns between consumers' travel purposes and hotel characteristics. For example, we find that travelers on a romantic trip, relative to other types of travelers, tend to place more importance on online customer reviews (i.e., both the valance of online ratings and the volume of reviews). In addition, consistent with prior research (Ghose et al. 2012), we find that business travelers are the least price-sensitive, whereas tourists tend to be more sensitive to hotel price. We provide the detailed information on the estimated interaction effects in Table F2 ("R3").

Table F2. Estimation Results - Robustness Test (3)
Search Model with Interaction Effects Between Travel Purposes and Summary-Page Variables

Variable	**Mean Effect (Std. Err)**^{R3}	*Family*	*Business*	*Romance*	*Tourist*	*Kids*	*Senior*	*Pets*
PRICE$^{(L)}$	-1.169* (.024)	**-.118*(.033)**	**.331*(.025)**	.123(.081)	**-.219*(.049)**	----	-.314(.257)	----
PAGE	-.212* (.021)	.005(.020)	**-.041*(.010)**	**.035*(.003)**	.021(.024)	----	----	----
RANK	-.288* (.034)	**-.037* (.008)**	**-.025* (.003)**	.022 (.021)	.154(.166)	-.011(.028)	----	----
CLASS	1.432* (.036)	**.067*(.021)**	**.092*(.022)**	**.065*(.016)**	**-.179*(.023)**	.200 (.363)	----	----
RATING	2.487* (.021)	.183(.226)	-.368(.399)	**.395*(.033)**	.040 (.057)	**.291*(.026)**	**.202*(.053)**	----
REVIEWCNT(	1.315* (.043)	**.177*(.023)**	-.256(.219)	**.301*(.042)**	**.123*(.026)**	----	----	----

(L) Logarithm of the variable. * Statistically significant at 5% level.

R3: Robustness Test 3 (Search Model with Interaction Effects).

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

[4] After writing an online review for a hotel, a reviewer is asked to provide additional demographic and trip information—e.g., "What was the main purpose of this trip? (Select one from the eight choices.)" We derive the distribution of T_i based on reviewers' responses to this question.

Online Appendix G.
Model Comparisons – Details on the Alternative Models

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. In particular, we considered a set of alternative benchmark models using different data sets or modeling mechanisms. We discuss the details of the alternative modeling mechanisms in this appendix.

(1) Alternative Model I: Use the purchase data only (Mixed Logit Model).

In reality, due to the unavailability of the individual-level click stream information, we may have access to only the purchase information. Classical static demand estimation models (e.g., Mixed Logit) are used to infer consumer preferences from the purchase data only. However, static demand estimation models do not consider the endogenous and limited nature of search-generated choice sets. With the recent growing pervasiveness of Internet, Web 2.0 and storage technologies, businesses have started tracking individual-level data beyond the final purchase to analyze a consumer's "path-to-purchase." However, individual-level click stream data are often at a much larger scale than the traditional purchase data and hence require more advanced methodologies and computing power for analysis.[5]

To examine how well our proposed search model performs by incorporating the additional click information in understanding consumer preferences, we consider a model that is widely used in the static demand estimation: the Mixed Logit model (e.g., McFadden and Train 2000). [6] To account for the variation in choice sets, we model the consumer decision process under the actual searched (limited) choice set, rather than under the universal choice set available in the market. Note the major difference between a static Mixed Logit model with actual choice sets and our proposed model is that our model captures not only the limited nature of the choice sets, but also the sequential and endogenous formation process of the choice sets. A static model typically takes the choice set as exogenously given.

Interestingly, we find that using a static model without accounting for consumers' search behavior can lead to an overestimation of the price elasticity (2.973 vs. 1.619, $p<0.05$). 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

[5] For example, our original click stream data set contains approximately a total of seven million observations, whereas focusing only on the purchase data will reduce our total number of observations to about only eight thousand.

[6] Note that the conditional purchase probability $\eta_{i,j}$ here has a close form as defined in the Mixed Logit model, because the consideration set in this case is not endogenously generated, but exogenously given. Therefore, there is no selection issue and the error term follows its initial Type I EV distribution.

show that when consumers face a highly differentiated market (e.g., product differentiation or retailer differentiation), they are more likely to focus on non-price factors during search. Hence the estimated price elasticity is lower when incorporating consumers' search behaviors into the model. On the contrary, when a market is less differentiated, consumers become more price-sensitive and tend to focus on price search. Thus a 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). The estimation results of this model are shown in Table G1, column 3 ("A1"). For easy comparison with the main model, we also provide the estimation results from our main model in Table G1, column 2 ("M").

(2) Alternative Model II: Use the purchase data only (Mixed Logit Model + Additional Search Cost Variables).

