

# Consumer Data and Consumer Welfare: Evidence from the Hotel Booking Market.

Andrey Minaev\*

This version: November 15, 2020

[Please click here for the latest version](#)

## Abstract

I study how the information a search intermediary has about consumer preferences impacts the market. Consumers participate in costly search among different sellers' products, relying on the rankings order provided by the intermediary based on their preferences. Better product targeting affects consumer search and purchases, which, in turn, changes the seller pricing incentives. I considered these aspects by modeling both sides of the market under various ranking algorithms used by the intermediary. On the demand side, I developed a joint model of consumer costly search and purchase. On the supply side, I considered the sellers' pricing competition. To estimate the demand and supply models, I utilized a rich dataset provided by Expedia, which includes consumer search and purchase data and information on the hotels and prices they charge. I find that if the intermediary uses data on consumers' preferences to provide them personalized rankings of products, consumers, on average, experience a 3.6% (\$4.9) utility decrease due to increased transaction prices, a 0.8% (\$1.1) utility gain due to a reduction in search spending, and 0.5% (\$0.7) utility gain due to finding a better-fitted hotel.

**JEL classification:** D12, D83, L13, L83.

**Keywords:** Big Data, Consumer Search, Online advertising, E-commerce, Intermediaries, Platforms.

---

\*Department of Economics, University of North Carolina at Chapel Hill. E-mail: andrey@unc.edu. Acknowledgments: The earlier version of this paper was presented at the 2020 Annual Oligo Workshop. I would like to thank my advisors Fei Li and Brian McManus for providing immeasurable guidance and support. I am grateful to Gary Biglaiser, Luca Maini, Jonathan Williams and Andrew Yates as well as participants of the industrial organization seminar at the University of North Carolina at Chapel Hill for providing helpful feedback. I deeply appreciate Alina Malkova for helpful comments and support. Any errors are mine.

# Introduction

Platforms like Google, Amazon, Facebook, Expedia, etc., collect enormous amounts of data about consumers' preferences and behavior. Although they claim using these data to provide better services to customers, a discussion has recently been raging about should we allow these tech giants to collect and use our personal data. While the central part of this discussion is about privacy and human rights, it also brings economic questions. How market outcomes change if the platform better knows consumers' preferences? How these changes reflect on consumers and the whole economy? Does it really help provide the best service to consumers or simply increase tech giants' potential to control markets?

In many markets, consumers are initially uninformed about the quality of the available products. They may conduct a costly search to learn about product qualities, and in many cases, these searches are facilitated by information intermediaries. For example, online platforms as Amazon and Expedia provide consumers ranked lists of products. As economic activities are increasingly connected through the Internet, consumers can have access to more products at lower search costs, but they also face a much larger set of products to choose from. Thus, consumers are increasingly dependent on intermediaries to guide their search (in some deliberate order) for sellers and products by providing rankings of products. Using personal consumer data on preferences helps platforms to provide more accurate rankings to consumers. This paper highlights how the market outcomes change if platforms collect and use consumer personal data on preferences.

Due to the presence of the search frictions, consumers explore not all products before making purchase decisions. As a result, the platform's ranking algorithm's change leads to a change in consumer demand function since consumers are more likely to explore products on higher positions in the ranking *ceteris paribus*. Better ranking helps consumers easier and faster find better-suited products, reducing search expenditures, and procuring a better product match. However, if consumers change search behavior, in equilibrium, sellers also change their behavior. With a better ranking, consumers find well-suited products higher in the list and have lower incentives to search further, which shrinks their consideration sets and changes the demand elasticity, which, in turn, relaxes competition between sellers and changes their pricing strategies. Thus the effect of better ranking on consumer welfare is ambiguous without additional analysis.

In this paper, I address how the market prices, consumer and economic welfare, and the quality of the purchased products change if the platform can provide consumers better product rankings based on the consumer personal preferences. I compare market outcomes in two different cases: in the first case, the platform provides the personalized rankings of products to consumers based on their personal preferences; in the second one, the platform provides the common ranking to all consumers based on the aggregated data of all consumers preferences.

To address these questions, I utilize the dataset provided by Expedia<sup>1</sup>. It includes consumers' search and purchase data as well as information on the hotels observed by consumers after filling a search query. I provide the equilibrium model to investigate market outcomes' change under different platform's ranking mechanisms. I construct the structural model of optimal consumer choice with the search frictions based on the classical [Weitzman \(1979\)](#) model and estimate consumers' demand. Conditional on this demand, I model hotels' pricing game and use it to estimate hotels' cost. Last, using estimation results, I run simulations to evaluate the market outcomes under different platform's ranking mechanisms.

This paper is the first attempt to estimate the equilibrium model in such a setting. Previous empirical works do not model firms' strategic pricing response on the change of platform's ranking mechanism and estimate only the welfare effects due to the change in consumers' search and purchase behavior. Part of the reason for that is computational difficulty in simulating the change in firms' pricing decisions due to the complicated nature of the demand correspondence accounting for search frictions. I overcome this difficulty by applying modern findings of [Choi et al. \(2018\)](#) and [Moraga-González et al. \(2018\)](#), which allows me to translate the pricing game among the sellers into a familiar discrete-choice problem. The equilibrium model allows me to estimate market prices change and get more accurate results. In contrast to previous research, I show that personalized ranking is harmful to consumers despite the decrease in search expenditures.

I find that under the personalized ranking consumers experience in average .8% (\$1.1) utility gain due to a reduction in search intensity comparative to the common ranking case, since consumers find better-suited product on higher positions. Besides, due to better ranking, consumers on average are able to find better-suited hotels, which increases their utility on average by 0.5% (\$0.7). On the other hand, consumers utility reduce in average by 3.6% (\$4.9) due to increased prices in the case of personalized ranking comparative to the common ranking case. Resulting effect is summarized as average lost of 2.3% (\$3.1). At the same time consumers who are less price sensitive might experience more than 11% (\$15) utility gain and consumers who are more price sensitive lose more than 15% (\$20) of utility.

This study results might argue in the discussion of policy implementation regarding collecting and using personal consumer data. In contrast to previous research, my results show that personal data usage is harmful on average for consumers. Although they might help provide better service to consumers, the market power shifts toward the supply side disproportionately, increasing market prices by higher amounts than consumers' gain. Simultaneously, consumer personal data usage raises economic welfare by reducing search expenditures and helping consumers find better-fitted products. Hence to forbid platforms collecting and using consumers personal data might be not the optimal solution

---

<sup>1</sup>The dataset was originally provided for [the Kaggle competition](#). Expedia provided the allowance to use the dataset for academic purposes after the competition was finished.

because it would lead to reduce in economic welfare. Direct money transfers to consumers for the data they share with companies might be a better solution.

## Insights from prior research

In many markets, consumers search among alternative options before making a purchase. The way that choices are presented to consumers can then have a substantial impact on their search behavior and hence competition and market performance. For example, if (as in standard search models) choices are viewed *ex ante* symmetrically by consumers, then they may search randomly among alternatives, and symmetric outcomes as between firms will tend to result. However, if one purchase option is somehow more prominent than others, consumers are likely to consider that offering first. There is abundant evidence that the way options are presented can significantly influence choice. [Ho and Imai \(2006\)](#) and [Meredith and Salant \(2013\)](#) observe that being listed first on a ballot paper can increase a candidate's vote share. [Einav and Yariv \(2006\)](#) present evidence that economists with surname initials earlier in the alphabet have more successful careers, and discuss various reasons why such researchers may be more "prominent".

[Ursu \(2018\)](#) studies the mechanism of prominence in the settings when consumer participates in the costly product search guided by the intermediary. There are three main components that affect consumer behavior in such settings: the part of utility observed prior to a search, the expectation of part of utility observed after a search, and the search cost. [Ursu \(2018\)](#) shows that placing the product on a higher position does not change the first two and affects only the search cost – consumers face lower search costs to explore the products which appear on the higher positions. [Shen \(2015\)](#) shows that in settings when the price is observed prior to search and consumers search for product quality, firms charge higher prices in the case of lower search costs. The logic is as follows: if the price is observed prior to a search, then the price affects the order in which a consumer conducts a search. Therefore firms have the incentive to lower the price to attract consumers to visit them first. This incentive to become a prominent firm by reducing the price increases with the search cost because consumers visit fewer firms. Furthermore, [Choi et al. \(2018\)](#) show in the setting of firm heterogeneous search cost, that firms with lower inspection cost could charge higher prices than firms with higher inspection costs.

Therefore, if the platform can observe consumer's personal preferences and place better-suited to him products on higher positions, making them prominent and reducing the cost of their inspection, that leads to a reduction of a number of options the consumer considers, which, in turn, leads to decrease of market competition and price increase.

The rest of the paper is organized as follows: In section [1.1](#) I provide the motivating example. Section [2](#) introduces the empirical demand and supply model used in this study. The details of dataset is discussed in section [3](#). Section [4](#) provides the results of

estimation. In section 5 I provide the main results – market simulations under different data allowance policies. Section 6 is a concluding remark.

