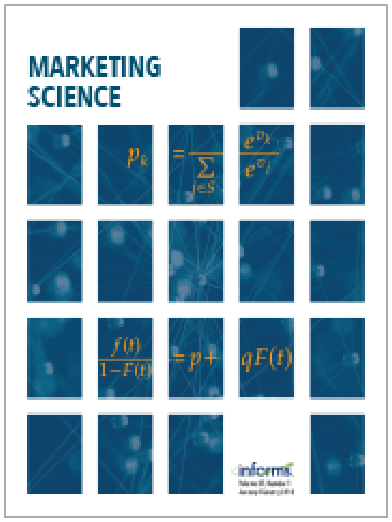


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Modeling Competition and Its Impact on Paid-Search Advertising

Sha Yang, Shijie Lu, Xianghua Lu

MARKETING SCIENCE
Vol. 33, No. 1, January–February 2014, pp. 134–153
ISSN 0732-2399 (print) | ISSN 1526-548X (online)
http://dx.doi.org/10.1287/mksc.2013.0812

Modeling Competition and Its Impact on Paid-Search Advertising

Sha Yang, Shijie Lu
Marshall School of Business, University of Southern California, Los Angeles, California 90089
{shayang@marshall.usc.edu, shijielu@usc.edu}

Xianghua Lu
School of Management, Fudan University, Shanghai, China 200433,
lxhua@fudan.edu.cn

Paid search has become the mainstream platform for online advertising, further intensifying competition between advertisers. The main objective of this research is twofold. On the one hand, we want to understand, in the context of paid-search advertising, the effects of competition (measured by the number of ads on the paid-search listings) on click volume and the cost per click (CPC) of paid-search ads. On the other hand, we are interested in understanding the determinants of competition, that is, how various demand and supply factors affect the entry probability of firms and, consequently, the total number of entrants for a keyword. We regard each keyword as a market and build an integrative model consisting of three key components: (i) the realized click volume of each entrant as a function of the baseline click volume and the decay factor; (ii) the vector of realized CPCs of those entrants as a function of the decay factor and the order statistics of the value per click at an equilibrium condition; and (iii) the number of entrants, the product of the number of potential entrants multiplied by the entry probability; the entry probability is determined by the expected revenue (a function of expected click volume, CPC, and value per click) and the entry cost at the equilibrium condition of an incomplete information game. The proposed modeling framework entails several econometric challenges. To cope with these challenges, we develop a Bayesian estimation approach to make model inferences. Our proposed model is applied to a data set of 1,597 keywords associated with digital camera/video and their accessories with full information on competition. Our empirical analysis indicates that the number of competing ads has a significant impact on the baseline click volume, decay factor, and value per click. These findings help paid-search advertisers assess the impact of competition on their entry decisions and advertising profitability. In the counterfactual analysis, we investigate the profit implication of two polices for the paid-search host: raising the decay factor by encouraging consumers to engage in more in-depth search/click-through and providing coupons to advertisers.

Key words: paid-search advertising; competition; Internet marketing; Bayesian estimation
History: Received: January 6, 2012; accepted: July 2, 2013; Preyas Desai served as the editor-in-chief and K. Sudhir served as associate editor for this article. Published online in *Articles in Advance* October 24, 2013.

1. Introduction

Paid (or sponsored) search has grown rapidly during the last decade, driving the Internet to become the second-largest media for advertising spending in the United States. This type of advertising format not only provides the largest source of revenue to traditional search engines such as Google but also has been extended to other business platforms such as online retailers (e.g., Amazon) and online market makers (e.g., eBay, Priceline) serving as a host for paid-search advertising. The paid-search host plays the role of directing users to relevant sponsored ads based on user-generated queries. When an Internet user enters a query, she receives search results containing both the organic links and paid links. If a user clicks on a paid link, she is directed to the advertiser's site, and the advertiser pays the search host a fee (i.e., cost per click, or CPC) for sending a potential customer.

As paid search becomes the mainstream platform for online advertising today, competition further intensifies. In paid-search advertising, a keyword is often regarded as a market reflecting a unique pattern of demand and supply. The demand captures the click volume of each ad on the sponsored search listings for that keyword, whereas the supply captures advertisers' decision of whether to enter the market (by purchasing the keyword) and how much to bid for their ads. Intuitively, the number of entrants or competing ads appearing on the paid-search listings for a given keyword affects consumer search and buying behavior, which will in turn influence advertisers' expected click volume, value per click, and CPC for entering such a market. At the same time, the number of entrants is related to advertisers' entry probability, which is often determined by the expected profit from paid-search advertising.

The main objective of this research is twofold. First, we want to understand, in the context of paid-search advertising, the effects of competition (the number of competing ads) on three key latent constructs that determine click volume and cost per click of paid-search ads: baseline click volume, decay factor, and value per click, where the decay factor can be interpreted as the conditional probability for consumers to click on the next ad on the paid-search listings. Second, we are interested in understanding the determinants of competition—that is, how various demand and supply factors affect the entry probability of firms and, as a result, the total number of entrants for a keyword.

The understanding of competition and its impact is important to both advertisers and the paid-search host. On the one hand, it helps advertisers more precisely evaluate the impact of competition on their expected profit and thus make better decisions about keyword choices. Furthermore, studying the separate effects of competition on the three key constructs can help improve paid-search advertisers' bidding effectiveness in the generalized second-price (GSP) auction. For example, advertisers should adjust their bids based on the number of competitors if competition affects click volume, decay factor, and/or value per click. This is because these parameters play a major role in determining the equilibrium CPCs in the GSP auction (Edelman et al. 2007, referred to as EOS hereafter). On the other hand, this analysis helps the paid-search host better understand how competition affects its own profit. If the paid-search host does benefit from competition, what policy changes can be made to improve its profit by influencing the competition? Our research provides a general framework to help the search host address this important question.

Utilizing a unique data set with full information on competition, we propose a structural framework to characterize competition and analyze its impact on click volume and CPC. Specifically, our integrative modeling framework has three major components: (i) the realized click volume of each entrant as a function of the baseline click volume and the decay factor; (ii) the vector of realized CPCs of those entrants as a function of the decay factor and the order statistics of the value per click at an equilibrium condition of the GSP auction; and (iii) the number of entrants as the multiplication of the number of potential entrants and the entry probability, and the entry probability is determined by the expected revenue (a function of expected click volume, CPC, and value per click) and the entry cost at the equilibrium condition of an incomplete information game.

We make several modeling contributions. Unlike most of the previous studies where click volume and CPC are modeled in a reduced-form fashion, we structurally model these two variables to allow inferences of the key underlying structural parameters. We also structurally model the number of competing ads as a reflection of both demand and supply conditions. The proposed modeling framework entails several econometric challenges. Specifically, order statistics stemming from the unobserved value per click induces cross-position correlation on CPCs of those entrants. Furthermore, the occurrence of the decay factor in both models of click volume and CPC requires a joint estimation. Finally, the number of entrants takes an implicit functional form, requiring numerical calculation of the Jacobian to simulate the likelihood function. To cope with these challenges, we develop a Bayesian estimation approach to make model inferences.

Several key findings emerge from our analysis. These include (i) the number of entrants (ads) positively affects the baseline click volume, (ii) the number of entrants has an inverse-U relationship with the mean decay factor, (iii) the number of entrants has a negative and convex relationship with the mean value per click of a keyword, and (iv) competition generally hurts advertisers but benefits the paid-search host.

Our structural analysis provides the paid-search host with some guidelines to improve its profitability. We conduct two counterfactual analyses as a demonstration. First, we show that the paid-search host could raise the decay factor by encouraging users to engage in more in-depth search/click-through on paid-search listings; such a policy change could help the paid-search host increase profit. Second, our analysis can also help the search host determine the optimal face value of entry coupons distributed to advertisers to increase the search host's profit.

The rest of this paper proceeds as follows. Section 2 reviews the relevant literature and positions our study in relation to previous studies. Section 3 describes the data and background information and develops the model. Section 4 provides an empirical application where we apply the model to real-world data collected from a large paid-search advertising host, and we discuss our findings. Section 5 provides managerial implications and presents two counterfactual analyses. Section 6 concludes this paper.

2. Relevant Literature

Our work is based on the growing literature on paid-search advertising. A series of papers have analytically examined advertisers' bidding behavior in the

GSP auction. EOS and Varian (2007) are the first two papers that characterize the bidding equilibrium in the GSP auction. EOS proved that the GSP auction is incentive incompatible; that is, bidding one's true value is not optimal. By studying a corresponding generalized English auction, they found that there exists a unique envy-free Bayes Nash equilibrium. Moreover, the ex post bids corresponding to this generalized English auction also satisfy the Nash equilibrium conditions of the GSP auction. Varian (2007) independently derived a similar equilibrium condition, which shows that the vector of equilibrium bids of the GSP auction can be expressed in a recursive form. The theoretical study on the GSP auction is further developed by Katona and Sarvary (2010), who extended the model to account for the heterogeneity of click-through rate (i.e., the ratio of actual clicks to the number of impressions) across competing ads and build the link between sponsored ads and organic ads. Athey and Nekipelov (2012) recently introduced advertisers' uncertainty in quality scores to the GSP auction and presented theoretical conditions for the existence of a unique Nash equilibrium. They also proposed a computation algorithm to infer the bounds of bidders' valuations and applied their method to the historical data of several keywords as a demonstration.