One concern towards the validity of *Alternative Model I* is that the static Mixed Logit model does not consider the search cost, so any difference in the estimation is likely to be caused by the missing variables that appear in the search cost from the search model. For example, the estimated parameter increase in price coefficient may be due to the correlation between search and prices—consumers search more for high-priced goods. Hence, inferences regarding price effects are likely to be biased when one does not control for quality (which may be related to the attributes that show up in the search-cost function). Therefore, we consider an alternative Mixed Logit model by incorporating the additional search cost variables. We provide the corresponding results in Table G2, column 2 ("A2").

We find the estimates from *Alternative Model II* are qualitatively consistent with our main model estimation results. Moreover, we also see a similar trend that the price elasticity from *Alternative Model II* is larger than the one from the main search model. This additional model provides further support that in a highly differentiated market (e.g., hotel market), ignoring consumers' search information in the demand model can lead to an overestimation of the price elasticity.

(3) Alternative Model III: Use the click data only (Click Model).

On the other hand, we sometimes have access to only the publicly available click-through data (but no purchase information). Therefore, finding out, given only the publicly available data, how well our search model can predict consumers' click behavior is important. For this purpose, we consider a "click model" in the robustness test. In particular, we include only the click-sequence-related information in the likelihood function using the click data only, as shown below in Equation (G1). We estimate this click model using a similar simulated maximum likelihood approach based on only the click probability. We find that the estimated coefficients are qualitatively consistent with the main results. For more details on the estimates from the click model, we provide the results in Table G1, column 4 ("A3").

$$\begin{aligned} Likelihood &= \prod_i \Pr(all\_clicks_i,\ all\_noclicks_i) \\ &= \prod_i \left\{ \prod_{r(n)\in S^{i,clicked}}^{1..N} \pi_{i,r(n)} * \prod_{r(m)\in S^{i,unclicked}}^{N+1..J} \left(1-\pi_{i,r(m)}\right) \right\}. \end{aligned} \tag{G1}$$

(4) <u>Alternative Model IV:</u> Use both the click and the purchase data (Joint Probabilistic Model of Click and Purchase + Additional Search Cost Variables, But No Click Sequence Information).

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, it is not clear the improvement in the model performance is attributed to the advantage of our holistic search model or simply to the use of more data. To examine this issue, we consider another alternative model − a joint probabilistic model of both click and purchase. 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). The results of this model are illustrated in Table G2, column 3 ("A4").

We find that the estimation results are qualitatively consistent with our main findings. However, interestingly we find that 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 (1.954 vs. 1.619, p<0.05). This result indicates that not only the final click or purchase decisions matter, but also the sequential click path of consumer search 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.

Table G1. Estimation Results - Main Model and Alternative Models (I) & (III)

