Search with Refinement

Yuxin Chen Song Yao [1]

June 1, 2012

[1]Yuxin Chen is the Polk Brothers Professor in Retailing and Professor of Marketing at the Kellogg School of Management, Northwestern University (email: yuxin-chen@kellogg.northwestern.edu). Song Yao is an Assistant Professor of Marketing at the Kellogg School of Management, Northwestern University (email: s-yao@kellogg.northwestern.edu). The authors would like to thank seminar participants at the Ohio State University, University of Chicago, and Washington University at St. Louis as well as Günter Hitsch, Dmitri Kuksov, and Ting Zhu for their feedback.. The authors thank an anonymous travel website and Wharton Customer Analytics Initiative for providing the data.

Abstract: Search with Refinement

Preliminary. Please do not cite or distribute

The development of online search technology has profound impacts on consumers, business, and the society. One important feature of online search is that consumers are able to refine the search results using tools such as sorting and filtering. Albeit such refinement tools have significant effects on consumer behavior and market structure, there is little empirical research documenting and measuring the effects. We propose a structural model of optimal consumer search that coherently integrates the decisions of consumer search and refinement as well as the structure of uncertainty about the products. The model is estimated using a unique data set of individual level search activities of hotels provided by a large travel website. We find that the refinement tools encourage 57% more searches and enhance the utility of the purchased product by nearly 20%. However, most websites by default rank search results according to the results' qualities or relevance to consumers (e.g., Google). We find that if consumers are uninformed about such default rules for ranking, they may engage in disproportionally more searches using the refinement tools. As a consequence, the overall welfare surplus may deteriorate by up to 8%. In contrast, when consumers are informed that the default ranking of the results already reflects the qualities, they search less and the welfare surplus increases 3.3%. We also find the refinement leads to a less concentrated market structure.

Keywords: consumer search, information asymmetry, market structure, electronic commerce, consumer behavior

1 Introduction

The advance of online search technology in the Internet era has made profound impacts on consumers, business, and society. According to a recent report by McKinsey, the global value of search technology is $780 billion annually with $540 billion direct contribution to global GDP.[1] As an important feature of modern search technology, it allows users to search products or services with multiple attributes using tools such as sorting and filtering to refine the search results. For example, an individual searches a hotel on a travel website may sort the search results by price in an ascending order and in the meanwhile filter out the hotels with star ratings below three. As another example, an academic researcher may conduct a keyword search for journal articles in an online library, filtering out all the non-peer-reviewed ones and having the results sorted by publication dates. Given the ubiquitous use of such refinement tools in search, it is surprising that there is little empirical studies on the refinement tools' value to consumers and impact on consumer search behavior and the market structure. The objective of this paper is to fill this gap by empirically examining the effect of refinement tools such as sorting and filtering on consumer search and quantifying their value to consumers and impact on market structure. We start out by building a structural model of consumer sequential search with the availability of sorting and filtering on multiple product/service attributes. Specifically, in our model, consumers engage in costly search to resolve the uncertainty about the attributes of products/services. They may apply the refinement tools to alter search cost and the distribution of attributes. The model parameters are then calibrated using a unique data set of individual level search activities and choices of hotels provided by a major travel website. Our empirical findings and the subsequent counterfactual analysis suggest that, with the aid of refinement tools, search cost drops and consumers search increases by about 57%. Furthermore, a consumer may achieve about 20% higher of utility for the product she chooses. On the other hand, while most websites by default rank search results according to the results' qualities or relevance to consumers (e.g.,

[1] "The impact of Internet technologies: Search", July 2011, Mckinsey.

Google), it is crucial for the websites to educate consumers about such default ranking rules. We find that consumers engage in disproportionally excessive searches using the refinement tools when they are uninformed about such default ranking rules. Consequently, consumers' overall welfare surplus is lower than the level if the refinement tools were disabled. The welfare loss due to the excessive searches can be up to 8%. In contrast, when the consumers understand that the search results by default are ranked according to the qualities, they search less and the welfare surplus becomes higher than the level without the refinement tools. We also find that the market becomes less concentrated owing to the existence of refinement tools because: (1) the lowered search cost allows consumers to search with a greater depth and find more hotels, and (2) heterogeneous consumers are able to locate hotels that match their preference better. Such better matches would be too costly to achieve without the refinement tools.

In addition to the aforementioned findings, this paper also advances the growing structural empirical modeling literature on measuring search costs and understanding consumer search behavior. To our best knowledge, this is the first empirical research using a structural modeling approach to investigate consumer sequential search for products/services with multiple unknown attributes given the availability of refinement tools. Hong and Shum (2006) and Hortacsu and Syverson (2004) develop structural approaches to estimate the distribution of consumer search costs. Their approaches utilize the parameter restrictions implied by the equilibrium price distribution derived from the supply side. Koulayev (2010b) allows the possibility that consumers do not know the price distribution during the search and learn the distribution during the search. Santos et al. (2011) compare alternative classical search models using web browsing and purchase data. They focus on consumer search among homogenous products with price uncertainty. In the context of choosing among heterogeneous products, Mehta et al. (2003), Kim et al. (2010), Koulayev (2010a), Seiler (2011) and Honka (2011) propose structural models for the formation of consideration sets as the result of consumer search and model consumer purchase conditional on the consideration sets. While

sequential search assumption is adopted in Kim et al. (2010) and Koulayev (2010a), simultaneous search assumption is considered in Mehta et al. (2003), Seiler (2011) and Honka (2011). A distinguishing feature of Honka (2011) is that the actual consideration sets of consumers are observed in the data. The actual search process and search behavior, however, are not observed in the data used in nearly all those studies mentioned above.[2] In contrast, we consider the sequential search behavior (including the usage of search refinement tools) of consumers as well as their purchase behavior in the context of choosing among heterogeneous services with multiple attributes (e.g., hotels). This enables us to build a structural model in which consumer decisions on search, use of search refinement tools, and purchase are derived from a unified framework of utility maximization. It also makes it feasible to evaluate the impact of search refinements.

Our paper is also related to Yao and Mela (2011) which explicitly models consumer decisions of using sorting and/or filtering functions in online search. Their model is constructed from the perspective of sellers. To be consistent with the information structure of the sellers, the model aggregates individual consumer choices up to the market level. In contrast, our model addresses the search at the individual consumer level, enabling us to address some subtle issues such as how the refinement affects the number of searches and how previous search results affect then-current consumer decisions.

The rest of the paper is organized as follows. First, we detail the structural model of consumer optimal search with the ability of refining search results. We then describe the data used for the estimation. We discuss the estimation strategy in section 4. Next, we present the results, model fit information, and some robustness tests. In section 6, we consider a few counterfactual simulations, exploring the impact of the refinement tools on consumer search behavior, welfare, and market structure. We conclude with a discussion of main findings and some future research directions.

[2]Santos et al. (2011) is a notable exception, which use browsing data from comScore to infer consumer searching behavior. One limitation of comScore's data is that the price information is missing unless there was a purchase. As a result, most of the price information during a consumer's search has to be imputed from other sources.

2 A Model of Search with Refinement

The travel website who provides us with the data hosts the information of a comprehensive list of hotels for a given city. When visiting this website to find a hotel, the consumer needs to first specify the criteria of hotels she is looking for, such as the location, checkin and checkout dates, etc.. The website shows the consumer a list of hotels that satisfy the criteria. The list of hotels is sorted according to some default algorithm by the website. Based on our communication with the website's management team, the default algorithm is based on the numbers of bookings of hotels at the website for a given period (i.e., the frequency of purchases). However, this default ranking rule was undisclosed to the consumers during the period of observation. The default ranking was vaguely named as "[Website] Picks", making no implication about the default ranking rule being used.[3] Even when consumers made inquires to the customer service regarding this default ranking, they would not get the straight answer. Instead, they would be told some obscure explanations such as it is "the summary from the most affordable price, highest guest rating, highest star rating, and the hotel nearest to the airport, to the expensive price, lowest guest rating, lowest star rating, and the hotel farthest to the airport."