These studies have provided important theoretical foundations to modeling advertisers' bidding strategies in the GSP auction. However, the theoretical results regarding paid-search advertiser's bidding behavior have not been empirically investigated. In this paper, we characterize the CPC formation based on the equilibrium condition provided by EOS and Varian (2007). We jointly estimate the CPC, click volume, and the number of entrant advertisers, and we infer the distributions of baseline click volume, mean value per click, and mean decay factor. We also identify keyword characteristics that affect these three parameters. Our proposed structural model allows us to derive insights that cannot be obtained from a reduced-form model and fits the data better.

On the empirical side, several papers have examined marketing-related issues in the context of paid-search advertising. For example, Ghose and Yang (2009) simultaneously modeled click-through, conversion, CPC, and ad position using keyword-specific data from one retailer. Yang and Ghose (2010) extended their previous work by analyzing consumer click-through and conversion on both sponsored search listings and organic search listings for the same keyword. Rutz and Bucklin (2011) explored the potential spillover effects between activities associated with generic and branded keywords in paid-search advertising, using data from a hotel chain. Goldfarb and Tucker (2010) empirically studied the price variation on paid-search ads related to a legal service on Google and found evidence of substitution of online advertising for off-line advertising. These aforementioned studies generally employed a reduced-form approach and focused on predicting paid-search ad performances.

Few studies have empirically examined the underlying competition of paid-search advertising. Two important pieces of work need to be mentioned here. Chan and Park (2013) studied the influence of sequential search behaviors of consumers on the value of click-throughs in sponsored search advertising. They modeled the position competition in the context of a first-price auction with a buy-it-now option, which allows advertisers to acquire a position without submitting a bid. Since a unique equilibrium cannot be obtained in such an auction mechanism, they used the moment-inequality estimation approach to avoid imposing restrictive assumptions on equilibrium selection and infer advertisers' value per click from the observed ad positions. Yao and Mela (2011) also modeled the position competition in the first-price auction with a sorting/filtering function available to users. They emphasized the dynamics in forward-looking advertisers' bidding strategies and used the Markov perfect equilibrium to characterize advertisers' bids. They estimated the model by applying the two-step estimators developed by Bajari et al. (2007), assuming the existence and uniqueness of the equilibrium.

Our paper differs from this line of work in several ways. First, our paper focuses on the effect of competition (the number of competing ads) on click volume and CPC through three latent variables: baseline click volume, decay factor, and value per click, whereas previous studies have not examined the separate effects of competition on these variables. Second, we develop an integrative model of click volume, CPC, and number of entrants. However, the previous two studies have not considered the entry decisions of advertisers and therefore have not modeled the number of entrants for a given keyword. Third, unlike the previous two studies, which looked at paid-search advertising in the first-price auction with a small number of ad positions, we study the GSP auction without capacity constraint, which is the most popular type of paid-search mechanism. Fourth, the characterization of CPC in our paper is built on the Nash equilibrium condition, which is theoretically proved and derived by EOS. This equilibrium condition enables us to more closely examine the realized CPCs of advertisers and estimate the distribution of value per click. We develop a Bayesian estimation algorithm to cope with the econometric challenges of the CPC model based on this equilibrium condition.

Knowing that the number of entrants not only affects click volume and CPC but also reflects the demand and supply conditions associated with a keyword, we model number of entrants as an aggregate outcome of entry decisions made by potential entrants. Since we model advertisers' simultaneous entry decisions, our paper is related to the literature of simultaneous-move game. The pioneering work in this line of research includes Bresnahan and Reiss (1990, 1991) and Berry (1992). The first two find that the multiple-equilibrium issue is prevalent in the simultaneous-move game, and they suggest that one way to bypass the multiple equilibrium is to focus on the total number of entrants in a market rather than the vector of individual entry decisions. One important finding in Berry (1992) is that the number of entrants is uniquely determined if the profit function strictly decreases with the number of entrants. Recently, more complicated simultaneous-move entry models have been developed by endogenizing either firms' product differentiation (Mazzeo 2002, Seim 2006) or spatial differentiation (Zhu and Singh 2009), or both (Datta and Sudhir 2011), accounting for the spillover effect within a market (Vitorino 2012) or across markets (Jia 2008), and incorporating the effect of zoning regulations into market structure (Datta and Sudhir 2013).

Because of the large number of potential entrants in our empirical context, we follow the previous literature to model the entry decisions of advertisers as a simultaneous-move game with incomplete information (Seim 2006, Zhu and Singh 2009, Datta and Sudhir 2011, Vitorino 2012). In other words, the profitability of each advertiser in a keyword is private information, and only its distribution is common knowledge among competitors. We further assume advertisers to be symmetrical for the following reasons. First, because one of our main objectives is to study the impact of number of entrants on three underlying constructs of click volume and CPC, it is reasonable to make this assumption to build an internally consistent model. Second, following Berry's (1992) idea to assume that advertiser's expected profit is a decreasing function of number of entrants, we can prove the existence of a unique equilibrium for number of entrants. Finally, because of the heavy computational burden, current methods can only handle a small number of heterogeneous players (e.g., Datta and Sudhir 2011, Vitorino 2012). However, because there are a large number of potential entrants in our empirical context, it is infeasible to adopt the same modeling approach.

3. Data and Proposed Model

3.1. Description of the Data

We obtain data from a leading online market maker outside the United States who hosts paid-search advertising. We regard each keyword as a market, and consequently, an advertiser is named an entrant to a keyword if she decides to advertise her product using this keyword. For this paid-search host, ad positions are auctioned in the second-price fashion and are entirely determined by the rank of bids submitted by entered advertisers. Each advertiser then pays the highest bid among all bids below hers for each click. In other words, the auction mechanism used in our data is exactly the same as the GSP auction defined in EOS.

Our data include aggregate information on 1,573 keywords of digital camera/video products and related accessories in June 2010. There are 359 advertisers that advertised through a subset of these 1,573 keywords. According to the paid-search host, advertisers in that market often review their keyword lists and make purchase decisions monthly. These 1,573 keywords are further classified into three main categories: digital camera, digital video, and accessory. In addition to this primary categorization, each keyword also belongs to one or several of 44 subcategories.[1]

We create several keyword attributes. First, we define three variables: *DV*, *Accessory*, and *Coverage*, based on the categorical information of a keyword. The variable *Coverage* measures the number of subcategories a keyword belongs to, which indicates the market breadth of the keyword. Second, we define several keyword attributes based on the product-related information: *Brand* (whether the keyword contains a brand name), *General* (whether the keyword includes a general feature that could apply to different products), and *Specific* (whether the keyword includes a specific feature such as model/series number that exclusively refers to a product). In addition, we also have the length information for each keyword. The variable *Length* indicates the total number of characters included in the keyword. Finally, we create a dummy variable, *Promotional*, if a keyword includes promotional terms. Taking the keyword "Nikon D700 HD Cheap" as an example, a brand name (Nikon), a specific word (D700), a general feature (HD, for "high definition"), and a promotional term (Cheap) are included. Table 1 reports the summary information of keyword characteristics.

For each keyword, the paid-search host informed us of the composition of its competition set (i.e., the

[1] Each subcategory can be regarded as a refined classification of the main category. For example, the keyword "Canon camera" belongs to two subcategories: the "ordinary digital camera" and "professional SLR (single-lens reflex) camera."

Table 1 Summary Statistics of Keyword Characteristics

Variable	Mean	SD	Min	Max
DV	0.11	0.31	0	1
Accessory	0.42	0.49	0	1
Coverage	2.04	1.91	1	17
Length	5.90	2.40	1	18
Brand	0.47	0.50	0	1
General	0.22	0.42	0	1
Specific	0.53	0.49	0	1
Promotional	0.08	0.27	0	1
n	9.10	4.40	5	25
*N*_Potential	119.00	61.00	5	306

Notes. All keyword characteristics including n are mean centered in our empirical implementation. The variables n and n^2 are scaled by 10 and 100, respectively, in estimation.

Table 2 Summary Statistics of Click Volume and CPC Across Keywords

Variable	Mean	SD	Min	Max
Total click volume	195.79	464.46	10	8,246
Average click volume across positions	17.88	33.28	1.25	515.38
Average CPC across positions (in cents)	14.53	5.36	5.27	53.40

set of potential advertisers). The selection is mainly based on two criteria. First, for keywords that link to a specific product or brand, the set of potential entrants includes advertisers that carry this product or brand in their stores associated with this paid-search host. Second, for a keyword that does not link to a specific product or brand, the set of potential entrants includes those who bought other keywords that share similar subcategories (e.g., professional digital camera, single-lens reflex) with it. As shown in Table 1, the number of entrants ranges from 5 to 25, and the number of potential entrants ranges from 5 to 306. On average, each keyword belongs to two subcategories. For each keyword, we have information on the aggregate click volume, average CPC, and average positions for each entered advertiser.[2] Table 2 reports the summary statistics of the average click volume and CPC across keywords.

3.2. Model Setup

To fix the context, we have I advertisers and K keywords. Potential entrants/advertisers of each keyword are indexed by i, and keywords are indexed by k. Let C_k stand for the set of potential entrants of keyword k. To be consistent with the industry practice that advertisers choose the set of keywords on a monthly basis and then optimize their bids after entry, we model advertisers' keyword selections and bid decisions as the following two-stage sequential process.

In the first stage, potential entrants $i \in C_k$ decide whether to purchase a specific keyword k. We model the entry of advertisers as a simultaneous-move game with incomplete information, in which advertisers possess private information about their own profitability. Since potential entrants do not observe realized click volume, CPCs, and value per click before entry, they are assumed to form expectations on these variables as well as the entry decisions of others. Based on these expected values, each potential entrant will form expectations on the advertising revenue and entry cost. We further assume that potential entrants are symmetrical, and the entry probability is determined by the expected revenue and the entry cost in a keyword. This microlevel process determines the total number of entrants for a keyword at the aggregate level. The equilibrium number of entrants is modeled as the number of potential entrants multiplying by the entry probability.