## 1.1 Motivating example

As an illustration of the logic of the mechanism of how the ranking affects prices, here I discuss the toy example. The main objective of the example is to demonstrate the difference in the market outcomes in the case when platform can provide the personal ranking to each consumer based on consumer’s preferences and when the platform have to provide the common ranking to all consumers based on the aggregate preferences of these consumers. The main point of the example is demonstrated on Figure 1, which highlights that firms charge higher prices if the platform is allowed to rank products according to personal consumers’ preferences rather than use the common ranking to all consumers.

The economy consist of two firms  $A$  and  $B$ , selling products  $a$  and  $b$  respectively, unit mass of consumers and the platform. Each consumer has a unit demand and does not have any outside option. The platform is the only place where the consumers can purchase the product. Consumers do not observe the entire product matching quality and have to pay the search cost to explore it. Though, prior to search consumers observe the part of product’s matching quality and observe the second part after the search. The platform guides the consumers’ search process providing the ranking of products and placing products with higher potential matching qualities on top positions in ranking. More details about platform’s role is provided below. Firms compete in prices and set them optimally conditional on consumers behavior. Firms objective is maximizing the profit. The marginal cost of both products are normalized to zero.

If the consumer  $i$  purchases product  $j$ , his utility can be described as:

$$U_{ij} = u_{ij} - p_j = \delta_{ij} + \epsilon_{ij} - p_j,$$

where  $\delta_{ij}$  and  $\epsilon_{ij}$  are parts of utility observed prior and after the search respectively, and  $p_j$  is the price of product  $j$ .  $\epsilon_{ij}$  is assumed to be a random draw from the exponential distribution with parameter 1 and be uncorrelated among consumers and firms.

Consumers are different in their valuations of products.  $\epsilon_{ij}$  is iid across consumers and products, though consumers value differently  $\delta_{ij}$ , the product’s part of utility, observed prior to search. Two thirds of consumers (labeled Consumer 1) has preferences  $\delta_{ia} = \delta$ , and  $\delta_{ib} = 0$ , while the remaining one third of consumers (labeled Consumer 2) has preferences  $\delta_{ia} = 0$ , and  $\delta_{ib} = \delta$ . Consumers’ product values are illustrated in Table 1.

Table 1: Consumers’ products values

Products	Consumer 1	Consumer 2
a	$\delta + \epsilon_{ia}$	$0 + \epsilon_{ia}$
b	$0 + \epsilon_{ib}$	$\delta + \epsilon_{ib}$

As mentioned above, the platform guides consumer search process providing the ranking of products and placing on higher positions products with higher potential matching qualities. Due to  $\epsilon_{ij}$  is iid among consumers and products, the platform attempt to place on the higher position the product with higher  $\delta_{ij}$ . The goal of this exercise is to compare market outcomes in two scenarios: first, the platform can provide the personal ranking of products to each given consumer and second, the platform has to provide the same ranking to all consumers. In the first scenario, platform will place the product  $a$  on higher position for two thirds of consumers (Consumer 1) and product  $b$  for remaining one third of consumers (Consumer 2). In the second scenario, the best the platform can do is to place product  $a$  higher for all consumers. The rankings under two scenarios are represented in Table 2.

Table 2: Positions of products under common and personal rankings

Position	Common ranking		Personal ranking	
	Consumer 1	Consumer 2	Consumer 1	Consumer 2
1	a	a	a	b
2	b	b	b	a

In accordance with Ursu (2018), I let the search cost differ over positions. Consumers pay zero cost to explore  $\epsilon_i$  of the product placed on the first position, while consumer  $i$  have to pay search cost  $s_i$  to explore  $\epsilon_i$  of the product placed on the second position.  $s_i$  is assumed to be a random draw from standard uniform distribution  $U[0, 1]$  and be uncorrelated among consumers.

Choi et al. (2018) shows that as a result of optimal search and purchase decisions, rational consumer purchases the product with the highest  $w_{ij} - p_j$ , where  $w_{ij}$  is defined in Equation 1.

$$w_{ij} = \min\{u_{ij}, r_{ij}\}, \quad (1)$$

where  $r_{ij}$  is the reservation utility of product  $j$  for consumer  $i$ , i.e. such utility level that the consumer  $i$  is indifferent between obtaining utility  $r_{ij}$  immediately and visiting seller  $j$ . The mathematical definition of reservation utility  $r_{ij}$  is provided as a solution of Equation 2 in  $r_{ij}$ .

$$s_{ij} = \int_{r_{ij}}^{\infty} (u - r_{ij}) dF(u) = \int_{r_{ij}-\delta_{ij}}^{\infty} (\epsilon - r_{ij}) dF(\epsilon) \quad (2)$$

Due the assumption that  $\epsilon \sim \text{Exp}(1)$ , Equation 2 can be solve in the closed form and the reservation utility can be decomposed into a utility observed prior to search

component and a search cost component:

$$r_{ij} = \delta_{ij} + \log \left( \frac{1}{s_{ij}} \right) \quad (3)$$

As a result, the Equation 1 can be rewritten as

$$w_{ij} = \delta_{ij} + \min\{\epsilon_{ij}, \log \left( \frac{1}{s_{ij}} \right)\}, \quad (4)$$

Due to  $\epsilon_{ij}$  is iid over consumers and products and  $s_{ij}$  depends only on the position of the product in the ranking but not the identity of the product itself, the distribution of the second additive part in the equation above depends only on the position of the product in the ranking. If the product  $j$  is listed on the first position, then  $s_{ij} = 0$ , and hence  $\min\{\epsilon_{ij}, \log \left( \frac{1}{s_{ij}} \right)\}$  follows an exponential distribution with parameter 1. If the product  $j$  is listed on the second position, then  $s_{ij} \sim U[0, 1]$ , which makes  $\log \left( \frac{1}{s_{ij}} \right)$  to follow the exponential distribution with parameter 1, and hence  $\min\{\epsilon_{ij}, \log \left( \frac{1}{s_{ij}} \right)\}$  follows an exponential distribution with parameter 2. The distribution of  $w$ 's is summarized in Table 3.

Table 3: The distribution of  $w$ 's under different rankings.

Position	Common ranking		Personal ranking	
	Consumer 1	Consumer 2	Consumer 1	Consumer 2
1	a: $w_{1a} - \delta \sim \text{Exp}(1)$	a: $w_{2a} \sim \text{Exp}(1)$	a: $w_{1a} - \delta \sim \text{Exp}(1)$	b: $w_{2b} - \delta \sim \text{Exp}(1)$
2	b: $w_{1b} \sim \text{Exp}(2)$	b: $w_{2b} - \delta \sim \text{Exp}(2)$	b: $w_{1b} \sim \text{Exp}(2)$	a: $w_{2a} \sim \text{Exp}(2)$

As shown in Choi et al. (2018), each consumer purchases the product with higher realization of  $w$ . As a result, the demands of firm A and firm B can be expressed as shown in Equation 5.

$$\begin{aligned} D_A(p_A, p_B) &= \frac{2}{3} \Pr(w_{1a} + \delta - p_A > w_{1b} - p_B) + \frac{1}{3} \Pr(w_{2a} - p_A > w_{2b} + \delta - p_B) \\ D_B(p_A, p_B) &= 1 - D_A(p_A, p_B) \end{aligned} \quad (5)$$

Note that firms has different demand correspondents under the common and personal rankings due to  $w_{2a}, w_{2b}$  has different distributions under the common and personal rankings. Each demand correspondence are probabilities that one exponential variable with given parameter is lower than other exponential variable with other given parameter, hence it can be expressed as a probability distribution function of a random variable that follows the Laplace distribution.

As Quint (2014) showed, due to distribution of  $w$ 's is log-concave, there exists a unique equilibrium, which is in pure strategies, in the pricing game among the sellers. The price

equilibrium is determined by standard FOC conditions. In this setting FOC condition are transcendental equation and can not be solved in the closed form, so I provide the results of numerical solution on the [Figure 1](#).

As we see, depending on the value  $\delta$  of the level of products vertical differentiation, firms might charge higher or lower prices in the case of personalized ranking comparative to the case of the common ranking. This might be explained by the fact that the transition from the common ranking to the personalized ranking involves two changes in the firm's pricing incentives, summarized by the following two effects. The first, assorting, effect provides incentives to both firms to increase prices. In the case of the personal ranking, comparative to the common ranking case, consumers in average find well-suited product on the first position, which lowers their incentives to search further. This leads to decrease in the competition between firms and as a result, both firms have incentive to increase prices regardless of their position in the common ranking. The second, advertising, effect affects firms pricing decisions heterogeneously depending on their ranking position in the common ranking. As [Armstrong \(2017\)](#) shows, when prices are observed prior to search they can be used to influence a consumer's search order. Firm A which is shown on the first positions under the common ranking has zero search cost and do not need to keep prices low to attract consumers to explore its product. Under the personal ranking firm A is shown on the second positions for a third of consumers, which provides incentives to decrease the price. Firm B which is shown on the second positions under the common ranking needs to keep its prices low, otherwise consumers will not explore its product due to search costs. Under the personal ranking firm B is shown to a third of consumers on the first positions. As a result, it has lower incentive to keep prices low under the personal ranking. As the level of products vertical differentiation increases, the advertising effect becomes less important since consumers have stronger preferences toward one of the products. Hence as  $\delta$  increases firm A has more incentives to increase the price. For firm B both effects provides incentive to increase the price for any level of  $\delta$ , but for very low  $\delta$  firm B in equilibrium decreases price in response to dramatic decrease in price of product A.