The consumer can choose to refine the default search results using alternative sorting and/or filtering rules (e.g., sort by prices, filter by star ratings, etc.). After the refinement, if two hotels have the same level of the attribute used for the refinement (e.g., when being sorted by star ratings, two hotels both have four-star rating.), the two hotels will be ranked according to the default ranking algorithm.

The list of hotels can potentially be very long. The website only displays up to twenty-five hotels per webpage. The consumer can then choose to explore the next 25 hotels on the list by turning to the next webpage. On a given webpage, the consumer can directly read some hotel attributes of the 25 listed hotels, including the star ratings, consumer ratings, and

[3]We use the word [Website] to disguise the website's identity.

average prices. For other attributes such as the amenities of a hotel, however, the consumer has to search further by clicking through the hotel's link.

2.1 Utility

Consumer i's utility of booking hotel j is characterized as

$$u_{ij} = \mu_i(z_{ij}, x_j) + \nu_{ij}$$

with

$$\nu_{ij} \sim N(0, 1)$$

where $j = 1, 2, ..., N_i$, with N_i being the total number of hotels that satisfy i's criteria. z_{ij} is a row vector of the attributes that can be directly read without click-through. z_{ij} may vary across individuals since it includes the average price of the hotel, which depends on the individual's search criteria. x_j is a row vector of hotel attributes (mainly amenities) that have to be discovered with click-through on the hotel's link. ν_{ij} is a consumer-hotel specific error term. In particular, ν_{ij} is drawn from standard normal distribution and i.i.d. across consumers and hotels.[4] z_{ij} and x_j are drawn from some joint distribution of hotel attributes. Without losing generality, we normalize the mean utility of the outside option to 0 such that

$$\begin{aligned} u_{i0} &= \mu_{i0} + v_{i0} \\ &= v_{i0} \\ \mu_{i0} &= 0 \end{aligned}$$

[4]We normalize the standard deviation to 1 for identification purpose. As in most discrete choice models, the standard deviation of the error term is usually unidentified (e.g. Train (2003)).

2.2 Refinement and Search Cost

Following the classical economic literature (e.g., Nelson (1970)), we define consumer search as gathering information to resolve the uncertainty about her utility regarding a specific product. There are multiple sources of uncertainty about the utility. Before turning into the webpage containing hotel j, consumer i has uncertainty about z_{ij}; in addition, before clicking through the link of hotel j, consumer i has uncertainty about ν_{ij} and x_j. She can resolve ν_{ij}, x_j and z_{ij} by clicking through the link and/or turning the webpage. Correspondingly, a theoretically consistent definition of a search is that the consumer views a list of hotels (resolving z_{ij}) and clicks on a hotel link (resolving x_j and ν_{ij}).

The consumer incurs some search cost during this process. The search cost can be interpreted as the time and efforts spent on the investigation. Since the slot position of a hotel on the list may affect the accessibility of that hotel, it may affect the search cost (Yao and Mela (2011); Kim et al. (2010); Ansari and Mela (2003)).

At the focal website, consumers have the option of refining the search results using sorting and filtering. Refinement will affect the distribution of x_j, z_{ij} as well as the search cost. In particular,

- **The effect of sorting on the distribution of $\{z_{ij}, x_j\}$.** The default order of hotel list (the undisclosed algorithm) has no obvious ordering on hotels' attributes. The public do not know that the default ranking is based on the booking frequencies of hotels. As a result, from the perspective of the consumer it can be considered that $\{z_{ij}, x_j\}$ are randomly drawn from the attributes distribution.[5] In contrast, if the consumer sorts the hotel based on some attribute such as price, the sorted attribute becomes an ordered statistics. For example, if the hotels are sorted by price ascendingly and hotel j is at slot 2 on the list, then the consumer knows that hotel j has the second lowest price. Since other attributes are likely to be correlated with price, the sorting

[5] In Section 5, we implement a robustness check for this assumption and confirm its validity.

will also have an impact on the distributions of those attributes conditioned on the slot position.

- **The effect of sorting on search cost.** Sorting rearranges the order of the list. The slot position may affect the search cost of a given hotel. For example, suppose highly priced hotel j is at slot 1 by default. There is some search cost for reaching and clicking through the link to resolve the uncertainty about the hotel. If consumer i sorts the list by price ascendingly, hotel j is moved to slot $s > 1$. Since to reach slot s the consumer needs to move down the list and turn the pages, the time and efforts involved are greater than those for slot 1. Consequently, the search cost increases.

- **The effect of filtering on the distribution of** $\{z_{ij}, x_j\}$. Filtering on a specific attribute eliminates those hotels that do not meet the criterion. As a result, the filtering changes the distribution of the attributes of the listed hotels. For example, if the consumer uses the filter to show five-star hotels only, then the distribution of star ratings is truncated below five-star. Since star ratings and other attributes (e.g., price) are correlated, such a filtering also affects the distribution of other attributes.

- **The effect of filtering on search cost.** Filtering makes the list of hotels become shorter and hence affect search costs. For example, following the same logic of how sorting affects search cost, if hotel j's slot moves up on the list because some hotels previously above it are filtered out, the search cost for hotel j will drop.

To accommodate these effects of refinement on search costs and attributes distribution, we introduce the notations of sorting/filtering-specific expected search cost and attributes distribution. To be specific, suppose there are K refinement methods, including the no sorting or filtering option. Also denote S_{ij} as the set of hotels which the consumer has searched before searching hotel j, including the outside option. For a given sorting/filtering method k ($k = 1, 2, ..., K$), the corresponding expected search cost is c_{ij}^k. The search cost is

a function of the slot position:

$$c_{ij}^k = \int_{z,x} c_i(\text{Slot}_j^k) dP^k(z_{ij}, x_j | S_{ij})$$

where Slot_j^k is hotel j's slot position when the consumer uses sorting/filtering method k. $P^k(z_{ij}, x_j|S_{ij})$ is the distribution of hotel j's attributes under refinement method k.[6] Since the slot position depends on the levels of hotel attributes and the consumer has uncertainty about the attributes, the integral has the uncertainty integrated out.

Note that the distribution of the attributes depends on S_{ij} and k. For example, if the consumer is sorting hotels descendingly using price and has searched hotel j' with a price tag of \$150/night (i.e., $j' \in S_{ij}$). Hotel j has a position lower than j'. Since any hotels listed below hotel j' have prices lower than \$150/night, the price distribution of hotel j becomes truncated above \$150.

With these specifications, we now formalize what the customer does and does not know *before* the search on hotel j.

- The consumer gets the information of the outside product without incurring any search cost.
- The consumer knows the total number of available hotels N_i.
- The consumer knows the distribution of ν_{ij}.
- The consumer knows the conditional distribution of $P^k(z_{ij}, x_j|S_{ij})$.
- The consumer knows the expected search cost for hotel j before the search.
- On the webpage containing hotel j, before clicking through the link of j, the consumer directly observes z_{ij} from the webpage but does not know the values of ν_{ij} and x_j.

[6]Note that when the consumer is at the webpage containing hotel j, the values for z_{ij} is deterministic as the consumer can directly read the information off the page. For the ease of exposition, however, we will use the joint distribution notation throughout the paper.