In the second stage, after entry decisions have already been made (i.e., the number of entrants n has been realized and becomes common knowledge), entrant advertisers now determine how much to bid. The equilibrium CPCs and ad positions are then determined by realizations of value per click. The value per click of each entrant advertiser is assumed to be private information. In the same setup, EOS proved that there exists a unique Bayes Nash equilibrium. In this equilibrium, ad positions are determined by the descending order of value per click, and the associated CPCs are shown to be a recursive function of value per click. Click volume at each position is then realized as users' responses to paid-search ads.

We next model the three main equilibrium outcomes of this game backward. We first present the model of click volume conditional on the rank of advertisers. Then we discuss how to model advertisers' equilibrium CPCs conditional on the number of entrants and their value per click. Finally, we present the model of number of entrants.

3.3. Modeling Click Volume

Let n_k stand for the number of entrants in keyword k, and let Q_{ki} stand for the realized click volume for advertiser i displayed in the paid-search results of keyword k at its realized position j_{ki}. A smaller j_{ki} corresponds to a higher position (i.e., more toward the top). Following Feng et al. (2007), we assume that the expected click volume of an ad decreases exponentially with its position. We model the click volume as

$$Q_{ki} = \begin{cases} B_k \exp(\epsilon_{ki}) & \text{if } j_{ki} = 1, \\ B_k \prod_{l=1}^{j_{ki}-1} \delta_{kl} \exp(\epsilon_{ki}) & \text{if } j_{ki} \geq 2, \end{cases} \tag{1}$$

[2] We were informed by the data provider that there is very small fluctuation on these measures within the data period.

where B_k stands for the baseline click volume at the top position for keyword k; δ_{kj} is the decay factor, which stands for the ratio of click volume between position $j+1$ and position j for keyword k ($0 < \delta_{kj} < 1$); and ϵ_{ki} is the noise component distributed as normal with mean zero and variance σ_q^2. Here, ϵ_{ki} is a measurement error between the expected and realized log click volume and is unknown to advertisers.[3] As for δ_{kj}, we assume that these position-specific decay factors are common knowledge to entrants because each entrant of a keyword can easily learn δ_{kj} by experimenting bids to change positions. Furthermore, because the realized decay factors depend on consumers' search behavior given the search listings, we assume that δ_{kj} is observed by advertisers only after their entry. After taking log on both sides of Equation (1), we obtain

$$\ln(Q_{ki}) = \begin{cases} b_k + \epsilon_{ki} & \text{if } j_{ki} = 1, \\ b_k + \sum_{l=1}^{j_{ki}-1} \ln(\delta_{kl}) + \epsilon_{ki} & \text{if } j_{ki} \geq 2, \end{cases} \tag{2}$$

$$\epsilon_{ki} \sim \mathrm{N}(0, \sigma_q^2), \tag{3}$$

$$b_k = \gamma_1 X_k + \eta_{k1}, \tag{4}$$

where $b_k = \ln(B_k)$, X_k is a vector of keyword characteristics where the first column is 1 and one covariate is the number of entrants, and η_{k1} is a keyword-specific shock that is common knowledge to all potential entrants; η_{k1} accounts for the unobserved heterogeneity in keyword-specific log baseline click volume. Its distribution specification will be described in §3.5, along with other keyword-specific error terms. We parameterize the decay factor δ_{kj} with a logit transformation to ensure $0 < \delta_{kj} < 1$:

$$\delta_{kj} = \frac{\exp(\lambda_{kj})}{1 + \exp(\lambda_{kj})}, \tag{5a}$$

$$\lambda_{kj} \sim \mathrm{N}(\lambda_k, \sigma_d^2), \tag{5b}$$

$$\lambda_k = \gamma_2 X_k + \eta_{k2}, \tag{6}$$

where λ_{kj} is the transformed decay factor, which is assumed to be normally distributed, and λ_k stands for the mean transformed decay factor, which is commonly known to all potential entrants. Similar to η_{k1} in Equation (4), η_{k2} captures the unobserved heterogeneity in λ_k. The mean (transformed) decay factor λ_k is a key parameter that governs the click volume across positions. Generally speaking, the larger the mean (transformed) decay factor, the smaller the variation of click volume across two adjacent positions, and the less important the ad position is in attracting click volume, holding everything else constant.

[3] All error terms specified in our model are unknown or unobservable to researchers. Therefore, we only explain whether an error term is known to advertisers, and, if so, at what stage advertisers know about it.

3.4. Modeling Cost per Click

Conditional on the total number of entrants n_k, advertisers submit bids. These submitted bids determine the rank of ads as well as CPCs of advertisers according to the rule of the GSP auction. We assume that each advertiser's value per click is private information. As discussed in EOS, the unique envy-free Bayes Nash equilibrium derived in the corresponding generalized English auction provides a closed-form solution. More important, the realized bids from this generalized English auction also satisfy the Nash equilibrium conditions of the GSP auction and are independent of advertisers' beliefs on others' valuations. As EOS pointed out in their paper, "This equilibrium has some notable properties. The bid functions have explicit analytic formulas, which, combined with equilibrium uniqueness, make our results a useful starting point for empirical analysis" (p. 253).

Following EOS, we derive the equilibrium CPC in a recursive form by characterizing CPC at each position as a weighted sum of the CPC and the advertiser's value per click at the adjacent position below. We provide a detailed derivation of CPC equation and prove several equilibrium properties in Appendix A. Since the theoretical proposition indicates that the rank of ads is in line with advertisers' value per click at equilibrium, we model advertisers' value per click as order statistics generating from a distribution. Let CPC_{ki} stand for the CPC advertiser i pays at position j_{ki} for keyword k. Then the equilibrium condition can be written as

$$\mathrm{CPC}_{ki} = \delta_{kj_{ki}} \mathrm{CPC}_{ki'} + (1 - \delta_{kj_{ki}}) S_{ki'}, \tag{7}$$

$$\{S_{ki}\}_{j_{ki}=2}^{n_k} \text{ are } 1, 2, \ldots, n_k - 1 \text{ descending-order statistics of } \{\tilde{S}_{ki}\}_{i=1}^{n_k-1}, \tag{8}$$

where i' refers to the advertiser who stays right below advertiser i (i.e., $j_{ki'} = j_{ki} + 1$), and $\{\tilde{S}_{ki}\}_{i=1}^{n_k-1}$ are independent and identically distributed (i.i.d.) log-normal distributed. Since the CPC of the advertiser at the top position is independent of her value per click, we cannot infer the first advertiser's value per click from observed CPC data. We also treat the CPC at the bottom position as exogenous since CPCs at the bottom are five cents in our data, which is the minimum bid set by the search host. Similar to δ_{kj}, an advertiser's value per click S_{ki} is assumed to be known after entry because the realized value per click for an advertiser depends on the conversion rate of her ad. Such demand-side information can only be accurately learned by advertisers after entering the keyword market. However, unlike δ_{kj}, S_{ki} is assumed to be private information to advertiser i and is known to others only up to the distribution. Because each advertiser's value per click is always positive, we take a log transformation of $\tilde{S}_{ki}$ and incorporate the observed heterogeneity as follows:

$$\ln(\tilde{S}_{ki}) \sim \mathrm{N}(V_k, \sigma^2_{\mathrm{vpc}}), \tag{9}$$

$$V_k = \gamma_3 X_k + \eta_{k3}, \tag{10}$$

where V_k stands for the mean log value per click of keyword k, and η_{k3} accounts for unobserved heterogeneity in V_k. Similar to η_{k1} and η_{k2} above, η_{k3} is assumed to be publicly known to all potential entrants. Note that $\tilde{S}_{ki}$ are latent variables that need to be inferred.

As shown in Equations (2) and (7), decay factors δ_{kj} are the key parameters bridging the click-volume and CPC equations. Intuitively, the decay factor at position j can be interpreted as the probability for a consumer to further click on an ad conditional on her click on the jth ad. This implies that the decay factor at each position is determined by the marginal search benefit and cost of a consumer. The decay factor is higher when the consumer incurs higher marginal return from search. The click-volume decay factor also affects the distribution of CPCs across positions. Consider two extreme cases where $\delta_{kj} = 1$ and $\delta_{kj} = 0$ for all $j = 1, \ldots, n_k - 1$. When $\delta_{kj} = 1$, no advertiser cares about the result of the auction since each position is equally profitable. Hence, advertisers will all bid zero in the GSP auction. When $\delta_{kj} = 0$, only the top position receives clicks, and all advertisers will bid their true value, as predicted in our model (see Equation (7)). To some extent, a vector of small decay factors drives advertisers to bid more aggressively for the top position.

3.5. Modeling the Number of Entrants

We focus on the keyword entry of advertisers at the aggregate level by modeling the number of entrants for a number of reasons. First, recall that one of the main objectives of this research is to understand, in the context of paid-search advertising, the effect of competition (the number of entrants) on three key latent constructs that determine click volume and CPC: baseline click volume, decay factor, and value per click. It makes sense for us to model the number of entrants to control for the potential endogeneity.

Second, given that we have a large number of potential entrants, it is rather difficult to treat and model their entry decisions individually as a result of the multiple-equilibrium problem and computational burden. To the best of our knowledge, the extant literature on entry game can handle only a handful of heterogeneous players (e.g., Datta and Sudhir 2011, Vitorino 2012). Therefore, we assume potential entrants to be symmetrical. We further assume that the expected profit function of an advertiser decreases with the number of entrants. Under these two assumptions, we later show that there exists a unique equilibrium for the number of entrants. We next outline the model we proposed on the number of entrants.