This example illustrates that the permutation of product positions in ranking alone is enough to change the market outcomes.

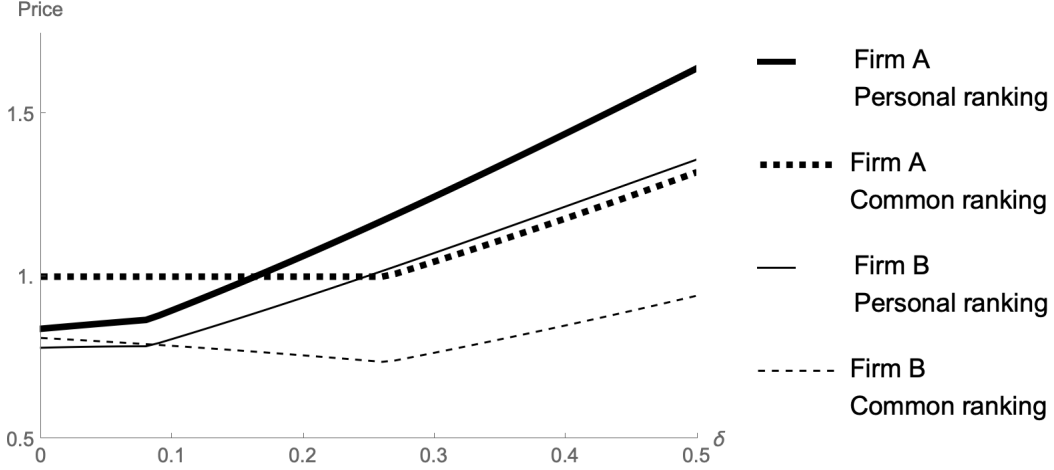
## Empirical Model

### 2.1 Modeling of the platform's information

In this section, I provide the main understanding of how I model the information about consumers' preferences the platform observes and might use to provide a ranking of the product under different ranking paradigms – the common ranking, the personal ranking, and the random ranking.



Figure 1: Prices as functions of  $\delta$ .



Note: This figure illustrates firms' prices in the case of personalized ranking and common ranking for the market settings, discussed in section 1.1. For high level of vertical product differentiation ( $\delta$ ) both firms has incentives to charge higher prices under the personalized ranking. For low  $\delta$  the firm A which is prominent in the common ranking case has incentives to charge lower price under the personalized ranking to attract consumers to explore its product.

By analogy with the example from the previous section, the platform observes  $\delta$ 's, the part of utility observed by the consumer prior to the search.  $\delta$  is a convolution of objective product characteristics weighted on consumer's sensitivity to them. More precisely, product's  $j$  utility that consumer  $i$  observes prior to search is

$$\delta_{ij} = \alpha_i p_j + \beta_i' \mathbf{x}_j,$$

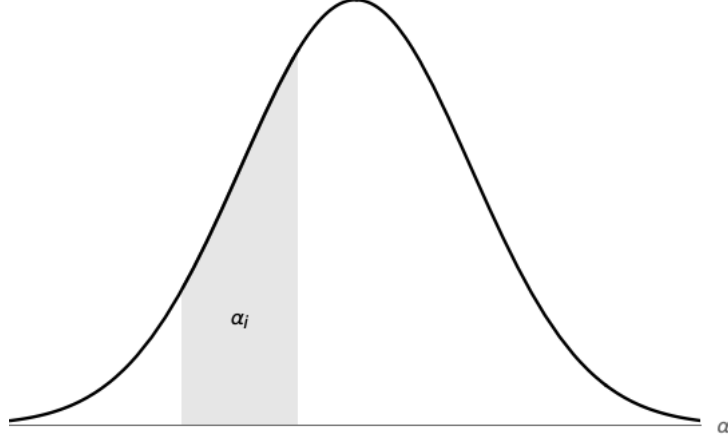
where  $p_j$  and  $\mathbf{x}_j$  are price and the vector of objective product's characteristics observed prior to exploring the product's page. In the case of hotels,  $\mathbf{x}_j$  might contain such characteristics as hotel star rating, review score, chain identity, location and etc.  $\alpha_i$  and  $\beta_i$  describe consumer's sensitivity to price and mentioned characteristics.

In general, two different consumers value differently the same objective properties of the product. In the case of hotels, different consumers might, for example, have different favorite hotel chains and have different sensitivity to the price of the hotel room. As a result, different consumers have different  $\alpha$ 's and  $\beta$ 's, labeled as  $\alpha_i$  and  $\beta_i$ , showing their affiliation to consumer  $i$ . The set of  $\alpha_i$ 's and  $\beta_i$ 's of all consumers on the market form distributions with means  $\bar{\alpha}$  and  $\bar{\beta}$  and variances  $\sigma_\alpha$  and  $\Sigma_\beta$ .

By saying that the platform knows personal consumer preferences I assume that the platform knows some information about individual  $\alpha_i$ 's and  $\beta_i$ 's. In the extreme case, the platform knows the actual values of  $\alpha_i$  and  $\beta_i$  for each given consumer. In a more realistic scenario, illustrated on Figure 2, the platform knows in what part of a distribution bell  $\alpha_i$  and  $\beta_i$  are positioned. Both scenarios allow estimating  $\delta_{ij}$  for each given consumer, which is different from the mean among all population's  $\delta_j$ .

If the platform is allowed to use this information about personal preferences to form rankings, the platform can rank the products to each given consumer  $i$ , placing products with higher  $\delta_{ij}$ 's on higher positions, what, as we saw in the previous section, leads to higher market prices. If the platform, on the contrary, is not allowed to use the information on the personal preferences, then it has to use only information on aggregated preferences,  $\bar{\alpha}$  and  $\bar{\beta}$ , which lead to identical ranking to all consumers.

Figure 2: Example of the platform's information



## 2.2 Demand side

The response to each consumer's query contains  $J$  different hotels (indexed by  $j = 1, 2, \dots, J$ ). The utility consumer  $i$  derives from hotel  $j$  is given by:

$$u_{ij} = \alpha_i + \beta_i' \mathbf{x}_j + \epsilon_{ij}, \quad (6)$$

where  $\alpha_i$  is a consumer-specific price coefficient, the variable  $p_j$  denotes the price of hotel  $j$  and the vector  $(\mathbf{x}_j, \epsilon_{ij})$  describes different hotel attributes from which the consumer derives utility. As usual,  $\mathbf{x}_j$  includes a 1 in order to allow for a constant term in the utility function. I assume that the consumer observes the hotel attributes contained in  $\mathbf{x}_j$  without searching. The variable  $\epsilon_{ij}$ , which is assumed to be independently and identically distributed across consumers and hotels, is a match parameter and measures the "fit" between consumer  $i$  and hotel  $j$ . Each  $\epsilon_{ij}$  is a draw from Gumbel distribution with location and scale parameters 0 and 1 (TIEV), as is common in choice models. I assume that  $\epsilon_{ij}$  captures "search-like" hotel attributes, that is, characteristics that can only be ascertained after visiting the hotel page. I assume that the econometrician observes the hotel attributes contained in  $\mathbf{x}_j$  but cannot observe those in  $\epsilon_{ij}$ .

Consumers differ in the way they value price and hotel characteristics. Parameters  $\alpha_i$  and  $\beta_i$  capture consumer heterogeneity in tastes for price and hotel attributes. These

parameters are assumed to follow the multivariate normal distribution, i.e.

$$\begin{bmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{bmatrix} = N \left( \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \sigma_\alpha & 0 \\ 0 & \boldsymbol{\Sigma}_\beta \end{bmatrix} \right), \quad (7)$$

where  $\boldsymbol{\Sigma}_\beta$  is a diagonal matrix, i.e. I assume that consumer's demand elasticities are independent.

I assume consumers must search to find out the exact utility they derive from each of the hotels available. To be more specific, I assume that before searching a consumer  $i$  knows (i) hotel characteristics  $p_j$  and  $\mathbf{x}_j$  for each hotel  $j$ , (ii) the distribution  $F(\epsilon)$  of match values  $\epsilon_{ij}$ , and (iii) the utility of his outside option  $\epsilon_{i0}$ . I thus regard the search of a consumer  $i$  as a process by which he discovers the exact values of the matching parameter  $\epsilon_{ij}$  upon visiting  $j$  hotel's page.

Consumers search sequentially with costless recall, i.e., they determine after each visit to a hotel's page whether to book any of the inspected hotels so far, to opt for the outside option, or to continue searching. Let  $s_{n_{ij}}$  denote the search cost of consumer  $i$  for visiting page of the hotel  $j$ , where  $n_{ij}$  is the position of the hotel  $j$  in the list of hotels shown to the consumer  $i$  by the platform. In section 2.2.2 I discuss why the cost of exploring the hotels depends on its position in the rank rather than the identity of the hotel.

### 2.2.1 Optimal consumer sequential search

The utility function in Equation 6 can be rewritten as

$$u_{ij} = \delta_{ij} + \epsilon_{ij}, \quad (8)$$

where  $\delta_{ij}$  is the mean utility consumer  $i$  derives from hotel  $j$  and  $\epsilon_{ij}$  is TIEV random shock. As explained above, the consumer knows  $\delta_{ij}$ , but has to search to discover  $\epsilon_{ij}$ . The match values  $\epsilon_{ij}$  follow TIEV distribution, which is the same for all consumers and hotels, and is given by  $F(\epsilon)$  with pdf  $f(\epsilon)$ .