2.3 Expected Marginal Gain of Searching an Additional Hotel

At a given point of the search process, define the maximum utility among those already searched hotels (including the outside option) as u_i^*.[7] Also denote the CDF of the conditional distribution of u_{ij} under refinement k as $F^k(u_{ij})$. The distribution of u_{ij} depends on (1) the method k used for refinement (2) the distribution of $P^k(z_{ij}, x_j|S_{ij})$ (3) the distribution of ν_{ij}. The expected marginal net gain from searching hotel j, using refinement method k, and then stopping the search and choosing any hotel with the highest utility is given by (Weitzman (1979)):

$$
\begin{aligned}
Q_{ij}^k =& \underbrace{\rho_i \int_{-\infty}^{u_i^*} u_i^* dF^k(u_{ij})}_{u_{ij}<u_i^*} + \underbrace{\rho_i \int_{u_i^*}^{\infty} u_{ij} dF^k(u_{ij})}_{u_{ij}\geqslant u_i^*} - c_{ij}^k - u_i^* \qquad (1)\\
=& \rho_i \int_{-\infty}^{u_i^*} u_i^* dF^k(u_{ij}) + \rho_i \int_{u_i^*}^{\infty} u_i^* dF^k(u_{ij})\\
&- \rho_i \int_{u_i^*}^{\infty} u_i^* dF^k(u_{ij}) + \rho_i \int_{u_i^*}^{\infty} u_{ij} dF^k(u_{ij}) - c_{ij}^k - u_i^*\\
=& \rho_i \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) dF^k(u_{ij}) - (1-\rho_i) u_i^* - c_{ij}^k \qquad (2)
\end{aligned}
$$

where ρ_i is the discount factor since the consumer has to spend more time for an additional search. The first term in equation 1 stands for the expected utility the consumer may get if $u_{ij} < u_i^*$; the second term stands for the expected utility if $u_{ij} \geqslant u_i^*$; the "$-c_{ij}^k$" term is because the consumer needs to pay the search cost for searching hotel j; the "$-u_i^*$" is because the consumer may stop now and get u_i^* without waiting and paying the search cost.

Since the time interval for the search is short, we set the discount factor $\rho_i = 1$. The expected marginal gain can then be simplified to

$$
Q_{ij}^k = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) dF^k(u_{ij}) - c_{ij}^k \qquad (3)
$$

[7] Note that u_i^* is not fixed during the search process and may change after each search.

2.4 Optimal Sequential Search Strategy with Refinement

The sequential search problem in our setting can be described as a consumer faces multiple search opportunities (Weitzman (1979)). Each search opportunity is the combination of a hotel that has not been searched and a specific refinement method (e.g., hotel j and sorting/filtering method k). Consumer gets the information of the outside option for free. Then by paying the search cost (c_{ij}^k), the consumer can search hotel j to discover the exact utility of booking that hotel (u_{ij}).

Before characterizing the optimal search strategy, we first define the consumer's reservation utility R_{ij}^k, which is the utility level that makes the consumer indifferent between (1) choosing an already-searched hotel with the utility of R_{ij}^k, and (2) search hotel j using refinement method k. That is, R_{ij}^k solves the implicit function

$$Q_{ij}^k = \int_{R_{ij}^k}^{\infty} (u_{ij} - R_{ij}^k) dF^k(u_{ij}) - c_{ij}^k = 0$$

or

$$c_{ij}^k = \int_{R_{ij}^k}^{\infty} (u_{ij} - R_{ij}^k) dF^k(u_{ij}) \tag{4}$$

In the Appendix we show that Q_{ij}^k is monotonically decreasing in R_{ij}^k and a unique solution of the reservation utility exists.

The optimal search strategy contains the following two steps, a stopping rule to determine when to stop searching and a selection rule for choosing which hotel to search (Weitzman (1979)):

Step 1: Stopping Rule (when to stop searching): Calculate the reservation utility for each alternative search opportunity (the combination of a hotel and a refinement method). If there is not a reservation utility that is higher than the then-current maximum utility u_i^*, stop the search and choose that hotel with the highest utility u_i^*. Otherwise, proceed to the next step.

Step 2: Selection Rule (how to search): Search the alternative with the highest reservation utility and go back to Step 1.

This optimal strategy can be interpreted as the following: The consumer will continue searching if the expected marginal gain of doing so is greater than 0. In particular, she will choose to search hotel j using refinement method k if such a search has the highest reservation utility.[8] If the consumer decides to stop searching, then she will book the hotel with the highest utility among those already searched, including the outside option.

3 Data

We analyze our model using data provided by a travel website, one of the major online travel products providers in the United States. Note that we consider the analysis of this data set a particular demonstration of a more general model. The model we developed can be easily modified and applied to other consumer search contexts.

The data set records consumer hotel searching and booking activities for the city of Cancun, Mexico and surrounding area. We estimate the model on a set of 215 randomly selected consumers who conducted the searches between October 1 and October 15, 2009 and were interested in the checkin dates between November 1 and November 15, 2009. We later use a holdout sample of the consumers whose intended checkin dates were between October 25 and October 31, 2009, to validate the model. Table 1 presents summary statistics of consumer searching and booking activities. On average a consumer makes 2.86 searches. However, there is a large variance of the number of searches, which implies that there may be substantial heterogeneity in search cost. The booking activity is rather sparse. Among the 215 consumers, only 3 consumers booked a hotel in the end. Per the discussion with the managers of the website, this is consistent with the industry average.

[8]The stopping rule implies that when the search continues, $R^k_{ij} > u^*_i$. Since $R^k_{ij} > u^*_i$, the integral term with u^*_i as the lower bound will be greater than the one with R^k_{ij} as the lower bound. Because $Q^k_{ij}(S_{ij}, R^k_{ij}) = 0$ by definition, the marginal gain $Q^k_{ij}(S_{ij}, u^*_i) > 0$.

Consumers are also very diversified in their refinement activities. The number of refinement among consumers range from 0 to 9. The diversity of refinement methods used indicates the consumers may be heterogeneous in their preferences regarding hotel attributes. We also find that in the data, the top seven methods used for the refinement account for over 95% of all methods used, including: (1) sort by price ascendingly, (2) sort by consumer rating descendingly, (3) sort by stars descendingly, (4) filter out hotels below four-star, (5) sort four- and five-star hotels by price ascendingly, (6) sort four- and five-star hotels by consumer rating descendingly, (7) sort four- and five-star hotels by stars descendingly.

Table 1: Summary Statistics of Consumers Activities

	Mean	Std. Dev.	Min.	Max.
Number of Searches per Consumer	2.86	3.09	1	16
Number of Hotel Booked per Consumer	0.01	0.12	0	1
Number of Refinement per Consumer	0.52	1.20	0	9

On the supply side, in total there are 91 hotels in the data. Table 2 reports the summary statistics of hotel attributes. The hotels on average have a daily price of $121.84, with a large variation across hotels. Star ratings and consumer ratings of hotels have similar mean levels but the latter has a larger variance. Among the hotel attributes reported, "Free Internet", "Air Condition", and "Swimming Pool" are not directly listed on the result page, which are the most popular amenities according to the website. Consumers have to engage in further search to discover the information of these amenities by clicking through the hotel link.

Table 2: Summary Statistics of the Hotels

	Mean	Std. Dev.	Min.	Max.
Average Daily Price ($)	121.84	101.40	0	438.00
Star Rating	3.39	0.82	0	5.00
Consumer Rating	3.32	1.72	0	5.00
Promotional Event	0.77	0.42	0	1.00
Price Information Missing	0.05	0.23	0	1.00
Free Internet Access	0.37	0.49	0	1.00
Air Condition in Room	0.97	0.18	0	1.00
Swimming Pool	0.96	0.21	0	1.00

4 Estimation

4.1 Utility and Sorting/Filtering

We estimate the model using maximum likelihood with random coefficients. Utility is defined as

$$\begin{aligned} u_{ij} &= \mu_i(z_{ij}, x_j) + v_{ij} \\ &= z_{ij}\alpha_i + x_j\beta_i + v_{ij} \end{aligned}$$

where z_{ij} is the vector of hotel attributes that consumers can directly observe when the consumer is at that webpage, which includes a hotel's chain brand, average daily price, star rating, average customer rating, whether the price is a promotional price and whether the price information is missing. α_i is the consumer's sensitivity with respect to z_{ij}.

x_j is the vector of hotel attributes that consumers cannot directly observe from the webpage without click-through, including hotel amenities such as free internet, swimming pool and air conditioning. The exact values of these attributes can only be discovered with click-through on the hotel link. β_i is the consumer's sensitivity regarding x_j.