We model advertisers' entry decisions as a simultaneous-move game with incomplete information. The advertiser's utility of purchasing a keyword consists of three parts: the expected revenue from search advertising, the management/entry cost that covers different types of cost associated with managing a keyword (e.g., cost of designing/revising ad title, ad copy, and landing page), and an idiosyncratic shock of entry cost, which is private information for each advertiser:

$$U_{ki} = \beta[\Pi_k(n_k, \theta_k) - F_k] + \xi_{ki}, \tag{11}$$

$$\xi_{ki} \sim \text{ Extreme value } (0, 1), \tag{12}$$

$$F_k = \gamma_4 X^F_k + \eta_{k4}, \tag{13}$$

where Π_k is the expected revenue of advertising through keyword k, and F_k is the keyword-specific entry cost, which depends on a vector of keyword attributes X^F_k; X^F_k is the same as X_k except that it does not include the number of entrants n_k in our empirical implementation; η_{k4} captures the unobserved heterogeneity in F_k and is assumed to be common knowledge to potential entrants; and ξ_{ki} is the private shock of entry cost known by advertiser i in keyword k and is assumed to follow a Type I extreme value distribution. The parameter β measures how advertiser's expected profit (on a monetary unit) affects entry probability.

The expected advertising revenue Π_k is formed based on the expectation of click volume, value per click, and CPC. These values further depend on three keyword-specific parameters: log baseline click volume (b_k), mean transformed decay factor (λ_k), and mean log value per click (V_k). We assume that all potential entrants of a keyword know these three keyword-specific parameters (b_k, λ_k, V_k) at the stage of entry. We believe this is a reasonable assumption because advertisers can learn these parameters from their previous advertising experience and keyword statistics provided by the search host.[4] Thus, the

[4] Most large search hosts such as Google and Bing provide advertisers with free tools to obtain aggregate information on click volume and CPC for various keywords. The search host in our data offers a similar service to advertisers.

expected advertising revenue for a potential entrant is

$$\Pi_k(n_k) = E_{Q,\lambda,S}\{Q(b_k(n_k), \lambda_k(n_k))[S(V_k(n_k)) - \mathrm{CPC}(S_k, \lambda_k(n_k))]\}, \quad (14)$$

where E is the expectation with respect to the vector of value per click of all entrant advertisers, the vector of decay factors across positions, and the vector of click volume across positions. In other words, the expectation in Equation (14) is taken with respect to the position- or advertiser-specific errors in Equations (3), (5b), and (9), but not with respect to the keyword-specific errors $(\eta_{k1}, \eta_{k2}, \eta_{k3})$ associated with (b_k, λ_k, V_k). More details on how to obtain an advertiser's expected revenue are provided in Appendix B.

Assuming ξ_{ki} follows a Type I extreme value distribution, the entry probability of a representative potential entrant is

$$P_k(n_k, F_k) = \frac{\exp[\beta(\Pi_k(n_k) - F_k)]}{1+\exp[\beta(\Pi_k(n_k) - F_k)]}. \quad (15)$$

Following the previous literature (Seim 2006, Datta and Sudhir 2011), we assume that the unobserved keyword-specific cost parameter F_k is adjusted to equate the observed number of entrants and the expected number of entrants predicted by the model. This assumption indicates that the entry cost F_k is just low enough for n_k entrants observed from the data. Therefore, n_k is implicitly determined by the following equation:

$$n_k = N_k P_k(n_k, F_k), \quad (16)$$

where N_k is the number of potential entrants for keyword k. Under the assumption that the expected revenue of advertisers decreases with competition (i.e., Π_k decreases with n_k), the entry probability specified in Equation (15) also decreases with n_k. This suggests that the left-hand side of Equation (16) increases with n_k, whereas the right-hand side decreases with n_k. Thus, we have shown that Equation (16) leads to a unique number of entrants n_k at equilibrium.

By substituting (15) into (16), we have

$$n_k = \frac{N_k \exp[\beta(\Pi_k(n_k) - F_k)]}{1+\exp[\beta(\Pi_k(n_k) - F_k)]},$$

from which F_k can be expressed as follows as a function of observed number of entrants and number of potential entrants:

$$F_k = \Pi_k(n_k) + \frac{1}{\beta}[\ln(N_k - n_k) - \ln(n_k)]. \quad (17)$$

The number of entrants can be endogenous in both click volume and CPC models since the keyword entry decisions of advertisers are based on their expected profits, which depend on the expected click volume, value per click, and CPC. To control for the potential endogeneity of n_k in click and CPC models, we need to simultaneously estimate the models of click volume, CPC, and number of entrants. We also assume the four keyword-specific errors in log baseline click volume, mean transformed decay factor, mean log value per click, and entry cost to be normally distributed and potentially correlated. Thus, the conditional likelihood of the number of entrants can be derived as the conditional likelihood of entry cost multiplying by the Jacobian:

$$(\eta_{k1}, \eta_{k2}, \eta_{k3}, \eta_{k4}) \sim \mathrm{MVN}(0, \Omega), \quad (18)$$

$$\Pr(n_k \mid b_k, \lambda_k, V_k, \Omega) = \Pr(F_k \mid b_k, \lambda_k, V_k, \Omega)\left|\frac{\partial F_k}{\partial n_k}\right|. \quad (19)$$

4. Estimation Results

4.1. Identification

We first show the theoretical identification of our model. We perform a simulation study in Appendix C by estimating the proposed model based on simulated data. The estimation results suggest that all parameters in our proposed model are identifiable and can be recovered within 95% confidence intervals of their true values. We next discuss the empirical identification of our model parameters.

A unique aspect of our data is that we observe both keyword entry decisions and post-entry information including CPCs and the realized click volume for each entrant advertiser. Using an analogy, CPC and click volume correspond to the price and sales data in the retailing context. These two pieces of information substantially help researchers identify the separate effects of number of competing firms on baseline click volume, mean decay factor, and mean value per click in search advertising context. Specifically, the competition effect on baseline click volume is identified through the variation in log click volume across positions and across keywords. We also observe variation in number of entrants in different keywords. Furthermore, because the decay factor appears in both click-volume and CPC equations, the click ratio between adjacent positions and the correlation between consecutive CPCs help us identify the decay factor. Given the decay factor, we can then identify the mean value per click.

We next illustrate the empirical identification of the equation of the number of entrants. Note that we are not estimating Equation (11), which is only used to derive the entry probability. Instead, we estimate Equation (17), in which F_k is a function of the observed number of entrants and number of potential entrants. Note that F_k is also a linear function of X_k^F, as specified in Equation (13). As a demonstration, we assume here the error term η_{k4} to be independent of other keyword-specific errors η_{k1}, η_{k2}, and η_{k3}, whereas we allow these four errors to be correlated in the model:

$$\eta_{k4} \sim \mathrm{N}(0, \sigma_4^2). \tag{20}$$

Based on Equations (13) and (20), Equation (17) can be rewritten as

$$\ln(N_k - n_k) - \ln(n_k) = \beta\gamma_4 X_k^F - \beta\Pi_k + \beta\eta_{k4}, \tag{21}$$

which is equivalent to the following equation after reparameterization:

$$\ln(N_k - n_k) - \ln(n_k) = \alpha X_k^F - \beta\Pi_k + e_k, \tag{22}$$

where $\alpha = \beta\gamma_4$ and $e_k = \beta\eta_{k4}$.

Note that the expected entry profit $\Pi_k(X_k, \theta_k)$ depends on both keyword attributes X_k and keyword-specific parameters θ_k (i.e., $\theta_k = (b_k, \lambda_k, V_k)$) estimated from the click-volume and CPC equations in a highly nonlinear fashion. Given that θ_k can be uniquely identified from the click-volume and CPC equations, Π_k can then be constructed. Because Π_k includes post-entry information in addition to keyword attributes, its coefficient β in Equation (22) can then be identified based on the empirical relationship between $\ln(N_k - n_k) - \ln(n_k)$ and Π_k. Once β is identified, γ_4 and σ_4^2 can be identified accordingly as in a regression. Intuitively, β should be positive, because as Π_k increases, we should expect more entrants and fewer nonentrants; that is, the ratio between nonentrants and entrants will decrease. In other words, $\ln(N_k - n_k) - \ln(n_k)$ and Π_k are expected to have a negative relationship. We later show that our estimate of β is positive, giving some empirical face validity about the model identification.

The discussion above aims to provide the basic idea of empirical identification of our model. The actual model is more complicated given that the expected profit Π_k is also a function of n_k. In that case, it is difficult to derive the explicit function of n_k, and therefore we numerically compute the Jacobian when deriving the likelihood function. To ensure that the proposed identification strategy also works for the more complicated model, we conduct a simulation study to confirm the theoretical identification of our model.

In sum, the identification strategy for competition effects in our study is not the same as the identification strategy applied in previous studies in which researchers only observe firms' entry or location choices (e.g., Zhu and Singh 2009, Vitorino 2012). Those studies typically rely only on the existence of exclusion restrictions to exploit the variation in the number of rivals across markets to identify strategic interaction effects (see Bajari et al. 2010 for a detailed discussion). In our study, we identify the competition effects not only from the observed keyword entry but also from the variation in subsequently realized click volume and CPC across positions and across keywords. This follows the identification strategy adopted by Datta and Sudhir (2011), who first articulated that in addition to the market entry data, the post-entry revenue and price data are essential for the identification of different types of spillover effects.