Since I allow for consumer-specific taste parameters, the distribution of consumer  $i$ 's utility  $u_{ij}$  from a given hotel  $j$  differs across consumers. This leads to the usual aggregation problem I need to deal with. Since the utility shock  $\epsilon_{ij}$  is an iid draw from TIEV distribution, the utility distribution for hotel  $j$  faced by consumer  $i$  is

$$F_{ij}(u) = F(u - \delta_{ij}), \quad (9)$$

that is, the distribution of  $u_{ij}$  is Gumbel location parameter  $\delta_{ij}$  and scale 1.

Following Weitzman (1979), I define  $H_{ij}(r)$ , the expected gains to consumer  $i$  from

exploring the hotel  $j$  when the best utility the consumer has found so far is  $r$ :

$$H_{ij}(r) = \int_r^\infty (u - r) dF_{ij}(u) \quad (10)$$

If consumer  $i$ 's expected gains are higher than the cost  $s_{n_{ij}}$  he has to incur to explore the hotel  $j$ , then he should explore the hotel  $j$ . Correspondingly, I define the reservation value  $r_{ij}$  as the solution to equation

$$H_{ij}(r_{ij}) = s_{n_{ij}} \quad (11)$$

Notice that  $H_{ij}$  is strictly decreasing so Equation 11 has a unique solution. Therefore  $H_{ij}$  is an invertible function.

$$r_{ij} = H_{ij}^{-1}(s_{n_{ij}}) \quad (12)$$

Note that  $r_{ij}$  is a scalar, and that for each consumer  $i$  there is one such scalar for every hotel  $j$ . Moraga-González et al. (2018) shows that the reservation value can be decomposed into a mean utility component and a search cost component:

$$r_{ij} = \delta_{ij} + H_0^{-1}(s_{n_{ij}}), \quad (13)$$

where

$$H_0(r) \stackrel{\text{def}}{=} \int_r^\infty (u - r) dF(u) = \gamma - r + \int_{e^{-r}}^\infty \frac{e^{-t}}{t} dt, \quad (14)$$

where in the last equation the fact that  $\epsilon_{ij}$  is TIEV random variable is used.  $\gamma$  here is the Euler constant.

Weitzman (1979) demonstrates that the optimal search strategy for a consumer  $i$  consists of visiting sellers in descending order of reservation values  $r_{ij}$  and stopping search as soon as the best option encountered so far (which includes the outside option) gives a higher utility than the reservation value of the next option to be searched. This optimal search strategy can be characterized by following search rules:

1. **Selection rule.** If a search is to be made, the option with the highest reservation utility should be searched next.
2. **Stopping rule.** Search should terminate when the maximum utility observed exceeds the reservation utility of any unsearched option, i.e.

**2.1** If the consumer makes a search, then her reservation utility from that hotel must exceed her utility from all hotels searched so far.

**2.2** All unsearched hotels must have a lower reservation utility than the maximum utility of the searched ones.

**3. Choice rule.** Once the consumer stops searching, she will choose the option with the highest utility among those searched.

The rules **2.2** and **3** rely only on the information what hotels consumer explored and which one finally booked, while rules **1** and **2.1** requires the data of the order in which consumers explores alternatives. The dataset provided by Expedia does not include information on order in which the consumer visits hotels' pages. [Jeziorski and Segal \(2015\)](#) showed that users click ads in a nonsequential order which makes it unreasonable to assume any given order of search (e.g. assume that consumers search in the order of ranking positions). Given the fact that some consumers explore up to 25 hotels, the number of possible search orders for these consumers is  $25! \approx 10^{25}$ , which makes it computationally impossible to model the search order. To address this challenge, I adapt recent findings from the theoretical search literature by [Armstrong \(2017\)](#) and [Choi et al. \(2018\)](#) and its application by [Moraga-González et al. \(2018\)](#) that make it possible to compute the buying probability of a given alternative without having to go explicitly through the myriad of possible ways in which a consumer may end up considering the alternative in question.

### 2.2.2 The effect of ranking

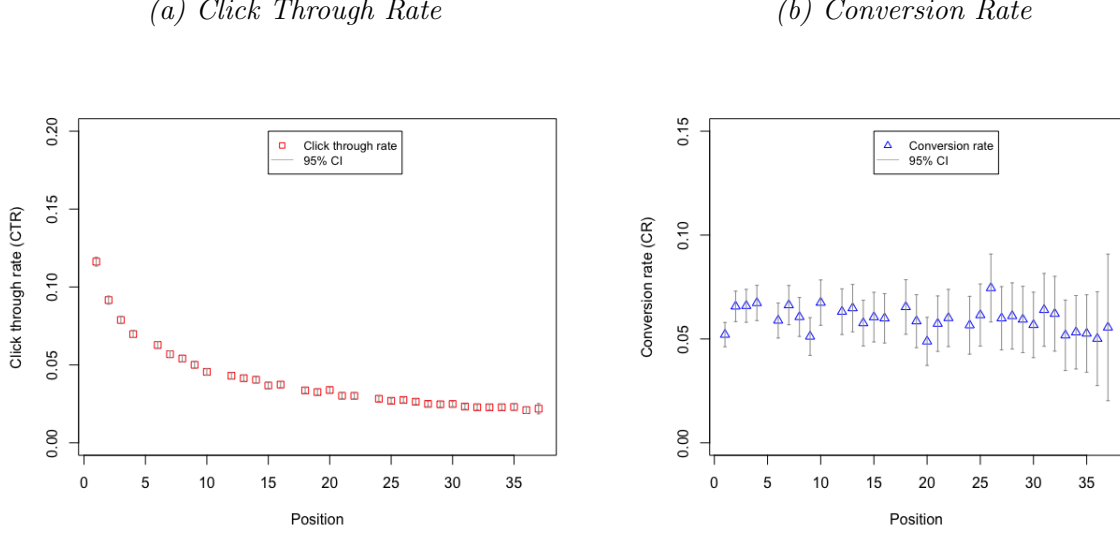
As discussed in [section 3](#), dataset provided by expedia contains impressions where the hotels were sorted randomly, which allows to separate the effect of hotels' positioning on the consumer behavior from the effect of hotels' attractiveness. [Figure 3a](#) shows that the probability consumer explores the hotel depends on the position of the hotel on the page. Hence hotel's position in the ranking affects consumers' search and purchase decisions.

Given consumer's optimal search strategy, described in the [section 2.2.1](#), the effect of the ranking on consumers' choice can be rationalized only in one of the following situations. The ranking affects either consumers' search behavior by affecting reservation utilities  $r_{ij}$  associated with the hotels. Or it affects consumers' purchasing behavior through affecting the actual utilities  $u_{ij}$  consumers derive from booking the hotels. According to [Equation 8](#) and [Equation 13](#), there are three potential ways how the ranking can affect the reservation or actual utilities – by affecting the utility prior to search ( $\delta_{ij}$ ), the portion of utility realized after the search ( $\epsilon_{ij}$ ), and the search cost ( $s_{n_{ij}}$ ).

[Ursu \(2018\)](#) showed using the dataset discussing in this study, that the rank of a hotel in the list provided to consumer's query has the effect only on the search cost associated with the hotel and does not have any effects on  $\delta_{ij}$  and  $\epsilon_{ij}$ . Therefore, the ranking affects the reservation utility and, in turn, the optimal searching and purchasing decisions only through an effect on the search cost of the displayed hotel, which is the model used in this paper. Ursu's arguments are mostly based on the observation that probability the

consumer books the hotel, conditional on exploring it, does not depend on the position of the hotel, as shown on Figure 3b. She makes the conclusion that the position of the does not have any effect on how the consumer values the hotel and only affects the probability the hotel is in the consumer’s consideration set.

Figure 3: Hotels’ Click Through and Conversion rates. Randomly sorted queries.



*Note: These figures illustrate the Click-through rate and the conversion rate (the purchase rate conditional on click) over positions for the case when the lists of hotels, presented to consumers were formed randomly without accounting to the utility provided by hotels. The right panel shows that the conversion rate does not depend on the position itself, which is an argument to the fact the position the hotel is presented does not affect consumers’ utility. The left panel shows that Click-through rate is decreasing over positions which is an argument to the fact that position of the hotel affects consumer’s search behavior.*

### 2.2.3 Probabilities of purchase

For each consumer  $i$  and hotel  $j$  define a random variable

$$w_{ij} \stackrel{\text{def}}{=} \min\{u_{ij}, r_{ij}\} = \delta_{ij} + \min\{\epsilon_{ij}, H_0^{-1}(s_{n_{ij}})\}.$$

Choi et al. (2018) showed that if the consumer conducts a sequential search, he purchases product  $i$  with the highest value of  $w_{ij}$  among all products. It suggests that each consumer’s eventual purchase decision can be represented as in canonical discrete-choice models. The only difference is that consumers’ purchase decisions are made based, neither on true values  $u_{ij}$  nor on reservation values  $r_{ij}$  but on newly identified values  $w_{ij}$ , which is referred to as *effective values*. Clearly,  $w_{ij}$  is related to underlying values  $\delta_{ij}$  and  $u_{ij}$ . In particular,  $w_{ij}$  converges to  $u_{ij}$  as  $s_{n_{ij}}$  tends to 0 (in which case  $H_0^{-1}(s_{n_{ij}})$  approaches to  $\infty$ ) and is determined only by  $H_0^{-1}(s_{n_{ij}})$  as  $s_{n_{ij}}$  tends to infinity (in which

case  $H_0^{-1}(s_{n_{ij}})$  approaches to  $-\infty$ ). Intuitively, if there are no search costs, each consumer makes a fully informed decision and purchases the product that offers the largest net utility (i.e.,  $w_{ij} = u_{ij} \forall i$ ). To the contrary, if search costs grow arbitrarily large, then consumers' purchase decisions depend only on values observed prior to search. In general, search frictions make consumers' match values  $\epsilon_{ij}$  imperfectly reflected in their purchase decisions. The problem becomes more severe, and consumers rely less on  $\epsilon_{ij}$ , as search frictions increase.