The random error term v_{ij} follows a standard normal distribution that is i.i.d. across individuals and hotels.[9]

Finally, we consider the top seven refinement methods, accounting for more than 95% of search with sorting/filtering. The seven methods are: (1) sort by price ascendingly, (2) sort by consumer rating descendingly, (3) sort by stars descendingly, (4) filter out hotels below four-star, (5) sort four- and five-star hotels by price ascendingly, (6) sort four- and five-star hotels by consumer rating descendingly, (7) sort four- and five-star hotels by stars descendingly.

4.2 Heterogeneity

To capture the heterogeneity of consumer preference, we consider a random coefficient structure for (α_i', β_i'). To be specific,

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Omega_u\right) \tag{5}$$

where (α', β') are the mean levels of the utility coefficients; Ω_u is a diagonal matrix with the attribute-specific variances on the diagonal, which capture the degree of heterogeneity of consumers.

[9]One concern for the assumption of independent error terms is the potential endogeneity bias caused by the possible correlation between the error and the variables. We introduce flexible fixed effects at the hotel chain level with a random coefficient structure to control the heterogeneity, which to some extent may extenuate the problem. A similar treatment for the potential endogeneity biases is considered in Kim et al. (2010). We are aware that this is an important issue. A complete solution may involve modeling the strategic behavior of the hotels and the website, which is beyond the scope of this paper.

Another level of consumer heterogeneity comes from consumer's search cost. The search cost is specified as[10]

$$\begin{aligned} c_{ij}^k &= \int_{z,x} c_i(\text{Slot}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \\ &= \int_{z,x} \exp(\gamma_{i0} + \gamma_{i1} \text{Slot}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \end{aligned}$$

We assume that $(\gamma_{i0}, \log\gamma_{i1})$ follow normal distribution such that

$$\begin{pmatrix} \gamma_{i0} \\ \log\gamma_{i1} \end{pmatrix} \sim N \left(\begin{pmatrix} \gamma_0 \\ \log\gamma_1 \end{pmatrix}, \Omega_c \right) \tag{6}$$

where $(\gamma_0, \log\gamma_1)$ are the mean levels of the coefficients; Ω_c is a diagonal matrix with the variances of $(\gamma_0, \log\gamma_1)$ on the diagonal, capturing the deviation from the means. The log-normal specification of γ_{i1} ensures that the cost is non-decreasing in slot position (Ansari and Mela (2003); Yao and Mela (2011)).

4.3 Calculation of the Reservation Utilities

In Appendix B we show that the following equation holds:

$$c_{ij}^k = \{(1 - \Phi(R_{ij}^k - \mu_{ij}^k))(\mu_{ij}^k - R_{ij}^k + \frac{\phi(R_{ij}^k - \mu_{ij}^k)}{(1 - \Phi(R_{ij}^k - \mu_{ij}^k))})\} \tag{7}$$

As a result, it becomes clear that, for a given pair of $\{c_{ij}^k, \mu_{ij}^k\}$, we can calculate the corresponding reservation utility R_{ij}^k by solving equation 7. The computation can be simplified by constructing a look-up table of the triple $\{c, \mu, R\}$, with the grid up to a substantial fine level.[11] Since the table holds for all (i, j, k), we drop the subscripts (i, j) and superscript k. Note that this grid does not depend on the parameter values. We can first create this table

[10] We also consider two alternative search cost specifications in section 5.3 and confirm the current specification is appropriate.

[11] A similar approach is used in Kim et al. (2010), without the consideration of attributes uncertainty and refinement methods.

outside of the estimation step. Then during the estimation for each given pair of $\{c_{ij}^k, \mu_{ij}\}$ we can check the table for the corresponding value of R_{ij}^k, potentially with a nonparametric interpolation step if the table does not contain the exact pair of $\{c_{ij}^k, \mu_{ij}\}$.[12] The uncertainty of $\{z_{ij}, x_j\}$ and hence μ_{ij} and c_{ij}^k is addressed by repeatedly making draws from their distribution and calculating the corresponding expectation.

4.4 Likelihood

Based on the optimal search strategy described in Section 2.4 and the data, we can obtain the following likelihood on search behavior. First, denote S_{ij} as the set of already searched hotels right before hotel j being searched. Also denote S_i as the set containing all of the hotels the consumer searched before she stops the search.

1. When the consumer search hotel j's link using a specific refinement method k, the reservation utility of that search opportunity (the combination of the hotel and the refinement method) is greater than the utilities of all searched hotels, including the outside option. The corresponding probability is

$$\begin{aligned} & \Pr(R_{ij}^k \geq u_{ir}, \forall r \in S_{ij}) \\ = & \prod_{\forall r \in S_{ij}} \Pr(R_{ij}^k \geq u_{ir}) \\ = & \prod_{\forall r \in S_{ij}} \int_{\alpha,\beta,\gamma_0,\gamma_1} \Phi(R_{ij}^k - (z_{ir}\alpha_i + x_r\beta_i)) dP(\alpha, \beta, \gamma_0, \gamma_1) \end{aligned} \tag{8}$$

   where $P(\alpha, \beta, \gamma_0, \gamma_1)$ is the heterogenous distributions of parameters from equation 5 and 6. The terms inside the integral is due to the assumptions that $\nu_{ir} \sim N(0, 1)$.

2. If the consumer stops the search and books a hotel (including the outside option, not booking any hotels), then that hotel's utility is greater than the utilities of all searched hotels and the reservation utilities of all unexplored search opportunities.

[12] We use a spline regression in our implementation.

The corresponding probability of this condition is

$$\begin{aligned}
&\Pr(u_i^* \geq u_{ij}, \forall j \in S_i)\Pr(u_i^* \geq R_{ij'}^k, \forall k, \forall j' \notin S_i) \qquad (9)\\
&= \prod_{\forall j \in S_i} \Pr(u_i^* \geq u_{ij}) \prod_{\forall k, \forall j' \notin S_i} \Pr(u_i^* \geq R_{ij'}^k)\\
&= \prod_{\forall j \in S_i} \int_{\alpha,\beta,\gamma_0,\gamma_1} \Phi\left(\frac{(z^*\alpha_i + x^*\beta_i) - (z_j\alpha_i + x_j\beta_i)}{\sqrt{2}}\right) dP(\alpha, \beta, \gamma_0, \gamma_1)\\
&\quad \prod_{\forall k, \forall j' \notin S_i} \int_{\alpha,\beta,\gamma_0,\gamma_1} \Phi((z^*\alpha_i + x^*\beta_i) - R_{ij'}^k) dP(\alpha, \beta, \gamma_0, \gamma_1)
\end{aligned}$$

where "*" is to index the hotel being booked.

Correspondingly, the likelihood of the model can be written as

$$\begin{aligned}
L_i &= \prod_{\forall j \in S_i} \prod_{\forall r \in S_{ij}} \Pr(R_{ij}^k \geq u_{ir}) \qquad (10)\\
&\quad \times \prod_{\forall j \in S_i} \Pr(u_i^* \geq u_{ij} \forall j \in S_i) \prod_{\forall k, \forall j' \notin S_i} \Pr(u_i^* \geq R_{ij'}^k)\\
L &= \prod_i L_i \qquad (11)
\end{aligned}$$

The model is estimated using maximum likelihood.[13]

4.5 Identification

We overview the empirical identification of the model in this section.