4.2. Model Estimation

Several challenges arise in model estimation. First, values per click are unobserved. On top of that, they are based on order statistics, which induces unobserved correlations across CPCs at different positions. Second, decay factors appear in models of both click volume and CPC. Third, because parameters associated with the baseline click volume, mean decay factor, and mean value per click affect advertisers' entry decisions, we need to simultaneously estimate models of click volume, CPC, and the number of entrants. Finally, the number of entrants takes an implicit functional form, requiring numerical calculation of the Jacobian to simulate the likelihood function.

To address these challenges, we adopt the Bayesian estimation approach to make model inferences. The key advantage of the Bayesian approach lies in its convenience in handling complex models. As shown, the model involves a large number of latent variables. Moreover, the Bayesian estimation approach can help obtain keyword-specific estimates for parameters $\{b_k, \lambda_k, V_k, F_k\}$ as a by-product of the Markov chain Monte Carlo (MCMC) algorithm. These keyword-level estimates are essential for the paid-search host to evaluate the impact of policy changes on its revenue.

We estimate the proposed model by implementing MCMC methods. We iteratively generate draws from the model parameters for 100,000 iterations and inspect the time series of model parameters to assess convergence. We use the first 50,000 draws as the burn-in and keep every 100th draw of the remaining 50,000 draws to estimate the posterior mean and standard deviations. Appendix C provides details of the MCMC algorithm (e.g., prior and posterior distribution).

4.3. Empirical Findings

To validate our proposed model on click volume, CPC, and the number of entrants, we compare it with a benchmark model where all three variables are modeled in a reduced form with the same set of information. In the benchmark model, the log-transformed click volume is modeled as a function of rank and keyword-specific attributes, and the rank coefficient (i.e., decay factor) is also modeled as a function of keyword attributes. CPC is modeled in a similar way except for the inclusion of an autoregressive term. We allow unobserved heterogeneity across keywords and correlation on all random coefficients in both the click volume model and the CPC model. We model number of entrants for keyword k as a linear regression:

$$n_k = \gamma_5 W_k + \eta_{k5}, \tag{23}$$

where W_k includes keyword-specific attributes (X_k^F), average click volume, and average CPC in a keyword, as well as two instrumental variables. The first instrument variable is the number of potential entrants (*N_Potential*), and the second one is the average number of entrants in similar keywords (*N_Similar*) for a given keyword k. Keywords are defined to be similar to each other if they belong to the same subcategory. We find that these two variables significantly predict n_k. On the other hand, when Internet users make click decisions, we believe that they view only the information from those entered advertisers, and therefore *N_Potential* is less likely to affect their clicking behavior. Furthermore, we believe that the user click intention is mainly influenced by the information on the keyword level rather than by the information on the category level. Therefore, the second instrument we present is also unlikely to be correlated with the error term in the click equation. A similar argument can be applied to the CPC equation. By adopting Equation (23) and allowing η_{k5} to be correlated with the unobserved heterogeneity error terms in equations of click volume and CPC, we control for the potential endogeneity of n_k in the benchmark model using the limited information approach (Villas-Boas and Winer 1999, Yang et al. 2003).

Our analysis suggests that our proposed model (log-marginal = 31,943) outperforms the reduced-form model (log-marginal = 57,083) substantially. We report the coefficient estimates from the reduced-form benchmark model in Tables 3 and 4. In the reduced model, the number of competing ads affects click volume and CPC for an individual advertiser directly in the form of a standard regression. As shown in Table 3, the number of entrants positively affects baseline click volume and CPC and negatively affects the click-volume decay factor.

Tables 5 and 6 report the coefficient estimates from the proposed model. Our proposed structural model differs from the reduced model in three ways. First, the click-volume decay factor δ_{kj} appears in both the click-volume equation and CPC equation, bridging the two important elements that will jointly affect an advertiser's expected revenue from a keyword. Second, it allows us to examine the different mechanisms under which number of entrants affects an individual advertiser's CPC. For example, the number of entrants can affect an advertiser's CPC

Table 4 Variance–Covariance Estimates from the Reduced-Form Model

Panel A: Standard deviations

Variable	Mean	SD
ln(*Click*)	**1.098**	**0.007**
CPC	**3.936**	**0.027**

Panel B: Variance–covariance matrix

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
ϕ_1 (Intercept in ln(*Click*))	**1.207** (**0.096**)	**−0.071** (**0.010**)	**0.130** (**0.059**)	**−0.076** (**0.020**)	**0.047** (**0.009**)	**−0.358** (**0.024**)
ϕ_2 (Intercept in decay factor)		**0.022** (**0.001**)	**−0.008** (**0.004**)	**0.004** (**0.002**)	**−0.004** (**0.001**)	**0.016** (**0.004**)
ϕ_3 (Intercept in CPC)			**0.041** (**0.013**)	−0.008 (0.006)	0.003 (0.003)	**−0.041** (**0.018**)
ϕ_4 (Rank coeff. in CPC)				**0.075** (**0.007**)	**−0.034** (**0.003**)	**0.036** (**0.006**)
ϕ_5 (Lag CPC coeff. in CPC)					**0.041** (**0.002**)	**−0.026** (**0.004**)
ϕ_6 (Error term in n_k)						**0.167** (**0.006**)

Notes. Posterior mean (posterior standard deviation) is reported for each parameter. Estimates that are significant at 95% are in bold.

Table 3 Estimation Results of the Reduced-Form Model

Keyword attributes	ln(*Click volume*)	Decay factor	CPC	Number of entrants
Intercept	**2.293** (**0.040**)	**−0.091** (**0.009**)	**4.044** (**0.299**)	**−0.092** (**0.011**)
Rank			**−0.472** (**0.032**)	
Lag CPC			**1.050** (**0.014**)	
DV	−0.193 (0.111)	−0.016 (0.020)	**0.643** (**0.189**)	−0.038 (0.036)
Accessory	−0.140 (0.089)	−0.004 (0.015)	0.030 (0.186)	−0.012 (0.033)
Coverage	−0.335 (0.172)	0.027 (0.028)	−0.621 (0.358)	**0.242** (**0.067**)
Length	**−1.337** (**0.186**)	**0.065** (**0.031**)	0.153 (0.402)	**0.120** (**0.057**)
Brand	−0.074 (0.079)	0.014 (0.013)	**−0.359** (**0.129**)	**0.068** (**0.024**)
General	**−0.395** (**0.094**)	−0.003 (0.016)	−0.215 (0.166)	**0.065** (**0.027**)
Specific	**0.758** (**0.095**)	**−0.049** (**0.017**)	0.130 (0.183)	**−0.170** (**0.028**)
Promotional	−0.121 (0.127)	−0.016 (0.024)	0.339 (0.229)	−0.056 (0.040)
$n/10$	**0.622** (**0.133**)	**−0.091** (**0.026**)	**1.806** (**0.317**)	
$n^2/100$	0.101 (0.096)	−0.005 (0.021)	−0.287 (0.364)	
Avg_Click				**0.005** (**0.000**)
Avg_CPC				**0.025** (**0.002**)
N_Potential				**0.049** (**0.023**)
N_Similar				**0.048** (**0.012**)

Notes. Posterior mean (posterior standard deviation) is reported for each parameter. Estimates that are significant at 95% are in bold.

through the mean transformed decay factor (λ_k), the mean log value per click (V_k), or both. Third, explicitly modeling the number of entrants (n_k) based on the revenue-cost approach enables us to infer the keyword-specific management cost, which cannot be inferred from the reduced-form approach. Because of these differences, we cannot directly compare the estimates from these two models. However, all significant effects of keyword attributes on the baseline click volume and decay factor reported from the reduced-form model are also significant and have the same signs in Table 5. Furthermore, we find positive effects of the number of entrants on baseline click volume in both models. The negative squared-term effect of the number of entrants on the mean decay factor suggests that as the number of entrants becomes large, the overall effect of n_k is the same as what is found in the reduced-form model. As for the effect of n_k on CPC, we show later in §5.1 that the simulated CPC based on estimation results from the proposed model is positively correlated with n_k. These findings support the internal validity of our proposed model.

We next discuss our findings based on the proposed structural model on click volume, CPC, and the number of entrants. Since one of the main objectives of this research is to understand how competition affects click volume and CPC through three underlying constructs—baseline click volume, mean decay factor, and mean value per click—we begin with a summary of findings on these relationships. (i) The number of entrants positively affects baseline click volume. (ii) The number of entrants has an inverse-U relationship with mean decay factor. (iii) The number of entrants has a negative and convex relationship with the mean value per click of a keyword.

We first explain the effects of competition on baseline click volume, which measures the level of consumer interest in clicking on ads related to a keyword. Intuitively, a larger number of paid-search ads on the result page are likely to navigate consumers' attention to the search listings and therefore encourage consumers to start clicking on ads. In some sense, this logic is similar to the effect of assortment size on the consumers' purchase intention in a retailing setting. Consistent with this conjecture, we find a positive competition effect on baseline click volume from estimation results of both reduced-form and structural models. Next we explain the competition effect on mean decay factor. When more ads are displayed on the result page, the higher marginal search benefit from a larger n_k encourages consumers to engage in a more in-depth search behavior and thus leads to a larger mean decay factor. However, when the number of ads becomes too large, it becomes costly for consumers to click through multiple ads from top to bottom. Thus, consumers are less likely to reach ads at lower positions on the paid-search listings, and the mean decay factor declines accordingly. Finally, the intuitive explanation for the negative

Table 6 **Variance Covariance Estimates from the Proposed Structural Model**

Panel A: Standard deviations

Variable	Mean	SD
Click volume: SD of ε (σ_q)	**1.135**	**0.010**
Decay factor: SD of ς (σ_d)	**1.299**	**0.009**
Value per click: SD of $\ln(S_{ki})$ (σ_{vpc})	**0.464**	**0.005**
Profit coeff. in utility of entry (β)	**12.493**	**2.158**

Panel B: Variance–covariance-matrix

Ω	η_1	η_2	η_3	η_4
η_1	**0.441**	**−0.219**	**−0.061**	**1.145**
	(0.028)	**(0.023)**	**(0.013)**	**(0.078)**
η_2		**0.218**	**0.043**	**−0.727**
		(0.020)	**(0.010)**	**(0.055)**
η_3			**0.075**	**−0.070**
			(0.007)	**(0.034)**
η_4				**4.189**
				(0.320)

Note. Posterior mean (posterior standard deviation) is reported for each parameter. Estimates that are significant at 95% are in bold.