Accordingly, the probability that buyer  $i$  books hotel  $j$  can be expressed as:

$$P_{ij} = Pr(w_{ij} \geq \max_{k \neq j} w_{ik}) = \int \left( \prod_{k \neq j} F_{ik}^w(x) \right) f_{ik}^w(x) dx \quad (15)$$

The distribution of  $w_{ij} = \min\{u_{ij}, r_{ij}\}$  can be obtained by computing the CDF of the minimum of two independent random variables. This means that

$$F_{ij}^w(x) = 1 - (1 - F_{ij}^r(x))(1 - F_{ij}^u(x)) \quad (16)$$

where  $F_{ij}^w$  and  $F_{ij}^r$  are the CDF's of  $w_{ij}$  and  $r_{ij}$ , respectively. Recall that  $F_{ij}^u(x)$  is the CDF of  $u_{ij}$ , which has been specified above in [Equation 9](#).

To obtain the distribution of the reservation values, I use [Equation 12](#).

$$F_{ij}^r(x) = Pr(r_{ij} < x) = Pr(H_{ij}(r_{ij}) > H_{ij}(x)) = Pr(s_{ij} > H_{ij}(x)) = 1 - F_{ij}^s(H_{ij}(x))$$

Substituting this into [Equation 16](#) gives

$$F_{ij}^w(x) = 1 - F_{ij}^s(H_{ij}(x))(1 - F_{ij}^u(x)) \quad (17)$$

[Equation 17](#) provides a relationship between the search cost distribution and the distribution of the  $w$ 's. Assuming the right search costs distribution, any needed distribution of  $w$ 's can be obtained. [Moraga-González et al. \(2018\)](#) shows that if

$$F_{ij}^s = \frac{1 - \exp(-\exp(-H_0^{-1}(s) - \mu_{ij}))}{1 - \exp(-\exp(-H_0^{-1}(s)))}, \quad (18)$$

where  $\mu_{ij}$  is a consumer-hotel specific parameter of the search cost distribution, then CDF of  $w_{ij}$  is given by Gumbel distribution:

$$F_{ij}^w(x) = \exp(-\exp(-(x - (\delta_{ij} - \mu_{ij})))) \quad (19)$$

Given Equation 19,  $P_{ij}$  in Equation 15 has a closed form:

$$P_{ij} = \frac{\exp(\delta_{ij} - \mu_{ij})}{1 + \sum_{k \in J} \exp(\delta_{ik} - \mu_{ik})} \quad (20)$$

Finally, the unconditional choice probability can be obtained from  $P_{ij}$  in Equation 20 by integrating out the consumer-specific variables. Denoting by  $\theta_i$  the vector of all consumer-specific random variables in  $P_{ij}$ , the probability that hotel  $j$  is booked is the integral

$$P_j = \int P_{ij} dF^\theta(\theta_i) = \int \frac{\exp(\delta_{ij} - \mu_{ij})}{1 + \sum_{k \in J} \exp(\delta_{ik} - \mu_{ik})} dF^\theta(\theta_i) \quad (21)$$

As discussed in Section 2.2.2, consumer-hotel specific parameter of the search cost distribution  $\mu_{ij}$  depends not on the identity of the hotel, but its position in the ranking. I model  $\mu_{ij}$  as  $\mu_{ij} = \log(1 + e^{\gamma n_{ij}})$ , where  $n_{ij}$  is the position of the hotel  $j$  in ranking shown to the consumer  $i$ .

The integral in Equation 21 does not have a closed-form solution. Thus, I replace  $P_j(\theta)$  with the simulated choice probability  $P_j^\wedge(\theta)$ .

## 2.3 Supply side

Each hotel  $j$  sets the price  $p_{jt}$  for a given hotel room at a given night  $t$  to maximize the expected profit of such sale, conditional on the prices and characteristics of rivals and the opportunity cost  $c_{jt}$  and the hotel-specific ad-valorem fee  $f_j$  charged by the platform.

$$\Pi_{jt} = ((1 - f_j)p_{jt} - c_{jt})D_{jt}(p_{jt}) \quad (22)$$

The expected demand of hotel  $j$  can be expressed as

$$D_{jt}(p_{jt}) = \int P(buy|\theta)(p_{jt}) dF^\theta(\theta) \quad (23)$$

where  $P(buy|\theta)(p_{jt})$  is a probability that consumer with demand parameter  $\theta$  purchases the product of the firm  $j$ . This probability depends on the position of the hotel in the hotel ranking shown to the consumer. Equation 23 can be rewritten as

$$D_{jt}(p_{jt}) = \int \left( \sum_{positions} P(buy|\theta, position)(p_{jt}) \cdot \mathbb{1}(position|\theta)(p_{jt}) \right) dF^\theta(\theta), \quad (24)$$

where  $\mathbb{1}(position|\theta)(p_{jt})$  is an indicator function of the hotel  $j$  be shown on the position



$position$  in  $i$ 's consumer ranking and can be expressed as:

$$\mathbb{1}(position)(p_{jt}) = \begin{cases} 1 & \text{if } \delta_j = \delta^{(position)} \\ 0 & \text{if } \delta_j \neq \delta^{(position)} \end{cases}$$

where  $\delta^{(position)}$  is a  $position$  order statistic of  $\delta$ 's, shown to the consumer i.e.  $position$  largest  $delta$  among  $delta$ 's of hotels in the query response.

Since the platform tends to put better-fitted hotels on higher positions, the probability that the hotel  $j$  is shown on the given position depends on the utility the consumer  $i$  derives from booking this hotel. As a result, if the hotel increases the price of room, there are two effect on its demand. First, it decreases chances of the hotel to be shown on high position, and second, for any position it decreases the probability the hotel is booked, as described in Equation 25.

$$\begin{aligned} \frac{\partial D_{jt}(p_{jt})}{\partial p_{jt}} = & \int \left( \sum_{positions} \frac{\partial P(buy|\theta, position)(p_{jt})}{\partial p_{jt}} \cdot \mathbb{1}(position|\theta)(p_{jt}) + \right. \\ & \left. + \sum_{positions} P(buy|\theta, position)(p_{jt}) \cdot \frac{\partial \mathbb{1}(position|\theta)(p_{jt})}{\partial p_{jt}} \right) dF^\theta(\theta) \end{aligned} \quad (25)$$

Profit maximizing hotel  $j$  sets the price according to the following equation:

$$p_{jt}^* = \frac{c_{jt}}{1 - f_j} - \frac{D_j(p_{jt})}{\frac{\partial D_j(p_{jt})}{\partial p_{jt}}} \quad \forall j \quad (26)$$

## Data

For kaggle contest, Expedia has provided a [dataset](#) that includes searching and purchase data as well as information on price competitiveness. The data are organized around a set of "search result impressions", or the ordered list of hotels and their characteristics that the user sees after they search for a hotel on the Expedia website. In addition to impressions from the existing algorithm, the data contain impressions where the hotels were randomly sorted, to avoid the position bias of the existing algorithm. The user response is provided as a click on a hotel and a purchase of a hotel room.

The summary statistics are provided in Table 4. To describe the observables in this data set, in this paragraph, I explain the three-step consumer search process on Expedia. First, the consumer begins her search query on Expedia by specifying details of her trip, such as the destination (city, country), the travel dates, and the number of travelers and rooms requested. In addition, based on her query, the number of days before the beginning of the trip is recorded (booking window). The data set includes information on all these variables. Second, in response to her query, the consumer gets a search

impression of all of the hotels that match her request, distributed over multiple pages. From this search impression, I observe the first page of results displayed to consumers (the "list page"), which includes the hotel ID, its position in the ranking, and its characteristics (price, number of stars and reviews, location, a chain, and a promotion indicator). On Expedia, consumers can sort or filter results; however, the data set only contains those search impressions where consumers made choices from the ranking displayed. Third, after observing the list of hotels, the consumer can click on a particular one to observe more information. In this case, she navigates to a sub-page reserved for that hotel (the "hotel page"), where she can see additional pictures, previous customers' reviews, and so on. Then she can either return to the previous screen to click on another hotel, leave the site without purchasing, or she can purchase. I observe all clicks and purchases consumers make. However, I do not observe the additional information consumers see on the hotel's page. Also, observations are provided at the query level, which means searches made by the same consumer cannot be linked.