In the data we observe (i) hotel prices variation across hotels and consumers, (ii) other hotel attributes (excluding prices) variation across hotels, (iii) hotel slot position changes

[13]Potentially an additional condition can be introduced into the likelihood. That is, when the consumer makes a search on hotel j's using a specific refinement method k, the reservation utility of that search opportunity is the greatest among all unexplored search opportunities, i.e., $R_{ij}^k \geq R_{ij'}^{k'}, \forall k', \forall j' \notin S_{ij}$. We can construct a simulated probability of $\Pr(R_{ij}^k \geq R_{ij'}^{k'})$ by adding disturbance terms to R_{ij}^k (equation A3 in Appendix B). However, we do not see a theoretically appealing explanation for the source of such disturbance terms.

across consumers due to different refinement tools used, (iv) consumer search, refinement, and purchase behavior.

First, the mean levels of preference parameters (α, β) are identified from the correlation between (1) hotel prices and other attributes and (2) consumer search/refinement/purchase shares of hotels. If consumers have a higher sensitivity on a specific hotel attribute, then (1) the variation on that attribute tends to have a greater impact on which hotels are more likely to be searched and purchased, and (2) the sorting/filtering on that attribute are more likely to be observed.

Second, conditioned on the mean utility levels, the correlation between (1) consumer search/purchase frequencies of hotels and (2) hotel slot positions enable the identification of mean parameters of search cost (γ_0, γ_1). In particular, conditioned on the attributes of a hotel, the search cost as a function of the slot position affects how frequent the hotel is searched.

Third, we are able to separately identify preference and search cost. In the data, when hotels have the same average slot position hence similar search cost, the different search/purchase frequencies will be attributed to the variation of hotel attributes. Similarly, when hotels have similar attribute levels hence similar utility levels, the variation of search/purchase frequencies will be attributed to the difference in search cost (slot positions). Further, preference and search cost enter the reservation utility in a nonlinear fashion (equation 4), this further helps the identification.

Fourth, we observe various search, refinement, and purchase behavior across consumers even when they face similar hotel lists. Such observations show deviations from the implied behavior patterns based on the mean levels of preference and search cost parameters. As a result, these observations enable the identification of heterogeneity.

Finally, even though we include purchase data in our estimation, they are not essential for identification. We consider a robustness test by dropping the sparse purchase data from the estimation. The test shows that there is little effect in the estimates. This is because even

without the purchase component, the likelihood function is still properly defined (equation 11).

5 Results

In this section we report the results of the estimation and the fit information of the model.

5.1 Parameter Estimates

Table 3 reports the parameter estimates. For simplicity, we fit the model with the heterogeneity only on the attributes that the consumers may sort or filter, the cost parameters, and the hotel chain intercepts. The heterogeneous hotel intercepts helps to mitigate the potential endogeneity biases caused by the correlation between error terms and variables.

On average, the price and consumer rating have the highest impact on consumer utility. In particular, consumer rating has a premium for consumer utility. For a one-tenth unit increase in consumer rating, the effect on utility is equivalent to a daily price decrease of \$5.8 ($= 1.49/2.57 \times 10$). Other significant factors that affect utility are star rating, promotion, and air condition in the room.

Search cost has important implications for consumer search behavior. First, inferior slot ranking decreases a hotel's chance of being searched. Consider the following heuristic example. At this website, each webpage has 25 hotel links. If a consumer has already found some hotel with positive utility on the first page, assuming having the same level of hotel attributes and no uncertainty, the first hotel on the second page needs to lower the daily price by \$8.07 to attract a click ($= \exp(-1.09 + 0.07 \times 26)/2.57 \times 10$). Second, using sorting and filtering, a consumer may decrease her search cost by placing those hotels with high expected utility on more prominent positions. Hence the existence of refinement becomes especially beneficial for consumers due to the significant search cost.

We also find that consumers demonstrate considerable heterogeneity in utility. For example, although the mean level of price coefficient is higher than that of star rating coefficient, a consumer may be much more sensitive to star rating than to price due to preference heterogeneity. As a result, this consumer may use some alternative refinement methods that prioritize star ratings and choose different hotels than an average consumer. We will further explore the ramification of the heterogeneity on market structure in the policy simulation section.

Table 3: Model Estimates[a,b]

	Mean Parameters (Std. Err.)	Heterogeneity (Std. Err.)
Log(Search Cost)		
Constant	**-1.09 (0.04)**	**0.90 (0.20)**
Slot Position	**0.07 (0.01)**	**0.87 (0.04)**
Utility		
Average Daily Price ($10)	**2.57 (0.63)**	**1.01 (0.20)**
Star Rating[c]	**0.37 (0.14)**	**1.88 (0.09)**
Consumer Rating[d]	**1.49 (0.22)**	**0.57 (0.08)**
Promotional Event	**0.60 (0.26)**	–
Price Information Missing	-0.01 (0.58)	–
Free Internet Access	0.57 (0.87)	–
Air Condition in Room	**1.04 (0.31)**	–
Swimming Pool	0.12 (1.12)	–

Note: *a*. Bold fonts indicate the estimates being significant at 95% level.
b. To save space, we do not report the respective intercepts and the heterogeneity of hotel chains. None of the intercepts and heterogeneity is significant.
c. Star rating ranges from one to five with increments of half-stars.
d. Consumer rating ranges from 0 to 5 with increments of 0.1.

5.2 Model Validation

To examine the fit of the model, we consider three tests using a holdout dataset where the checkin dates are between October 25 and October 31 instead of between November 1 and November 14. As for the sample size, the number of hotels is comparable to the data used

for estimation and the number of consumers are roughly half of the size. The levels of hotel attributes and prices are also similar.

We begin by calculating the hit rate of hotel search using the holdout sample.[14] To be specific, for each individual consumer, we generate $R_1 = 100$ sets of preference parameter values and $R_2 = 100$ search cost parameter values using the model estimates. For each consumer, we also generate $R_3 = 100$ random utility shocks per hotel (v_{ij}). Conditioned on the observed hotel attributes levels, prices and slot positions, for each set of parameter values and consumer-hotel random shocks, we can infer which hotel is searched by consumer i using the optimal search model described in section 2. We repeat the exercise using all parameter draws and random utility shocks and then calculate the hit rate. We find the hit rate for the holdout sample is 0.77, suggesting our model captures the search behavior well.

Next we consider the predicted "market share" of search (or average probability of being searched) of each hotel. We again draw 100 sets of preference and search cost parameter values. For each consumer and each hotel, we also draw 100 sets of random utility shocks. For each consumer, the predicted probability of a hotel being searched can be obtained by integrating out the random draws of parameters and utility shocks. We then aggregate the predicted probability of search of each hotel over all consumers, obtaining the "market share" of search. We compute the correlation between the predicted shares and the observed shares. The correlation is 0.84.

Consumers have heterogenous sensitivities for each of the hotel attributes, which is the main reason why they use different refinement methods. To the extent that the refinement method used reflects the heterogeneity of consumers, we consider a third test to examine the model's ability of recovering the heterogeneity. Using a similar approach, we predict the

[14]Since the booking data are very sparse, we opt to calculate the hit rate of search rather than booking. To the extent that booking is predicated upon search, this test also shed some insight about the model fit regarding hotel booking.

"market shares" of sorting/filtering methods. We then calculate and compare the predicted shares with the observed ones. The correlation is 0.81.[15]

As for in-sample fit, the corresponding hit rate and the two correlation coefficients are 0.81, 0.88, and 0.83, respectively. Overall the model fits well.

5.3 Robustness Checks

In this section, we consider several robustness checks pertaining to the current specification of the model.