Table 5 **Estimation Results from the Proposed Structural Model**

Keyword attributes	Log baseline click volume b_k	Mean transformed decay factor (λ_k)	Mean log value per click (V_k)	Management cost (in 100 cents) (F_k)
Intercept	**2.326 (0.039)**	**2.682 (0.078)**	**3.596 (0.041)**	**4.465 (0.105)**
DV	**−0.327 (0.060)**	0.044 (0.049)	0.014 (0.031)	**−0.965 (0.230)**
Accessory	**−0.443 (0.034)**	**0.186 (0.052)**	−0.005 (0.032)	**−1.523 (0.162)**
Coverage	**0.713 (0.092)**	**−0.260 (0.048)**	**−0.349 (0.044)**	**0.979 (0.345)**
Length	**−0.934 (0.076)**	**0.254 (0.059)**	**0.238 (0.040)**	**−2.437 (0.265)**
Brand	−0.056 (0.041)	**0.130 (0.032)**	**−0.106 (0.028)**	**−0.445 (0.135)**
General	**−0.457 (0.041)**	**0.173 (0.060)**	0.004 (0.032)	**−1.417 (0.200)**
Specific	**0.290 (0.048)**	**−0.101 (0.031)**	**−0.055 (0.020)**	**0.853 (0.202)**
Promotional	**−0.387 (0.053)**	**0.282 (0.068)**	0.074 (0.040)	**−0.939 (0.254)**
$n/10$	**0.368 (0.079)**	**0.449 (0.054)**	**−0.330 (0.035)**	
$n^2/100$	0.054 (0.035)	**−0.315 (0.042)**	**0.097 (0.041)**	

Notes. Posterior mean (posterior standard deviation) is reported for each parameter. Estimates that are significant at 95% are in bold.

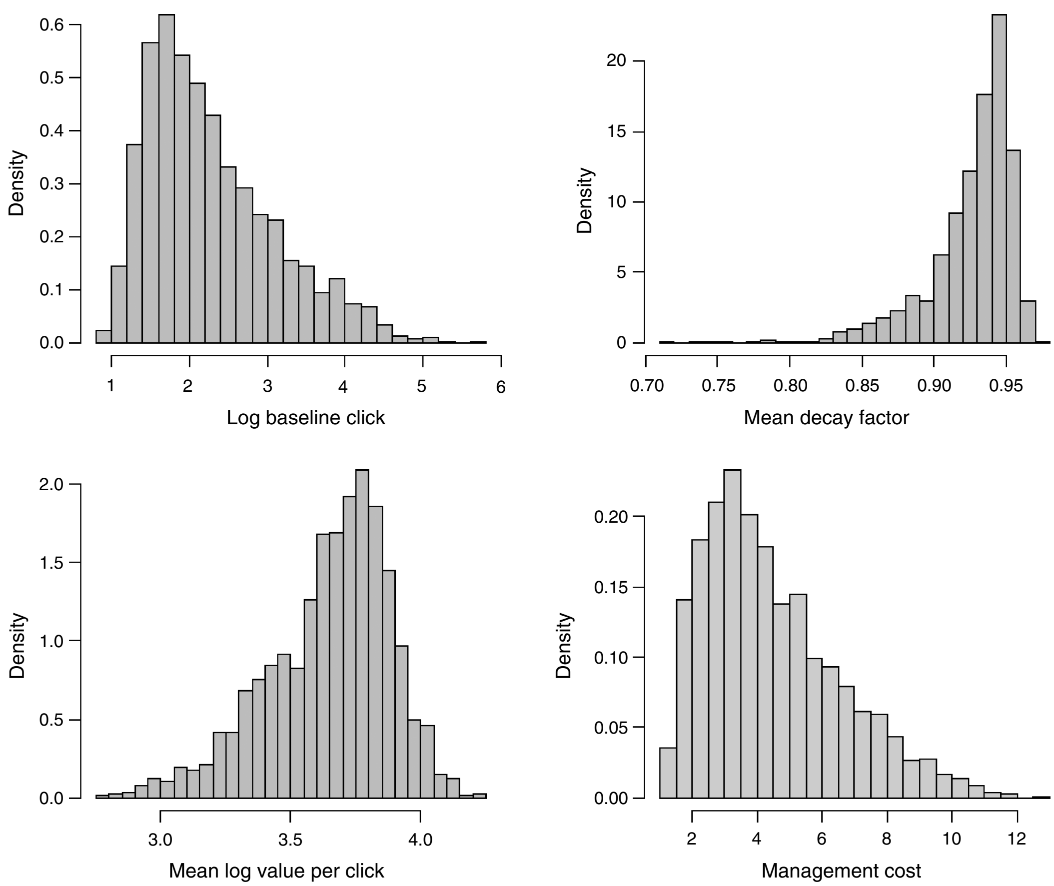
relationship between n_k and mean value per click is straightforward: when more competitors choose to advertise through the same keyword, an advertiser may expect a lower conversion probability and therefore the lower value from a click. Besides, the small but positive squared-term effect implies that the magnitude of the negative effect of n_k on mean value per click diminishes with n_k.

We next report the effects of keyword characteristics on the four parameters we are particularly interested in: baseline click volume, mean decay factor, mean value per click, and management cost. We start with our findings on the baseline click volume: (i) The baseline click volume is higher for keywords affiliated with more subcategories and negatively associated with keyword length. (ii) Keywords containing general descriptions tend to receive less baseline click volume, whereas keywords pointing to a specific product tend to receive more baseline click volume. One possible explanation is that most consumers shopping from this online market maker are knowledgeable about general product features associated with camera. Therefore, they are less likely to search for these general keywords, leading to less baseline click volume. On the other hand, those specific keywords tend to generate more demand. (iii) Keywords including promotional terms tend to receive less baseline click volume. This might be because the segment of price-sensitive consumers is relatively small in the category of digital-related products.

We next summarize our findings on the mean decay factor: (i) The mean decay factor is lower for keywords belonging to more subcategories and higher for longer keywords. (ii) Keywords with brand or general descriptions are associated with a higher mean decay factor, whereas keywords including words that are exclusive to the product tend to have a lower mean decay factor. This is consistent with our previous explanation for the finding on baseline click volume: consumers who are less knowledgeable about camera-related products tend to search more. (iii) Keywords with promotional terms tend to have a higher mean decay factor. One possible reason is that price-sensitive consumers are likely to search deeper along the ad listings to compare more products.

As for the effect of keyword characteristics on advertisers' mean value per click and management cost, we find that short keywords or keywords affiliated with multiple subcategories not only provide less

Figure 1 **Histogram of the Posterior Mean of Four Key Parameters Across Keywords (Log Baseline Click, Mean Decay Factor, Mean Log Value per Click, and Management Cost)**



per-click value to advertisers but are also associated with higher management cost. On the other hand, keywords containing fewer specific terms are associated with higher mean value per click and lower management cost.

In addition to the observed heterogeneity, we also find significant unobserved heterogeneity in these key model parameters. Figure 1 provides histograms of the posterior mean estimates of four key parameters across keywords. As shown, there is a substantial variation among keywords, suggesting a different demand pattern, CPC level, and entry cost associated with different keywords. The covariance estimates suggest two things. First, there are significant unobserved covariations among baseline click volume, mean decay factor, and mean value per click. For example, our results suggest that keywords with higher baseline click volume tend to have lower mean decay factor and mean value per click. Second, the negative correlation between keyword-specific unobserved terms in entry cost and mean value per click suggests a positive correlation between the number of entrants and mean value per click across keywords. This finding is consistent with the intuition that keywords with a higher mean value per click may attract more advertisers to enter. It also provides evidence of the number of entrants being endogenous in the CPC equation.

5. Managerial Implications

5.1. Implications for Paid-Search Advertisers

It is important for paid-search advertisers to understand how competition influences their own expected revenue from a keyword. For both advertisers who are already competing in a keyword-specific market and those who can potentially enter, the decisions of whether to continue the ad campaign and whether to start advertising through a keyword both depend on the evaluation of advertising revenue as a result of the number of entrants. Since an advertiser's expected revenue depends on expected click volume, CPC, and value per click, one advantage of our proposed structural model is that it can be used to infer the mean value per click of a keyword, and therefore it can help advertisers predict how the expected revenue changes with competition.

We simulate an advertiser's expected click volume, CPC, and advertising revenue in three keywords in Figure 2 based on the posterior distribution of keyword-specific baseline click volume, mean decay factor, and mean value per click. The three keywords, "Casio h10," "Sony w390," and "Fujifilm j38," are chosen based on the estimates of their mean decay factor, baseline click volume, and mean log value per click. After controlling for the effect of competition, the three keywords share a similar baseline click volume and mean log value per click, whereas the posterior means of their mean transformed decay factors are 1.227 (Casio h10), 1.978 (Sony w390), and 2.527 (Fujifilm j38), respectively.