*Table 4: Hotel and Query Summary Statistics*

	Observations	Mean	Median	SD	Min	Max
Hotel level						
Price	5,511,851	156.49	129.00	101.28	10	1000
Stars	5,383,647	3.31	3.00	0.88	1	5
Review Score	5,505,786	3.86	4.00	0.91	0	5
Chain	5,511,851	0.65	1.00	0.48	0	1
Location Score	5,511,851	3.09	3.00	1.52	0	7
Promotion	5,511,851	0.24	0.00	0.43	0	1
Query level						
Number of hotels displayed	206,657	27.12	31.00	8.10	5	38
Trip length (days)	206,657	2.42	2.00	1.98	1	40
Booking window (days)	206,657	39.26	18.00	53.89	0	498
Saturday night (percent)	206,657	0.50	1.00	0.50	0	1
Adults	206,657	2.00	2.00	0.90	1	9
Children	206,657	0.39	0.00	0.79	0	9
Rooms	206,657	1.12	1.00	0.44	1	8
Total clicks	206,657	1.12	1.00	0.61	1	25
Two or more clicks (percent)	206,657	0.07	0.00	0.25	0	1
Transaction	206,657	0.66	1.00	0.48	0	1
Random ranking (percent)	206,657	0.31	0.00	0.46	0	1

# Estimation

## 4.1 Demand side

### 4.1.1 Estimation strategy

The probability that a random consumer purchase the product of firm  $j$  was provided in section 2.2.3 in Equation 21 as  $P_j(\theta)$ , where  $\theta = (\alpha, \sigma_\alpha, \beta, \Sigma_\beta, \gamma)$  is a set of population distribution parameters.

Hence, the log-likelihood function can be obtained as:

$$LL(\theta) = \sum_i \sum_j d_{ij} \log(P_j(\theta)), \quad (27)$$

where  $d_{ij} = 1$  if the consumer  $i$  books the hotel  $j$  and zero otherwise. The integral in Equation 21 does not have a closed-form solution. Thus, I replace  $P_j$  with the simulated choice probability  $\hat{P}_j$ . This approach results in the following simulated log-likelihood:

$$SLL = \sum_i \sum_j d_{ij} \log(\hat{P}_j(\theta)). \quad (28)$$

I simulate  $\hat{P}_j(\theta)$  by drawing values of  $\theta$ , plugging these draws into  $P_{ij}$  and averaging over the resulting MNL probabilities. The MSL estimate of the parameter vector  $\hat{\theta}$  is a consistent estimator of  $\theta$  only if both the numbers of simulations and observations go to infinity. However, there are two (related) reasons why this consistency result should not interfere with inference based on  $\theta$ . First, Börsch-Supan and Hajivassiliou (1993) show that MSL provides accurate parameter estimates in polychotomous choice problems, even with a small number of simulations. Second, I use a number of simulations (10,000) that is not considered small for this type of problem.

### 4.1.2 Monte Carlo Simulations

In this section, I describe simulation results to show that the estimation strategy described in subsection 2.2.3 can be used to recover utility and search cost parameters in this model. To this end, I generate a data set of 1,000 consumers, each searching among 30 hotels. Hotel characteristics (Quality and Price) are assumed to be drawn from a multivariate log-normal distribution. The simulation results are given in Table 5. The first column shows the true parameters and the second column shows the estimated parameters. Based on the results we can conclude that provided estimation method is effective in recovering of true demand parameters, In the next section I apply the method to real data, provided by Expedia to estimate the utility and search parameters of consumers participated in the hotels search and booking.

Table 5: Monte Carlo Simulation Results

	True values	Estimated values
Utility		
Price	-1	-0.9608 (0.0525)
Price heterogeneity	0.3	0.2849 (0.0430)
Quality	2	1.8941 (0.0883)
Quality heterogeneity	0.6	0.5335 (0.0542)
Search cost		
Position	0.1	0.0856 (0.0065)

### 4.1.3 Empirical results

I apply the estimation strategy, derived in section 2.2.3 to estimate consumers demand, using the data, provided by Expedia. The results of estimation are provided in Table 6. The results show

Search cost is significant. It has important implications for consumer search behavior. Hotels that appear lower in the ranking of slots have lower chances of being searched. Placing hotels with high expected utility levels in more prominent positions may reduce the cost of each search. Hence, the existence of search cost makes ranking especially beneficial for consumers.

Consumers demonstrate considerable heterogeneity in their sensitivities for hotels attributes, especially in the hotel location and chain affiliation. As a result, using the personalized ranking might have a big impact on consumers search and utility. consumers may use alternative refinement methods that prioritize more important attributes. I will further explore the effect of heterogeneity on the market structure in the policy simulation section.

## 4.2 Supply side

### 4.2.1 Estimation strategy

The point of interest is hotels' opportunity costs  $c_{jt}$ , which vary among hotels and queries, and hotel-specific fees  $f_j$  charged by the platform which vary among hotels only. For simulation purpose it is not necessary to estimate both the cost and fees, but only the ratio  $\frac{c_{jt}}{1-f_j}$  because, as described in Equation 26, hotels set prices conditional on this ratio.

Under the existing Expedia algorithm, the position of the hotel does not depend on

Table 6: Demand estimation results

	Parameter	Standard Deviation
Utility		
Constant	4.1167*** (0.1742)	6.2731*** (1.152)
Price (\$100)	-2.4881*** (0.1203)	0.6499*** (0.1259)
Star rating	1.2369*** (0.0705)	0.0166 (0.1703)
Review score	0.1118 (0.0895)	0.0916 (0.171)
Location score	0.0737 (0.1006)	0.9509*** (0.0726)
Chain dummy	0.7346*** (0.2656)	1.6514*** (0.4475)
Search cost		
Position	0.0625*** (0.0032)	— —

Note: Stars indicate estimates significant at the 99% level.

the consumer's characteristics, hence the Equation 24 can be rewritten as

$$D_{jt}(p_{jt}) = \sum_{positions} \left( \int \left[ P(buy|\theta, position)(p_{jt}) dF^\theta(\theta) \right] \cdot \mathbb{1}(position)(p_{jt}) \right) \quad (29)$$

Choi et al. (2018) shows that as a result of optimal search and purchase decisions, rational consumer purchases the product with the highest  $w_{ij} - p_j$ , where  $w_{ij}$  is defined in Equation 1. Hence  $P(buy|\theta, position)(p_{jt})$  in Equation 29 can be expressed as

$$\begin{aligned} P(buy|\theta, position)(p_{jt}) &= Pr(w_{ij} \geq \max_{k \in J_i} w_{ik} | \theta) = \\ &= Pr \left( \delta_{ij} + \min(\epsilon_{ij}, H_0^{-1}(s_{n_j})) \geq \max_{k \in J_i} [\delta_{ik} + \min(\epsilon_{ik}, H_0^{-1}(s_{n_k}))] | \theta \right) = \\ &= \frac{\exp(\delta_{ij}(\theta) - \mu_{ij})}{1 + \sum_{k \in J_i} \exp(\delta_{ik}(\theta) - \mu_{ik})} \end{aligned} \quad (30)$$

Using the consumers demand characteristics estimates from section 4.1.3, Equation 30 can be expressed as a function of hotel's  $j$  price  $p_j$ . Under the assumption that the hotel knows on what position it is going to be shown conditional on price, the demand, described in Equation 29 can be expressed as a function of hotel's  $j$  price.

Finally, the first order condition, provided in Equation 26 can be used to estimate the

parameter  $\frac{c_{jt}}{1-f_j}$ .

$$p_{jt}^* = \frac{c_{jt}}{1-f_j} - \frac{D_j(p_{jt})}{\frac{\partial D_j(p_{jt})}{\partial p_{jt}}} \quad \forall j$$

These parameters are used later in section [Counterfactual simulations](#) to run simulations for different data allowance policies.

#### 4.2.2 Empirical results

The estimation strategy, described in the previous chapter allows to recover hotels' opportunity costs. The histogram of hotels' opportunity costs is represented on [Figure 4](#). I use this estimations in the next section to get counterfactuals results and estimate the change in hotels pricing under the personal and common rankings.