5.3.1 Alternative Information Structure

By default, the search results of hotels are ranked according to the frequencies of purchases. However, since the information about this default ranking rule is rather obscure at the website, we assume that the consumers do not know the rule. This assumption implies that when the consumers view the default list, they treat the hotel attributes uncorrelated with slot positions. To be consistent with this assumption in the estimation, when a consumer is facing the default ranking, for each slot we randomly draw the attributes from a joint distribution that is independent of the slot position. The distribution is obtained as the empirical distribution from the data.

On the other hand, it is possible that consumers know the default rule during the search. In particular, they may infer that top hotels on the default list are those on average being more preferred by the population (higher frequencies of purchases). This alternative assumption implies that, under the default ranking, consumers know that more preferable attribute levels are more likely to be observed at top positions than at inferior positions.

To test this alternative assumption, we consider two approaches:

[15]Since there are only 8 refinement methods considered (including the default one), to achieve a better accuracy of the correlation, we use bootstrapping to generate 25 sets of simulated shares (so there are 200 observations of simulated shares). We then replicate the observed shares 25 times to match the bootstrapped data.

1. We re-estimate the model under the alternative assumption, i.e., consumers understand the default ranking rule. In particular, when a consumer faces the default ranking, instead of drawing the attributes from a distribution that is independent of slot positions, we draw them from distributions that are slot-specific, obtained as the empirical distributions from the data. The model fit deteriorates as measured by BIC (711.16 vs. 740.10). The out-of-sample fit also becomes worse. We use the same three measures for out-of-sample fit as in section 5.2, (1) hit rate, (2) the correlation between the predicted shares of searched hotels and the observed ones, and (3) the correlation between the predicted shares of refinement methods and the observed ones. The measures change from 0.75, 0.83, and 0.80 to 0.69, 0.79, and 0.71. We take these as evidence that the original assumption (consumers do not know the rule) may be more appropriate for the data.

2. About 20% of the consumers in the data can be identified as "registered users" since they logged into their accounts before the search. The remaining 80% of the consumers were either new to the site or did not log in. It is reasonable to expect that, if some consumers understand the default ranking rule, it is more likely to be those 20% registered users. We pool all consumers together, with checkin dates from October 25 to November 14. We then use the 20% "registered users" (66 in total) to re-estimate the model under the original assumption and the alternative one. The model fit as measured by BIC is better under the original assumption (301.36 vs. 330.21). This result implies that even for those who are more likely to understand the default ranking rule, the original assumption seems more appropriate.

5.3.2 Alternative Search Cost Specifications

The search cost is specified as

$$\begin{aligned} c_{ij}^k &= \int_{z,x} c_i(\text{Slot}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \\ &= \int_{z,x} \exp(\gamma_{i0} + \gamma_{i1} \text{Slot}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \end{aligned}$$

i.e., the search cost is increasing in slot position since it takes more time and efforts for consumers to go down the hotel list. Under this specification, when a hotel moves down the list, the search cost increases. We further consider two alternative specifications regarding search cost:

1. The search cost of a hotel is determined by the page number of the webpage where it locates. For hotels on the same webpage, their search costs are at the same level. The cost increases with more page turning. Under this specification, we have

$$\begin{aligned} c_{ij}^k &= \int_{z,x} c_i(\text{Page}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \\ &= \int_{z,x} \exp(\gamma_{i0} + \gamma_{i1} \text{Page}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \end{aligned}$$

   Page_j^k is the webpage number where j locates. For example, since there are 25 hotels per page, hotels from slot 51 to slot 75 have the same cost level and $\text{Page}_j^k = 3$ for $j = 51, 52, ..., 75$.

2. The search cost of a hotel is determined by both the slot position and the number of page turning. The search cost increases with slot position and page turning. In particular,

$$\begin{aligned} c_{ij}^k &= \int_{z,x} c_i(\text{Slot}_j^k, \text{Page}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \\ &= \int_{z,x} \exp(\gamma_{i0} + \gamma_{i1} \text{Slot}_j^k + \gamma_{i2} \text{Page}_j^k) dP^k(z_{ij}, x_j | S_{ij}) \end{aligned}$$

where $1 \leq \text{Slot}_j^k \leq 25$ is the slot on a specific webpage, and Page_j^k is the webpage number where j locates. For example, if hotel j is positioned at slot 51 under refinement method k, then $\text{Page}_j^k = 3$ and $\text{Slot}_j^k = 1$ (i.e., hotel j locates on page 3, slot 1.).

Under these two alternative search cost specification, the utility estimates are essentially the same but the model fit deteriorates as measured by BIC (711.16 vs. 723.03 and 711.16 vs. 731.10, respectively). More importantly, the coefficients of Page_j^k are insignificant in both alternative specifications, potentially due to the sparse observations of page-turning among consumers.

6 Managerial Implications

6.1 Refinement and Consumer Welfare

6.1.1 The Effect of Refinement on Consumer Welfare

High search cost limits a consumer's search, potentially resulting in choosing options with lower utilities. As discussed earlier, the search cost for finding the more preferred hotels can be decreased due to the ability of refining the search results along the dimension that matters to the consumer the most. To empirically investigate this insight, we simulate the searching and booking outcomes of all consumers with the refinement ability and without the refinement ability. With the sorting/filter options, on average each consumer makes 2.91 searches during the searching process.[16] In comparison, when the sorting/filtering options are removed, the average drops to 1.85, a 36% decrease. In terms of the utility for the hotels booked, the average utility decreases by 17% when the refinement are removed.[17]

Although the final utility for the booked hotel becomes higher with the refinement, it is still unclear what the impact is on the overall welfare of consumers. If under the refinement

[16]This is consistent with the data where we observe on average 2.86 searches per consumer.

[17]Because the booking is very sparse in the original data, we use bootstrap to expand the data 50 times so as to measure the utility level of booked hotels.

the number of search increases disproportionally to the utility improvement, the accumulated search cost incurred during the search process may well outweigh the benefits, hence lowering the overall consumer welfare. To evaluate the overall consumer welfare, we further compute the net surplus of search as measured by the final utility of the booked hotel net the total search cost. Surprisingly, we find that on average the net surplus decreases by 4% with the existence of refinement and the loss of the welfare surplus has a 95% confidence interval of (-6.1%, -1.3%). To better understand this seemingly counter-intuitive result, recall that the default ranking of hotels already reflects the average utility levels among the population regarding these hotels. Even when there are no refinement tools, the baseline level of consumer welfare is fairly high if consumers make decisions according to the default ranking. As a result, the main reason of the welfare reduction in face of the refinement is that consumers do not understand the default ranking rule and disproportionally made more searches.

To further explore this insight, we consider an additional simulation. We again simulate the searching and booking outcomes of all consumers with and without the refinement ability. However, in this simulation we assume that consumers are educated about the default ranking rule. To be specific, under the default ranking of hotels, the attributes are drawn from distributions that are slot-specific. Through this simulation, we find that the average number of searches with the availability of refinement is 2.16, compared to 1.70 when refinement are unavailable. These numbers of searches are smaller than those when consumers are uninformed about the default ranking rule (2.91 and 1.85, respectively). We also find that net welfare surplus instead increases by 3.3% with the refinement, compared to the drop of 4% with uninformed consumers. The 95% confidence interval of the welfare improvement is (1.4%, 4.3%). This is consistent with our conjecture, i.e., uninformed consumers engage in disproportionally more searches, leading to the deterioration in net welfare surplus. In contrast, when they understand the default ranking rule, the refinement tools instead improve the net welfare surplus..