The implications for advertisers based on the simulation results in Figure 2 can be summarized as follows: (i) Advertisers' expected revenue is generally a decreasing and convex function of the number of entrants. This negative effect of competition on the expected revenue of advertisers is consistent with the assumption made in our model setup. (ii) Advertisers' expected click volume is also generally a decreasing function of the number of entrants. (iii) Advertisers' expected CPC is in a positive and concave relationship with the number of entrants. (iv) Conditional on the competition, an advertiser's expected revenue is positively associated with the mean decay factor since the expected click volume is higher, whereas the expected CPC is lower in keywords with a larger mean decay factor.

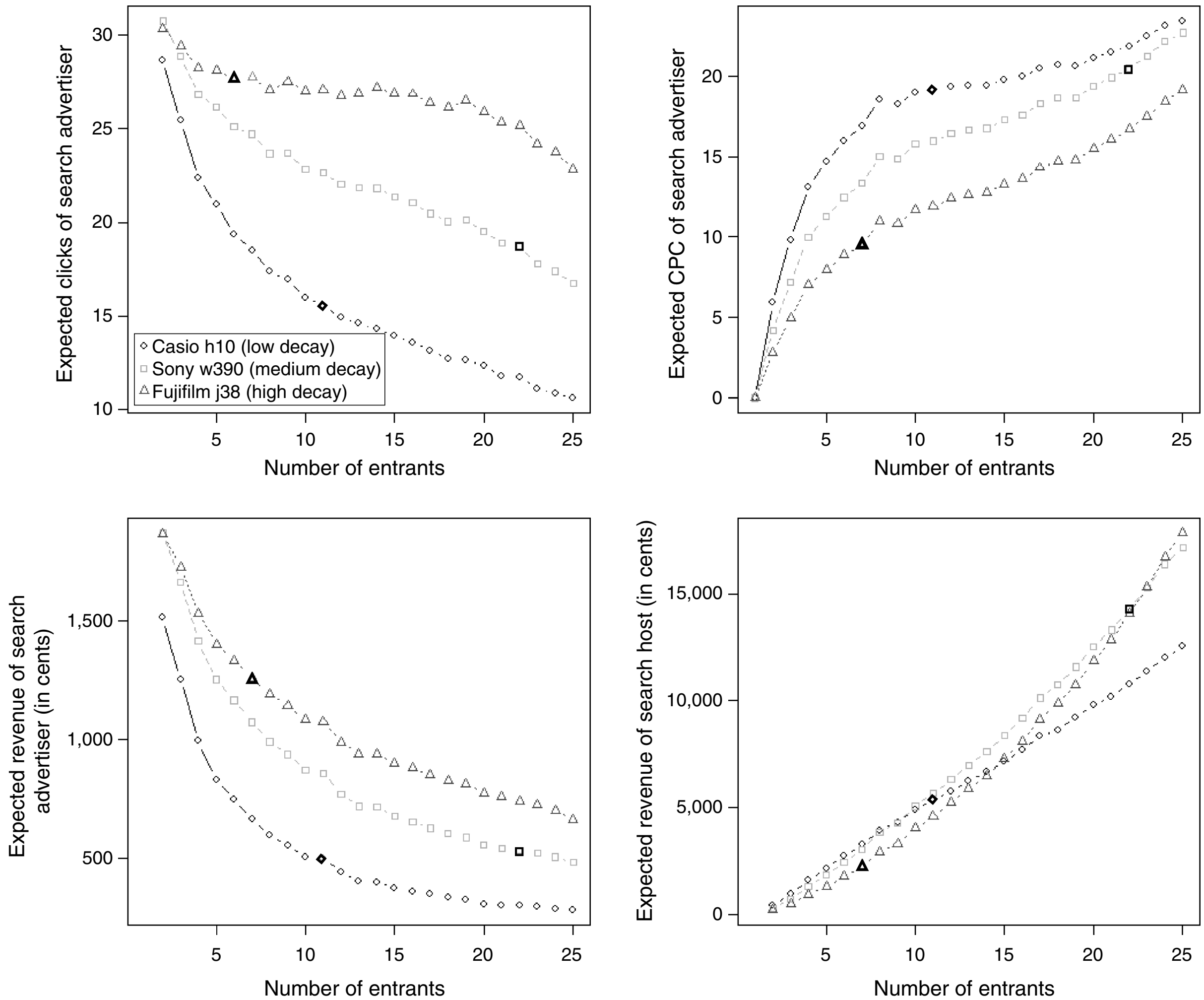
5.2. Implications for the Paid-Search Host

The structural nature of our proposed model can provide many insights for the paid-search host. As shown in Figure 2, in contrast to the advertiser, the paid-search host is interested in the market expansion for each keyword because its expected revenue is positively related to the number of entrants. Although the paid-search host does not have direct control over the number of entrants, our proposed model suggests two approaches for the paid-search host to attract advertisers. We conduct two counterfactual experiments to illustrate.

5.2.1. Counterfactual Experiment 1. We study how the paid-search host can improve profit by changing the decay factor in all keywords. Specifically, we allow the intercept of mean transformed decay factor—that is, the baseline mean transformed decay factor—to be changed from the current level by {−20%, −15%, −10%, −5%, 0%, 5%, and 10%}.[5] Since the mean decay factor is determined by consumer search patterns, the paid-search host is able to influence it through changing the page design. For example, the paid-search host can increase the mean decay factor by allowing advertisers to upload more content in their ads, which in turn could increase consumer search benefit, or setting a tighter space constraint for all ads shown on the paid-search results to lower consumer search cost.

[5] Given that the estimated baseline mean decay factor is already large (0.93), we restrict the upper bound of change in the baseline mean transformed decay factor to be 10% in this counterfactual experiment.

Figure 2 The Relationship Between the Number of Entrants and Four Key Advertising Metrics in Three Keywords (Expected Clicks, CPC, Revenue of Search Advertiser, and Expected Revenue of Search Host)



Note. The bold points refer to the realized number of entrants in each keyword.

Our proposed model suggests that conditional on the competition (i.e., the number of entrants), the click-volume decay factor influences the revenue of the paid-search host in two opposing ways: (i) A larger mean decay factor tends to increase the total click volume to the search host. (ii) A larger mean decay factor tends to lower average CPC paid by advertisers. Therefore, it is not straightforward whether the paid-search host should increase or decrease the baseline mean decay factor to improve its profit. Moreover, this question is further complicated by the fact that the change in baseline mean decay factor also affects the number of entrants at equilibrium, since the mean decay factor can influence the expected revenue of an advertiser. Therefore, it is difficult for the paid-search host to anticipate the overall impact of change in baseline mean decay factor on its profit. Our proposed model can help the paid-search host identify the right direction to influence the baseline mean decay factor through appropriate website design.

We next outline how we predict the paid-search host's revenue at equilibrium under different values of baseline mean decay factor. As the baseline mean decay factor changes, the advertiser's expected revenue of entry will change. Thus, the equilibrium number of entrants determined by Equation (16) will also change. After the equilibrium number of entrants is adjusted, the realized mean decay factor in each keyword is determined. We then predict click volume and CPC under this new equilibrium. Finally, the paid-search host's expected revenue in each keyword can be computed at this new equilibrium.

We report the paid-search host's average expected revenue as well as the average number of entrants across keywords under each policy change over the baseline decay factor in Table 7. We compare the counterfactual results with the paid-search host's average revenue and the average number of entrants at the current stage based on the real data. Our result shows that the paid-search host's revenue can be

Table 7 Counterfactual Results of Adjusting the Baseline Mean Decay Factor

	Percentage of change of baseline mean transformed decay factor (intercept in λ_k)						
	−20	−15	−10	−5	0	5	10
Average expected n across keywords	6.37	6.95	7.67	8.55	9.10	10.55	12.66
Average expected revenue of search host across keywords (in 100 cents)	17.42	19.35	21.60	24.44	27.09	33.48	40.94

improved substantially with an increase in the baseline mean decay factor. For example, the expected revenue of the search host increases by 51% if the baseline mean transformed decay factor is 10% higher than its current level.[6] Moreover, such a policy change increases the average number of entrants in keywords by 39%. Thus, it generates a net benefit to the paid-search host by offsetting the potential loss from a smaller average CPC as a result of a larger mean decay factor. Our result suggests that the paid-search host can improve its revenue by shifting advertisers' revenue function upward through the increase in baseline mean decay factor, but consequently, it intensifies the competition among advertisers.

One caveat of this counterfactual analysis is that a design change could reduce consumers' repeated usage of the paid-search host, which will have a negative effect on the search host's revenue. However, this particular counterfactual experiment is still meaningful because it provides an illustration of the effect of the changing mean decay factor (via change of search listing designs) on the paid-search revenue. The paid-search host needs only to identify the boundary conditions under which a design change will not significantly affect consumer loyalty and select the appropriate design (associated with a certain baseline mean decay factor) within this range to maximize its profit.

5.2.2. Counterfactual Experiment 2. We next demonstrate how our study helps the paid-search host select the right face value of coupons to draw more advertisers into search advertising and improve profit. Specifically, the paid-search host provides a coupon with common face value to advertisers when they launch ad campaigns in any keyword for the first time.[7] The goal of offering such coupons is to encourage more advertisers to enter the market. Each advertiser can redeem only one coupon for each keyword. In practice, Google has provided a similar coupon system for new customers when they sign up with Google AdWords. Our experiment selects the optimal entry coupon from five face values {10%, 20%, 30%, 40%, and 50%} of the average advertisers' management cost across keywords. The mechanism works in the following way. Given a face value of the coupon, we predict the equilibrium number of entrants of each keyword. Then the number of entrants at this new equilibrium will affect the baseline click volume, mean decay factor, and mean value per click, through which the realized click volume and CPC are also changed. In a manner similar to that used in the first counterfactual experiment, we can then calculate the paid-search host's expected revenue at the new equilibrium induced by a policy change.