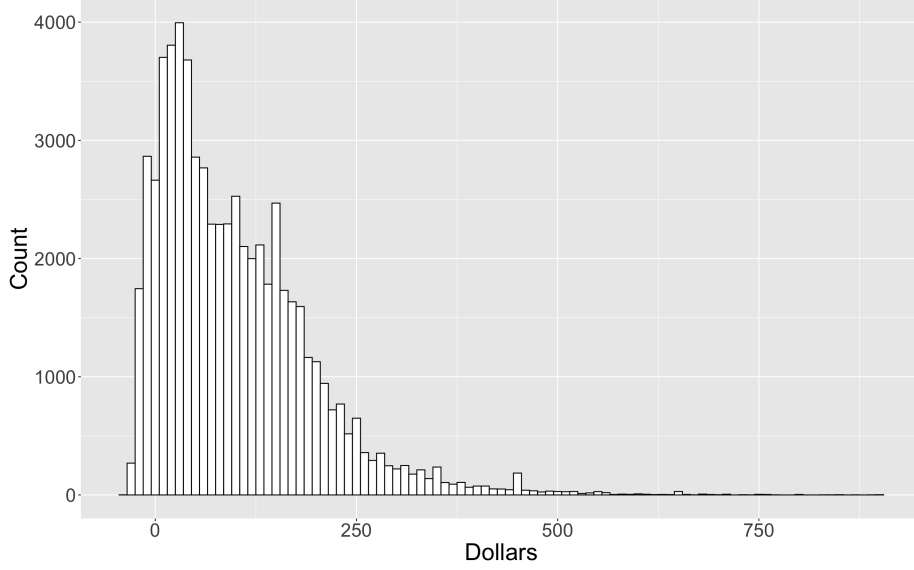
It's important to note that around 9% of the opportunity cost in the data is negative. This is possible because of the fundamental difference between the opportunity cost and the marginal cost. Cost estimated in this chapter captures the dynamic nature of the hotel's pricing problem and represents the cost of selling the room at the moment query was submitted. If the hotel expects that in the future the equilibrium price on the market is going to decrease, for example because of increase in competition, the opportunity cost of selling the room right now might be negative. Also, selling the room for a low price the hotel might expect the consumer write the positive review, which increases future hotel's competitiveness and might be considered as an investment.

## Counterfactual simulations

In this section I discuss the details of counterfactual simulations. Using the demand and firms' opportunity costs estimations, provided in sections [4.1.3](#) and [4.2.2](#) respectively, I simulate firms pricing decisions under two different data allowance policies and compare results. In the first one I allow the platform to use consumers personal data to provide the personal ranking to each consumer. In the second one the platform is allowed to use only aggregated data of all consumers and provide the same ranking to all consumers. In the first case consumers find better-suited hotels on higher positions, which affects consumers' search behavior and, as a result, hotels' demand function. This leads to different optimal prices under different raking mechanisms. [Figure 6](#) shows the histogram of the change in price each firm charges under the personal and common rankings.

In the case of the personal ranking, comparative to the common ranking case, consumers find better-suited hotels on higher positions, which lowers their incentives to search and decreases the average number of searched hotels. This effect leads to decrease in the competition between hotels and as a result, all hotels have incentive to increase the price regardless of their position in the common ranking. The second effect affects

Figure 4: Hotels' Opportunity Costs Histogram  $\left(\frac{c_{jt}}{1-f_j}\right)$



*Note: The figure provides the result of hotels' opportunity costs estimation. The opportunity cost of the hotel might be negative. If the hotel expects that in the future the equilibrium price on the market will decrease, for example because of increase in competition, the opportunity cost of selling the room right now might be negative.*

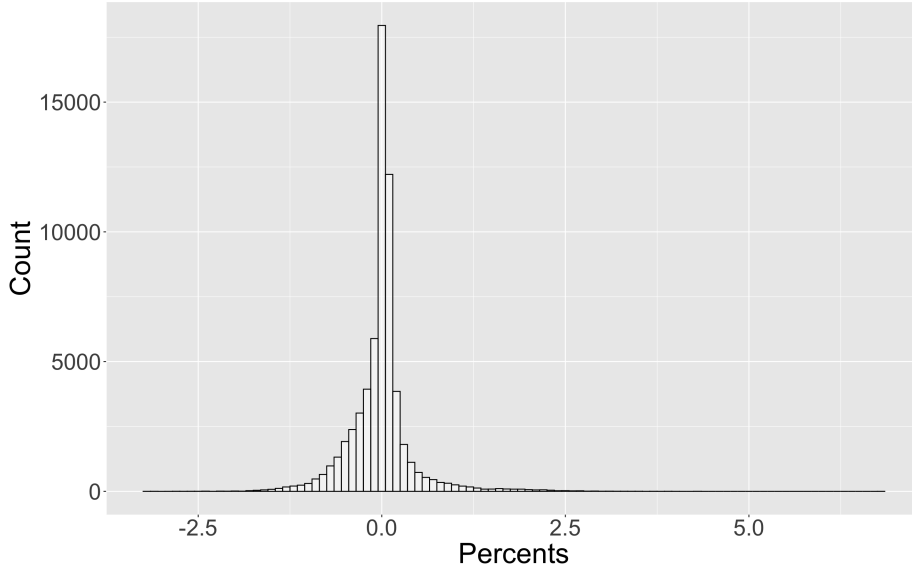
hotels pricing decisions heterogeneously depending on their ranking position in the common ranking. As (Armstrong, 2017) shows, when prices are observed prior to search they can be used to influence a consumer's search order. The hotels which are shown on high positions under the common ranking have low search cost and do not need to keep prices low to attract consumers to explore them. Under the personal ranking these hotels are shown on lower positions for some of consumers, which provides incentives to decrease the price. The hotels which are shown on low positions under the common ranking need to keep their prices low, otherwise consumers will not explore them due to their high search costs. Under the personal ranking these hotels are good-suited for some consumers and shown to them on the high positions. As a result, these hotels have lower incentive to keep prices low under the personal ranking.

Figure 6 and Figure 7 illustrate the heterogeneity of the sum of two effects over the positions of hotels in the common ranking. Figures show that hotels, which are on the higher positions in the common ranking has higher incentives to decrease the prices.

As discussed previously, all hotels have incentives to charge higher prices under the personal ranking due to consumers find better-fitted hotels on higher positions and explore fewer hotels, which lowers the competition between hotels. This effect increases with the level of hotels vertical differentiation. Figure 8 shows that if consumer observes higher variation of hotels utilities in the query, the first effect has bigger magnitude and the hotels have higher incentives to increase the price.

The change of the ranking mechanism has two effects on consumer utility. In addition

Figure 5: Percentage Price Change. Personal vs Common rankings



*Note: The figure provides the histogram of the percentage difference in hotel prices between the case of the personalized ranking and the common ranking. The price change can be explained by the combination of two effects: the assorting one and the advertising one. The assorting effect provides incentives to all hotels to increase the price due to decreased competition among hotels caused by the reduction of the consumers' search intensity. The advertising effect is heterogeneous by the position of the hotel in the common ranking. Hotels which are prominent under the common ranking have to decrease prices in the case of personalized ranking to attract consumers to explore their rooms. Hotels which were showed on the low positions in the common ranking becomes more prominent for some consumers under the personalized ranking and hence have lower incentives to keep low prices.*

to price change, discussed above, consumer finds better-fitted hotels on higher positions, which leads to reduction in search expenditures. The first effect is summarized on [Figure 9](#). In average, due to the price increase, consumers loose \$4, or 3% of his utility if the platform applies the personal ranking, comparative to the common one. More sensitive to price consumers loose more and less sensitive ones loose less utility as illustrated on [Figure 9b](#).

The second effect is represented on [Figure 10](#), which shows that the average position of the booked hotels decreases under the personal ranking. In average, consumers save \$1 of search expenditures if the platform applies the personal ranking, comparative to the common one.



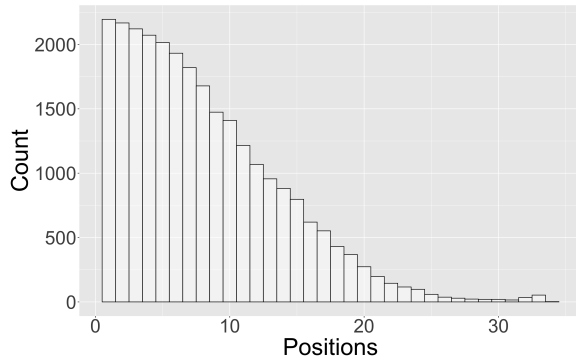
Figure 6: Percentage Price Change by position in the common ranking



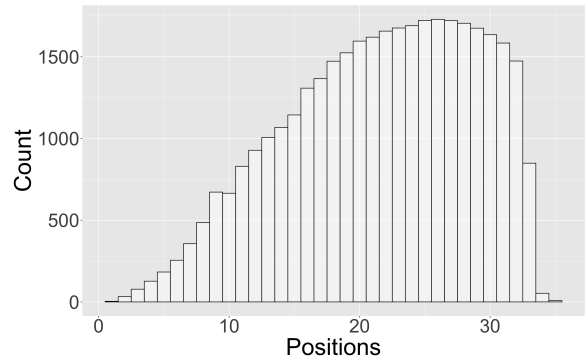
Note: The figure illustrates the percentage difference in hotel prices by positions between the case of the personalized ranking and the common ranking. The advertising affect is higher for the hotels which were more prominent under the common ranking, causing them to decrease prices more dramatic. The assorting effect causes all hotels to increase the price. The figure illustrates that the assorting effect has larger magnitude for less prominent hotels, while the advertising one has larger magnitude for more prominent hotels.