6.1.2 Alternative Default Ranking

Recall that if two hotels have equivalent ranking after the refinement, they will further be ranked according to the undisclosed default algorithm. In addition to educating consumers about the default ranking rule, it is possible to further enhance net consumer surplus with refinement by provide an alternative default ranking scheme using additional information. Ghose et al. (2012) show that a website can improve consumer welfare by directly ranking products by consumer utilities.

To investigate the effect of this alternative ranking method on consumer welfare, we consider the following policy simulation:

1. We generate $R_1 = 100$ sets of preference parameter values using the model estimates. Conditional on the observed hotel attribute levels, we obtain the mean utility of each hotel by integrating over the $R_1 = 100$ preference parameter values.

2. Instead of using the observed default ranking of hotels, we rank the hotels based on the imputed mean utilities of hotels. In particular, after sorting/filtering, if two hotels have the same implied ranking based on the refinement method, the ranking will further be determined by their mean utilities.

3. For a given consumer, we assume that this consumer knows this new default ranking rule. We then generate 100 sets of search cost parameter values as well as 100 random utility shocks per hotel.

    (a) Conditioned on the observed hotel attributes levels, prices, and the new positions based on the mean utility ranking, for this consumer and a given set of parameter values and hotel random error shocks, we can infer which hotel is searched by this consumer using the optimal search model described in section 2. We can then compute the utility of the booked hotel and the net welfare of this consumer.

(b) To compute the *expected* utility of the booked hotel and the net welfare of this consumer, we need to integrate over the distributions of preference, search cost, and random shocks by repeat Step 3(a) using all parameter draws and random utility shocks and then calculate the average.

4. We repeat Step 3 for all consumers and then sum the results to calculate the total net welfare and total utility of booked hotels.

In comparison to the observed default ranking of the website, the utility of the booked hotels increases by 5.2% with a 95% confidence interval of (3.0%, 6.1%). The total net welfare increases by 7% with a 95% confidence interval of (5.9%, 9.7%). The current simulation result implies that the combination of the refinement tools and the new ranking by mean utilities would not only further increase the utility of booked hotels, but also enhance the overall net welfare surplus by about 4% (= 7% − 3.3%). Even if we use the bounds of the 95% confidence intervals to calculate the net welfare change, the improvement in the net welfare is still significantly higher than that under the website's current default ranking algorithm (1.6% = 5.9% − 4.3%). We further compare the average number of searches under the alternative ranking method and the current default ranking method. We find that under the new ranking method the number of searches is 1.99, lower than the 2.16 searches under the current default ranking method. In other words, the improvement in the net welfare mainly comes from the decrease in the number of searches and further enhancement of the utility of the booked hotels.

6.2 Refinement and Market Structure

The ability of refining the search results may affect the market structure. First, the magnitude of search cost affects the market structure (e.g., Nelson (1970)). When there is no refinement options, hotels positioned low on the default list incur higher search cost and hence are less likely to be clicked by consumers. This renders those hotels on superior slots

on the default list higher market shares at the expense of the lower ranked hotels, making the market less competitive. Sorting and filtering lower the search cost and provide easier access to low-ranked hotels than otherwise. As a result, the market becomes more competitive.

Second, consumers face the same slot ranking under the default list. It may be too costly for a consumer to reach those more preferred hotels if they are ranked low on the default list. Without the refinement options, it is possible that most consumers will be limited to the top ranked hotels due to search cost even though they would have chosen differently otherwise. Only those consumers with relatively lower cost may search further down the list. Consequently, the top-ranked hotels on the default list have higher market shares. In comparison, when sorting and filtering are available, consumers will use different sorting/filtering methods as they have heterogeneous preferences. The choices are no longer only limited to the top hotels on the default list. Thus the market becomes more competitive.

To explore the impact of refinement on market structure, we start by calculating the Herfindahl-Hirschman Index of search shares under the current market condition: heterogenous consumers with refinement options.[18] Herfindahl-Hirschman Index (HHI) is a measure of the intensity of market competition, defined as

$$HHI = \sum_{j=1}^{H} s_j^2$$

where s_j is the market share of firm j. We calculate search shares using the same method in section 5.2. Under the current market condition, the HHI takes the value of 0.21, indicating moderate competition in the market.[19]

Next we remove the refinement options such that all consumers face the same default hotel list. Under this new market condition, the HHI increases to 0.32, showing a high level of market concentration among the top hotels on the default list. We further remove the

[18]We use search share rather than booking share because each consumer booked at most one hotel and many chose the outside option (do not book any hotel), resulting in many zeros in the booking data.

[19]Horizontal Merger Guidelines, US Department of Justice and the Federal Trade Commission, 2010.

heterogeneity of preference and search cost among consumers. In this case, the HHI increases by another 6%, reaching 0.34. In conclusion, the refinement, together with the heterogeneity of consumers, making the market less concentrated.

7 Conclusion

With the $780 billion global value of search technology generated annually, the impact of the technology on consumer and business is of great interest for both the industry and the academia. In particular, the ability of consumers sorting and filtering search results has substantial effects on consumer and firm behavior. Empirical research pertaining to disentangling and measuring such effects is surprisingly absent. This paper proposes a structural model of consumer optimal online search with the ability of refining the search results using sorting and filtering. Using a unique data set of individual online search activities with refinement, we are able to estimate the preferences and search costs of heterogeneous consumers, providing insights about consumer decisions in face of uncertainty about products and the ability of refining search results.

Our modeling approach has a few important features. First, the model is consistent with classical optimal information search theory and has a solid theoretical foundation. The model explicitly model consumer search as a utility maximization process. Second, although it may be unrealistic, many previous studies assume that consumers have perfect knowledge about product attributes for the sake of tractability. Instead, our model allows uncertainty to be resolved during the search. More importantly, a consumer's refinement decisions will affect how uncertainty being resolved by changing the search cost and the distribution of product attributes. As a result, the decisions of search and refinement are coherently integrated into the utility optimization framework. Third, the model can be applied to other online search contexts where consumers have the ability of refining results. The refinement is not just limited to sorting and filtering. The refinement can include many other search features

that can potentially lower search cost and affect the distribution of products. For example, at shopping websites such as Amazon.com, the website makes suggestions about categories and products related to a consumer's search as well as corrects misspelled keywords. These features can be easily incorporated into the model. As a result, the model has a broader applicability.

The model has a decent out-of-sample fit. In particular, it has the ability to recover the pattern of consumer heterogeneity. Conditioned on the estimates, we consider several policy simulations. First, we find that consumers have lower search cost with the aid of refinement tools. They make 57% more searches on average and are able to obtain about 20% higher utilities from the products they choose. Second, although the utility of the purchased product increases, the overall welfare surplus drops for consumers. The default rule reflects the booking frequency ranking of hotels among consumers. As a result, the baseline level of consumer welfare is fairly high even without the aid of refinement tools. The welfare reduction in face of the refinement is mainly because that consumers do not understand the ranking rule and disproportionally made excessive searches. To address such deterioration in net welfare, we show that by simply educating the consumes about this rule, it will improve consumer welfare. We further suggest a new ranking method that can further enhance consumer welfare. This new ranking rule uses the imputed mean utilities of products. The new default ranking method has the ability of improving both the utility of purchased product and the net welfare. Third, we also find the refinement tools make the market less concentrated. This is because (1) the lowered search cost enables the consumers to search with a greater depth and reach more hotels, (2) the refinement tools facilitate heterogeneous consumers to find hotels that match their preferences better. Such matches would be too costly without the refinement tools.

Several extensions to the current model are possible. First, in our model, although consumers do not know product attributes before the search and use search to resolve the uncertainty, we assume consumers know the distribution of attributes. This assumption is

reasonable in the current context as the attributes level of hotels across time are relatively stable. However, in other contexts when consumers face some unfamiliar product category, the distribution may also be unknown and consumers need to update their beliefs about the distribution based on every round of search. Adam (2001) proposes a theoretical optimal search model where the agent learns the profits distribution of alternative options during search by following some belief updating rules. Integrating learning into the model will certainly enhance our understanding of the consumer search process.