Variable	Mean Effect (Std. Err)^{M}	Heterogeneity (Std. Err)^{M}	Mean Effect (Std. Err)^{A1}	Heterogeneity (Std. Err)^{A1}	Mean Effect (Std. Err)^{A3}	Heterogeneit y (Std. Err)^{A3}
(Preferences)	$\overline{\alpha}, \overline{\beta}, \overline{\lambda}$	$\sigma_{\alpha}, \Sigma_{\beta}, \Sigma_{\lambda}$	$\overline{\alpha}, \overline{\beta}, \overline{\lambda}$	$\sigma_{\alpha}, \Sigma_{\beta}, \Sigma_{\lambda}$	$\overline{\alpha}, \overline{\beta}, \overline{\lambda}$	$\sigma_{\alpha}, \Sigma_{\beta}, \Sigma_{\lambda}$
PRICE$^{(L)}$	-1.252* (.022)	.417* (.074)	-2.391* (.038)	1.064* (.082)	-1.744* (.029)	.472* (.088)
PAGE	-.239* (.003)	.080 (.133)	-.283* (.002)	.142 (.261)	-.206* (.003)	.117* (.140)
RANK	-.314* (.008)	.132* (.067)	-.341* (.008)	.138 (.211)	-.268* (.009)	.190* (.037)
CLASS	1.516* (.023)	.935* (.181)	1.882* (.012)	1.060* (.271)	2.057* (.020)	.574* (.128)
AMENITYCNT(	.146* (.034)	.066 (.070)	.212* (.051)	.059 (.126)	.137* (.015)	.056* (.022)
ROOMS$^{(L)}$	.394* (.024)	.195 (.287)	.449* (.060)	.240 (.333)	.410* (.075)	.251 (.467)
EXTAMENITY	.165* (.036)	.041 (.046)	.207* (.049)	.041 (.107)	.199* (.051)	.046 (.039)
BEACH	1.539* (.028)	.561* (.099)	1.924* (.033)	.492* (.191)	1.227* (.077)	.388* (.104)
LAKE	-.663* (.116)	1.560* (.389)	-.745* (.081)	.974* (.267)	-.712* (.082)	1.568* (.235)
TRANS	1.336* (.140)	.192* (.064)	1.359* (.116)	.198 (.170)	1.503* (.182)	.193 (.216)
HIGHWAY	.447* (.093)	.068 (.061)	.464* (.080)	.057 (.109)	.374* (.092)	.053* (.011)
DOWNTOWN	1.198* (.061)	.471* (.093)	1.051* (.088)	.283 (.076)	.943* (.053)	.331 (.078)
CRIME	-.173* (.043)	.015 (.034)	-.189* (.041)	.036 (.067)	-.178* (.032)	.018 (.017)
RATING	2.661* (.015)	1.308* (.091)	2.361* (.017)	.926* (.083)	2.017* (.020)	1.334* (.092)
REVIEWCNT$^{(L)}$	1.230* (.107)	.369* (.069)	1.102* (.128)	.405* (.057)	1.182* (.158)	.438* (.059)
STAFF	.139* (.027)	.034 (.088)	.142* (.021)	.031 (.081)	.147* (.022)	.033 (.095)
FOOD	.225* (.038)	.136* (.002)	.234* (.043)	.141* (.009)	.251* (.039)	.146* (.005)
BATHROOM	.290 (.271)	.060 (.103)	.278 (.259)	.082 (.122)	.242 (.277)	.091 (.118)
PARKING	.097* (.008)	.075* (.011)	.092* (.005)	.071* (.009)	.088* (.008)	.079* (.013)
BEDROOM	-.175 (.232)	.253 (.269)	-.164 (.241)	.277 (.256)	-.189 (.237)	.270 (.244)
FRONTDESK	.065 (.103)	.021 (.076)	.077 (.112)	.016 (.068)	.073 (.125)	.028 (.071)
BRAND	Yes					
(Search Cost)	$\overline{\gamma}$	Σ_{γ}	$\overline{\gamma}$	Σ_{γ}	$\overline{\gamma}$	Σ_{γ}
Search Base Cost	-7.511*(.089)	.971* (.176)	----	----	-4.041*(.092)	.932* (.241)
COMPLEXITY	.541* (.094)	.398* (.115)	----	----	.219 (.104)	.525 (.781)
SYLLABLES$^{(L)}$	.678* (.115)	.721* (.106)	----	----	.582* (.165)	.633 (.958)
SPELLERR$^{(L)}$	.329* (.082)	.033 (.101)	----	----	.192 (.226)	.053 (.283)
SUB	.196* (.045)	.057 (.229)	----	----	.141 (.123)	.070* (.011)
SUBDEV	.342* (.056)	.119 (.273)	----	----	.284* (.084)	.169* (.030)
Maximum LL	-405,418		-114,003		-352,359	
Price Elasticity	-1.619		-2.973		-2.183	

(L) Logarithm of the variable. * Statistically significant at 5% level.

M: Main Model.

A1: Alternative Model I (Use Purchase Data Only – Mixed Logit Model, with Actual Limited Consideration Set).

A3: Alternative Model III (Use Click Data Only – Click Model).

Table G2. Estimation Results – Alternative Models (II) & (IV)