Figure 7: Positions of the hotels which increase and decrease prices respectively if the platform applies the personal ranking

(a) Positions of hotels which charges lower prices under the personal ranking

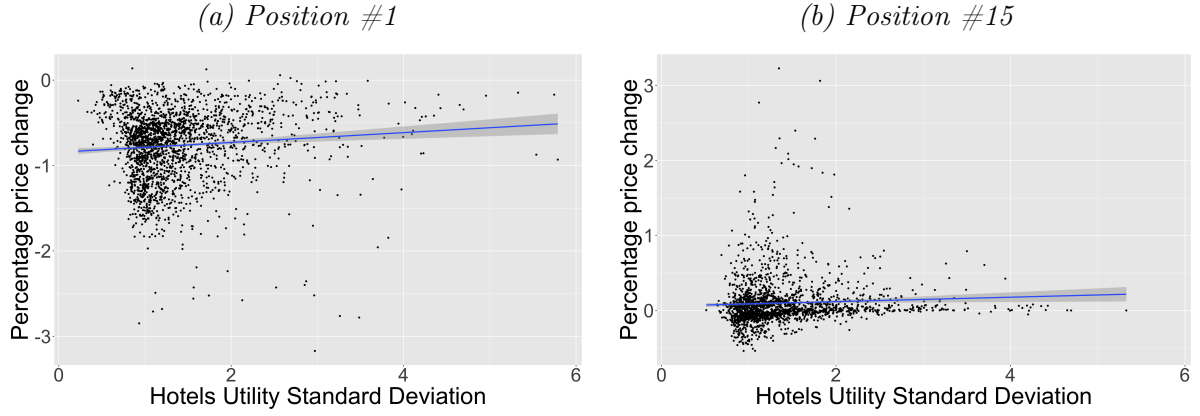


(b) Positions of hotels which charges higher prices under the personal ranking



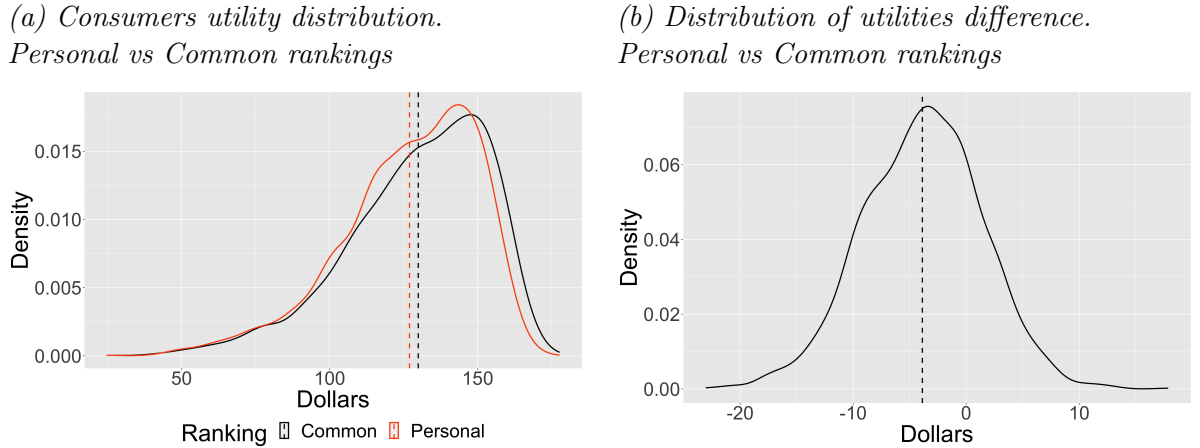
Note: In addition to Figure 6, this figure is another evidence to the fact that the assorting effect has larger magnitude for less prominent hotels, while the advertising one has larger magnitude for more prominent hotels.

Figure 8: Price change. Personal vs Common rankings



Note: The figure illustrates the percentage difference in hotel prices between the case of the personalized ranking and the common ranking as the function of the measure of the vertical differentiation of the hotels in query, measured as the standard deviation of utility among the hotels in a query. As the level of products vertical differentiation increases, the advertising effect becomes less important since consumers have stronger preferences toward one of the products, which leads to higher price difference.

Figure 9: Consumers' utility. Personal vs Common rankings

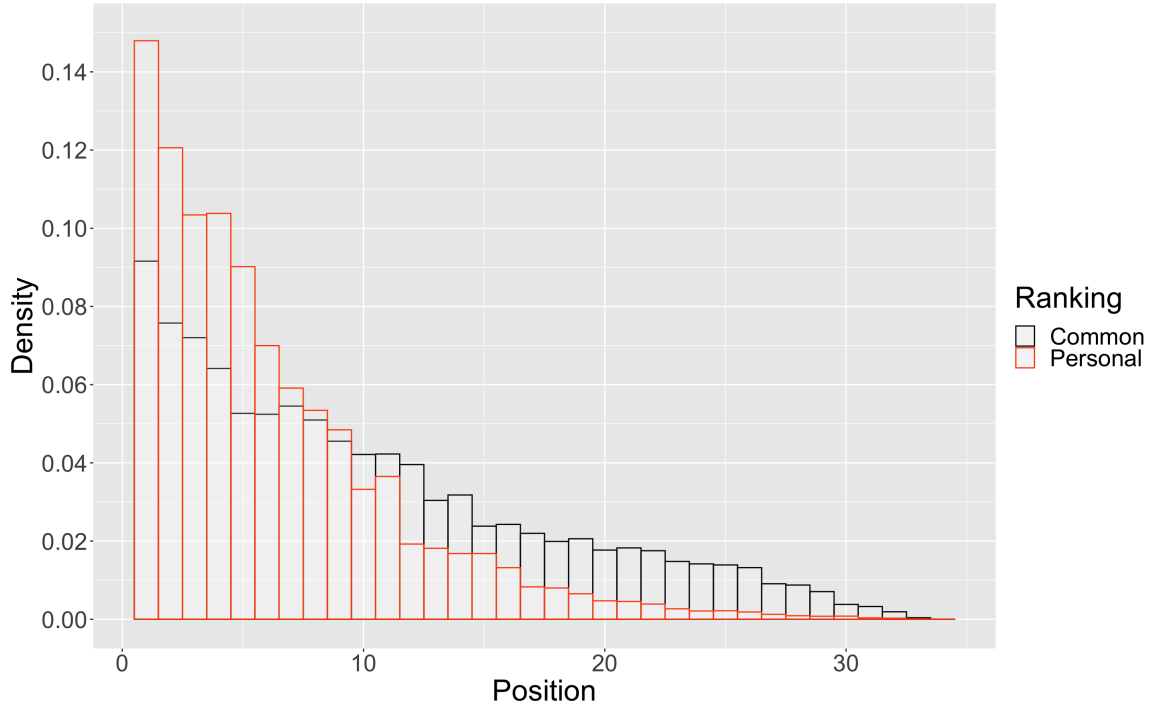


Note: The figure illustrates consumers' utility in the cases of personalized ranking and common ranking. The left panel provides distributions of consumers' utilities under two rankings. The right panel provides the distribution of the difference in consumers' utility under two rankings.

## Concluding remark

In this study I discuss the influence of the consumers' personal information aka big data on markets in the modern world. In many markets, consumers search costly among alternative options before making a purchase. The way that choices are presented to consumers has an impact on their search behavior and hence the competition and the market performance. Using the rich dataset which contains consumers search and purchase decisions,

Figure 10: Positions of booked hotels. Personal vs Common rankings



Note: The figure provides the histogram of positions of the hotels booked by consumers under personalized and common rankings. In the case of personalized ranking consumers more often book more prominent hotels.

I found that big data in average is harmful for consumers.

The fact that the platform uses consumer's personal preference data to provide him a better products ranking allows consumer spend less effort to find a suitable product and save in average .8% of utility (\$1.1) by the reduction of search expenditures and increase utility by .5% (\$0.7) by booking a better hotel. But the reduction of search intensity reduces the competition between firm, providing them incentives to raise prices. As a result, consumers lose 3.6% of utility (\$4.9) in average due the price increase. The resulting effect is negative in contrasts to all previous empirical studies, which didn't account for transaction price change.

## References

- Armstrong, Mark (2017), "Ordered consumer search." *Journal of the European Economic Association*, 15, 989–1024.
- Börsch-Supan, Axel and Vassilis A Hajivassiliou (1993), "Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models." *Journal of econometrics*, 58, 347–368.

- Choi, Michael, Anovia Yifan Dai, and Kyungmin Kim (2018), “Consumer search and price competition.” *Econometrica*, 86, 1257–1281.
- Einav, Liran and Leeat Yariv (2006), “What’s in a surname? the effects of surname initials on academic success.” *Journal of Economic Perspectives*, 20, 175–187.
- Ho, Daniel E and Kosuke Imai (2006), “Randomization inference with natural experiments: An analysis of ballot effects in the 2003 california recall election.” *Journal of the American Statistical Association*, 101, 888–900.
- Jeziorski, Przemyslaw and Ilya Segal (2015), “What makes them click: Empirical analysis of consumer demand for search advertising.” *American Economic Journal: Microeconomics*, 7, 24–53.
- Meredith, Marc and Yuval Salant (2013), “On the causes and consequences of ballot order effects.” *Political Behavior*, 35, 175–197.
- Moraga-González, José L, Zsolt Sándor, and Matthijs R Wildenbeest (2018), “Consumer search and prices in the automobile market.”
- Quint, Daniel (2014), “Imperfect competition with complements and substitutes.” *Journal of Economic Theory*, 152, 266–290.
- Shen, Jian (2015), “Ex-ante preference in a consumer search market.” *Available at SSRN 2826011*.
- Ursu, Raluca M (2018), “The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions.” *Marketing Science*, 37, 530–552.
- Weitzman, Martin L (1979), “Optimal search for the best alternative.” *Econometrica: Journal of the Econometric Society*, 641–654.