Second, a search engine can be interpreted as a platform of a two-sided market that facilitates the interactions between sellers and consumers (Yao and Mela (2011)). The current model focuses on the demand side. Extending the model by including the supply side will further enrich our insights about the interactions of the two sides of the market. The inclusion of the supply side also enables additional policy simulations such as the strategic interactions among the sellers.

Finally, over the long run, how consumers and firms adapt to the advance of search technology will be a fruitful avenue for future research. For example, the advance of search technology enables consumers to engage in more extensive search for lower prices. Consequently, this intensifies the price competition among firms. Ellison and Ellison (2009) show that, to minimize damages, firms may start to engage in information obfuscation, making obtaining their product information from the search engine more difficult for consumers. Overall, we hope this paper will inspire more future research on online search.

References

Adam, Klaus. 2001. Learning while searching for the best alternative. *Journal of Economic Theory* **101**(1) 252 – 280.

Ansari, Asim, Carl F Mela. 2003. E-customization. *Journal of Marketing Research* **40**(2) 131–145.

Ellison, Glenn, Sara Fisher Ellison. 2009. Search, obfuscation, and price elasticities on the internet. *Econometrica* **77**(2) 427–452.

Ghose, Anindya, Panagiotis G. Ipeirotis, Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Science* forthcoming.

Greene, William H. 2003. *Econometric analysis*. Prentice Hall.

Hong, Han, Matthew Shum. 2006. Using price distributions to estimate search costs. *Rand Journal of Economics* **37**(2) 257–276.

Honka, Elisabeth. 2011. Quantifying search and switching costs in the u.s. auto insurance industry. *Working Paper* .

Hortacsu, Ali, Chad 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**(2) 403–456.

Kim, Jun, Paulo Albuquerque, Bart J. Bronnenberg. 2010. Online demand under limited consumer search. *Marketing Science* **29**(6) 1001–1023.

Koulayev, Sergei. 2010a. Estimating demand in online search markets, with application to hotel bookings. *Working Paper* .

Koulayev, Sergei. 2010b. Search with dirichlet priors: estimation and implications for consumer demand. *Working Paper* .

Mehta, Nitin, Surendra Rajiv, Kannan Srinivasan. 2003. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Science* **22**(1) 58–84.

Nelson, Phillip. 1970. Information and consumer behavior. *The Journal of Political Economy* **78**(2) pp. 311–329.

Santos, Babur De los, Ali Hortacsu, Matthijs R. Wildenbees. 2011. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review* forthcoming.

Seiler, Stephan. 2011. The impact of search costs on consumer behavior: A dynamic approach. *Working Paper* .

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press.

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

Yao, Song, Carl F. Mela. 2011. A dynamic model of sponsored search advertising. *Marketing Science* **30**(3) 447–468.

Appendix

A Existence and Uniqueness of the Solution to $Q_{ij}^k(S_{ij}, R_{ij}^k) = 0$

It can be shown that equation 7 has a unique solution with respect to R_{ij}^k. To see this, for $Q_{ij}^k(S_{ij}, R_{ij}^k)$ take partial derivative with respect to R_{ij}^k, we have

$$\begin{aligned}\frac{\partial Q_{ij}^k(S_{ij}, R_{ij}^k)}{\partial R_{ij}^k} &= \frac{\partial \int_{R_{ij}^k}^{\infty}(u_{ij} - R_{ij}^k)dF^k(u_{ij})}{\partial R_{ij}^k} \\ &= -[(u_{ij} - R_{ij}^k)f^k(u_{ij})]_{u_{ij}=R_{ij}^k} + \int_{R_{ij}^k}^{\infty}(-1)f^k(u_{ij})du_{ij} \\ &= -(1 - F^k(u_{ij})) \leq 0\end{aligned}$$

That is, Q_{ij}^k is monotonically decreasing in R_{ij}^k if $F^k(u_{ij}) < 1$. Further, since

$$Q_{ij}^k(S_{ij}, R_{ij}^k) = \begin{cases}\infty, & \text{if } R_{ij}^k = -\infty \\ -c_{ij}^k, & \text{if } R_{ij}^k = \infty\end{cases}$$

As a result, there exists a unique solution to $Q_{ij}^k(S_{i(j-1)}, R_{ij}^k) = 0$.

B Calculating Reservation Utilities

According to the definition of reservation utility, we have[20]

$$\begin{aligned}c_{ij}^k &= \int_{R_{ij}^k}^{\infty}(u_{ij} - R_{ij}^k)dF^k(u_{ij}) \\ &= [1 - F^k(R_{ij}^k)]\int_{R_{ij}^k}^{\infty}(u_{ij} - R_{ij}^k)\frac{f^k(u_{ij})}{(1 - F^k(R_{ij}^k))}du_{ij}\end{aligned} \tag{A1}$$

[20]This method of calculating the reserve utility is adapted from Kim et al. (2010).

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

The utility is defined as $u_{ij} = z_{ij}\alpha_i + x_j\beta_i + \nu_{ij}$. Denote the mean utility level as $\mu_{ij} = z_{ij}\alpha_i + x_j\beta_i$. Given a set of attribute values of z_{ij}, x_j and the distribution of $\nu_{ij} \sim N(0,1)$, $u_{ij}|z_{ij}, x_j \sim N(\mu_{ij}, 1)$. So we may write

$$c_{ij}^k = \{(1 - \Phi(R_{ij}^k - \mu_{ij})) \int_{R_{ij}^k}^{\infty} (u_{ij} - R_{ij}^k) \frac{\phi(u_{ij} - \mu_{ij})}{(1 - \Phi(R_{ij}^k - \mu_{ij}))} du_{ij}\} \tag{A2}$$

Using the formulas of the expectation of truncated normal distribution, equation A2 can be rewritten as[21]

$$c_{ij}^k = \{(1 - \Phi(R_{ij}^k - \mu_{ij}))(\mu_{ij} - R_{ij}^k + \frac{\phi(R_{ij}^k - \mu_{ij})}{(1 - \Phi(R_{ij}^k - \mu_{ij}))})\} \tag{A3}$$

As shown in Appendix A, when $F^k(u_{ij}) < 1$, Q_{ij}^k is monotonically decreasing in R_{ij}^k. Consequently, for each pair of search cost c_{ij}^k and μ_{ij}, we can solve the reservation utility R_{ij}^k using equation A3. To facilitate the computation, we can utilize this result to construct a mapping table between (c, μ) and R outside the estimation step. Here we drop the subscripts (i, j) and superscript k since the table holds for all (i, j, k). This table does not depend on specific parameter values. Then during the estimation, for any given values of c and μ, we can use the table to check out the value of reservation utility R, potentially with an interpolation step. In particular, note that $\mu_{ij} = z_{ij}\alpha_i + x_j\beta_i$. To account for the fact that hotel attributes x_j and z_{ij} may be random and follows some distribution $P^k(z_{ij}, x_j|S_{ij})$, we can make arbitrarily large number of draws for the attributes, calculate corresponding μ's, then compute the reservation utility R with the uncertainty integrated out. Since the table is constructed outside the estimation step, a finer grid imposes no further computational burden for the estimation.

[21] Suppose $x \sim N(\mu, \sigma^2)$, $E(x|x \geq a) = \mu + \sigma\lambda(\frac{a-\mu}{\sigma})$, where $\lambda(\frac{a-\mu}{\sigma})$ is the hazard function such that $\lambda(\frac{a-\mu}{\sigma}) = \phi(\frac{a-\mu}{\sigma})/[1 - \Phi(\frac{a-\mu}{\sigma})]$ (e.g., Greene (2003)).