Variable	Mean Effect (Std. Err)^{A2}	Heterogeneity (Std. Err)^{A2}	Mean Effect (Std. Err)^{A4}	Heterogeneity (Std. Err)^{A4}
(Preferences)	$\overline{\alpha}, \overline{\beta}, \overline{\lambda}$	$\sigma_{\alpha}, \Sigma_{\beta}, \Sigma_{\lambda}$	$\overline{\alpha}, \overline{\beta}, \overline{\lambda}$	$\sigma_{\alpha}, \Sigma_{\beta}, \Sigma_{\lambda}$
PRICE$^{(L)}$	-2.036* (.034)	.989* (.086)	-1.663* (.027)	.421* (.052)
PAGE	-.240* (.004)	.124 (.248)	-.190* (.009)	.080 (.163)
RANK	-.309* (.004)	.130 (.186)	-.245* (.006)	.177* (.055)
CLASS	1.767* (.023)	1.020* (.244)	1.606* (.025)	.824* (.155)
AMENITYCNT$^{(L)}$	.201* (.045)	.050 (.105)	.144* (.038)	.059 (.075)
ROOMS$^{(L)}$	.425* (.037)	.211 (.309)	.320* (.035)	.159 (.292)
EXTAMENITY$^{L)}$	.193* (.051)	.043 (.128)	.174* (.043)	.036 (.049)
BEACH	1.958* (.023)	.491* (.199)	1.868* (.012)	.491* (.088)
LAKE	-.886* (.076)	1.026* (.202)	-.797* (.109)	1.318* (.361)
TRANS	1.161* (.089)	.186 (.195)	1.021* (.183)	.172* (.072)
HIGHWAY	.412* (.072)	.056 (.110)	.335* (.099)	.030 (.086)
DOWNTOWN	.991* (.083)	.263 (.083)	.918* (.055)	.389* (.082)
CRIME	-.159* (.041)	.034 (.061)	-.166* (.023)	.014 (.041)
RATING	2.215* (.011)	.883* (.098)	2.794* (.019)	.972* (.097)
REVIEWCNT$^{(L)}$	1.077* (.114)	.413* (.044)	1.208* (.192)	.408* (.079)
STAFF	.146* (.022)	.033 (.081)	.132* (.025)	.033 (.082)
FOOD	.241* (.045)	.145* (.007)	.230* (.040)	.133* (.005)
BATHROOM	.286 (.267)	.084 (.121)	.292 (.266)	.062 (.104)
PARKING	.091* (.004)	.074* (.007)	.090* (.003)	.079* (.012)
BEDROOM	-.160 (.240)	.282 (.273)	-.171 (.227)	.254 (.260)
FRONTDESK	.079 (.111)	.017 (.067)	.066 (.098)	.022 (.075)
COMPLEXITY	-.130* (.022)	.063 (.331)	-.149* (.025)	.093 (.443)
SYLLABLES$^{(L)}$	-.175* (.041)	.271* (.048)	-.171* (.037)	.246* (.033)
SPELLERR$^{(L)}$	-.083* (.003)	.019 (.032)	-.081* (.004)	.034 (.054)
SUB	-.109* (.008)	.038* (.006)	-.136* (.011)	.040* (.010)
SUBDEV	-.139* (.019)	.073 (.054)	-.165* (.012)	.067* (.033)
BRAND	Yes			
Maximum LL	-119,752		-388,106	
Price Elasticity	-2.393		1.954	

(L) Logarithm of the variable. * Statistically significant at 5% level.

A2: Alternative Model II (Mixed Logit Model, with Actual Limited Consideration Set + Additional Search Cost Variables).

A4: Alternative Model IV (Joint Probabilistic Model of Click and Purchase, Using Both Click and Purchase Data + Additional Search Cost Variables, But No Click Sequence Information).

Online Appendix H.
Breakdown Analysis for Predicted Search Engine Revenues

As shown in Table 4 from the policy experiments, under various different scenarios product search engines will experience a revenue increase when providing different sets of information on the search summary page. To examine where the revenue increase comes from (i.e., existing consumers or better market coverage), we conducted an additional analysis on the breakdown of the revenue in the simulation.

More specifically, the predicted total search engine revenues can be computed as follows:

$$\text{Predicted Total Revenues} = \sum_{All\ Sessions} \sum_{All\ Hotels} (\text{Price} \ * \ \text{Predicted Purchase Probability} * \text{Commission Rate } 20\%). \tag{H1}$$

Based on Equation (G1), we were able to separately compute the predicted revenues from all hotels for each session *i* as follows:

$$\sum_{All\ Hotels\ in\ Session\ i} (\text{Price} \ * \ \text{Predicted Purchase Probability} * \text{Commission Rate } 20\%). \tag{H2}$$

Then, we categorized all sessions from our observed data into two types: 1) sessions without any purchase, and 2) sessions with purchase. We separately computed the total predicted revenues from each of the two categories, and then compare the difference in the predicted revenues and the observed revenues.

We found that the revenue increase (i.e., increase under all scenarios except for "Existing - Price") came from both types of sessions. In particular, for sessions without any purchase, the predicted revenue increase indicates a potential increase in market coverage (i.e., consumers who did not purchase in the past become likely to purchase). This is consistent with our model intuition 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).

For sessions with purchase, the predicted revenue increase indicates a potential increase in spending from the existing consumers (consumers who were less likely to purchase in the past become more likely to purchase; or consumers who were likely to spend less in the past become likely to spend more). We provide more details on the breakdown of the revenue increase in Table H1 below for your convenience.

Table H1. Breakdown Analysis Results on Search Engine Revenue Increase

	Overall Search Engine Revenue	Revenue Increase		
		All Sessions	Sessions With Purchase	Sessions Without Purchase
Existing	$452,781	$0	$0	$0
Existing + Location Information	$553,136	$100,355	$91,323	$9,032
Existing + Service Information	$467,369	$14,588	$11,962	$2,626
Existing – Price Information	$420,132	*$-32,649*	*$-28,731*	*$-3,918*
Existing + Review Information (Text Features)	$507,160	$54,379	$47,853	$6,526
Existing + Review Information (Topic Entropy)	$490,063	$37,282	$30,572	$6,711

In addition, we also found that the revenue increase occurs for both hotels that have been purchased in the past (i.e., existing hotels) and hotels that have not been purchased in the past (i.e., new hotels). This finding provides additional 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.