Table 8 reports findings from this counterfactual experiment. We find that an appropriate selection of the face value of entry coupons (from 10% to 30% of average management cost) can increase the profit of the paid-search host from the current level. In the optimal situation, the paid-search host can improve its profit by 25% if it uses the entry coupon with a face value of 20% of average management cost. Compared with the alternative of influencing the click-volume decay factor through page design, the coupon system

Table 8 Counterfactual Results of Providing Entry Coupons

	Value of coupon in percentage to the average management cost					
	0	10	20	30	40	50
Average expected n across keywords	9.10	13.12	16.76	19.72	21.91	23.60
Average expected revenue of search host across keywords (in 100 cents)	26.98	38.79	48.90	57.94	66.13	76.03
Average expected cost of entry coupons across keywords (in 100 cents)	0.00	5.86	14.97	26.42	39.14	52.68
Average expected profit of search host across keywords (in 100 cents)	27.09	32.93	33.93	31.52	26.99	23.35

Note. The average management cost across keywords is 446.50 cents.

[6] We assume that the cost of changing the page design for the paid-search host is negligible.

[7] Although the paid-search host could design coupons with different values in different keywords or to different advertisers, it is less practical than the uniform coupon because of the potential discrimination issue. We therefore focus on the simple common-value coupon in this paper.

is under full control of the paid-search host and therefore is more practical for the search host to implement.

6. Conclusion

For paid-search advertisers to survive in a highly competitive market, it is crucial for them to understand the nature of competition and how it influences their profitability. As such, most paid-search information providers collect competition-related information, and a very important piece of this information is the number of competing ads on the same paid-search listings. Although advertisers have been using this measure as an indicator of competition intensity, they need more insights into how such a measure affects certain key factors determining their click volume and CPC. From the paid-search host's perspective, an in-depth understanding of the key determinants of competition among paid-search advertisers can help the paid-search host improve its profit through appropriate interventions and policy changes.

In this paper, we propose a structural framework to characterize competition and analyze its impact on click volume and CPCs through three key latent constructs: baseline click volume, decay factor, and value per click. We regard each keyword as a market and establish an integrative model of realized click volume, CPCs, and number of entrants. One of our key modeling contributions is that we link demand (i.e., click volume) and supply (i.e., CPCs) through the decay factor in a structural manner. We also structurally model the number of entrants as an aggregate equilibrium outcome based on a microlevel process.

One of the key findings from our analysis is that the number of entrants has significant impacts on baseline click volume, decay factor, and value per click, which determine the expected click volume, CPC, and revenue of search advertisers in a keyword. Therefore, our results suggest that advertisers should account for these effects in adjusting their bidding and keyword portfolio strategies with respect to competition. As for the paid-search host, our counterfactual analysis shows that the paid-search host can improve its profit by manipulating the click-volume decay factor or distributing coupons with the right face value to intensify the competition among advertisers.

Our paper has limitations that offer opportunities for future research. First, our modeling framework can be extended to a dynamic setting. A dynamic modeling framework will be especially powerful, for example, in examining advertisers' keyword portfolio composition and competition strategies over time. However, because of the relatively short time span of our data, where advertisers' keyword portfolio compositions do not change, we are not able to model the dynamic component of advertisers' keyword management strategies. Should longer-span data be available, researchers could enrich our modeling framework by incorporating dynamics. Second, our proposed model is applied to only one product category, and the generalizability of our empirical findings needs to be validated with more product categories. Third, because we do not observe how advertisers make the keyword entry decisions exactly, we assume the keyword entry of advertisers to be independent across keywords. Extending our model to allow joint keyword entry will be an interesting area for future research, even though we do not find strong evidence of such interdependent entry decisions across keywords in our empirical context.[8] Fourth, field experiments will be valuable to empirically test the effectiveness of our proposed policy changes in the real world. Notwithstanding these limitations, we hope that this study will generate further interest in exploring this important emerging area in marketing.

Acknowledgments
The authors thank the anonymous company that provided data for this study. They are grateful to the editor, associate editor, and two anonymous referees for extremely helpful comments. They also thank participants at the Marketing Science conference, the Annual UT Dallas FORMS Conference in 2012, as well as the seminar given at the University of Rochester for their valuable comments. The first author acknowledges the financial support from the National Natural Science Foundation of China [Grants 71128002 and 71090402]. The third author acknowledges the financial support from the National Natural Science Foundation of China [Grant 71172037].

Appendix A. Derivation of the CPC Equation and Proof of Equilibrium Properties

EOS found that there is a unique Bayes Nash equilibrium in the generalized English auction that corresponds to the GSP auction. More important, the realized bids of this generalized English auction satisfy the Nash equilibrium condition of the GSP auction. We first describe the rule of the generalized English auction. Suppose there is a price clock continuously increasing from zero, and each advertiser's decision is to choose at what price to drop out (i.e., this price will be her bid). The auction concludes when there is only one advertiser left. The vector of these bids is used to determine

[8] We tested the independence assumption in our empirical context in a reduced-form model following the idea of testing spatial autocorrelation (Anselin 1988). In the first step, we ran a reduced-form model by regressing the log of entry probability for keyword k (entry probability is defined as the number of entrants divided by the number of potential entrants) on keyword attributes. In the second step, we saved all the residuals obtained from the first step. In the third step, we constructed Moran's I statistic based on the residuals and a spatial contiguity matrix W (where the diagonal elements of W were set to zero, and the ijth element was equal to the total number of common attributes shared by keyword i and keyword j). The Moran's test statistic has a p-value of 0.157, which suggests that there is no significant unobservable interdependence in entry across keywords after controlling for the keyword attributes.

the rank and CPCs of ads based on the same rule of the GSP auction. As for the information structure, each advertiser knows her own value per click S_{ki} but only the distribution of other advertisers' valuations after their entry into keyword k. The vector of click ratio between adjacent positions $\{\delta_{kj}\}$ in a keyword is assumed to be common knowledge. Theorem 2 of EOS indicates that in the unique Bayes Nash equilibrium of this generalized English auction, an advertiser i drops out at price

$$p_{ki} = S_{ki} - \delta_{k,\,j_{ki}-1}(S_{ki} - b_{ki'}), \tag{A1}$$

where i' refers to the advertiser who drops right before advertiser i and thus stays right below advertiser i (i.e., $j_{ki'} = j_{ki}+1$). If we replace i and i' with i' and i'' (assume i'' is the one who stays below i'), Equation (A1) can be written as

$$p_{ki'} = S_{ki'} - \delta_{kj_{ki}}(S_{ki'} - b_{ki''}). \tag{A2}$$

EOS defines an advertisers' bid as the price at which she drops out from the auction, and therefore we have $p_{ki'} = b_{ki'} = \mathrm{CPC}_{ki}$ and $b_{ki''} = \mathrm{CPC}_{ki'}$. Equation (A2) can then be written as

$$\begin{aligned}\mathrm{CPC}_{ki} &= S_{ki'} - \delta_{kj_{ki}}(S_{ki'} - \mathrm{CPC}_{ki'})\\ &= \delta_{kj_{ki}}\mathrm{CPC}_{ki'} + (1-\delta_{kj_{ki}})S_{ki'},\end{aligned} \tag{A3}$$

which is exactly the same as Equation (7) in our model. We next formally show that this bidding equilibrium characterized by EOS has three important properties.

Property 1. *An advertiser with a higher value per click obtains a higher position.*

Proof of Property 1. Equation (A1) implies that given last advertiser's bid, the price at which each advertiser drops out increases with her value per click. Thus, among the remaining advertisers, the advertiser with the lowest value per click will drop out and obtain the lowest remaining position. This implies that at realized values per click, an advertiser with a higher value per click gets a higher position.

Property 2. *Equilibrium CPCs decline with positions, and each advertiser's CPC is below her value per click by construction.*

Proof of Property 2. We remove the subscript k for expositional convenience. Besides, with a slight abuse of notation, we denote advertiser i as the advertiser with the ith highest value per click among all n entrants. Property 1 implies that advertiser i stays at the ith position at equilibrium. We start the proof for the CPC of advertiser at the next to last position: $\mathrm{CPC}_{n-1} = \delta_{n-1}\mathrm{CPC}_n + (1-\delta_{n-1})S_n$, where CPC_n equals the minimum bid set by the search host and therefore $\mathrm{CPC}_n < S_n$. Then we have $\mathrm{CPC}_{n-1} - \mathrm{CPC}_n = (1-\delta_{n-1})(S_n - \mathrm{CPC}_n) > 0$. At the same time, we have $S_{n-1} - \mathrm{CPC}_{n-1} = (S_{n-1} - S_n) + \delta_{n-1}(S_n - \mathrm{CPC}_n) > 0$. Next, we prove that as long as $\mathrm{CPC}_i < S_i < S_{i-1}$, the relationship $\mathrm{CPC}_i < \mathrm{CPC}_{i-1} < S_{i-1}$ always holds. Similar to the analysis for the case when $i = n$, these two inequalities hold because $\mathrm{CPC}_{i-1} - \mathrm{CPC}_i = (1-\delta_{i-1})(S_i - \mathrm{CPC}_i) > 0$ and $S_{i-1} - \mathrm{CPC}_{i-1} = (S_{i-1} - S_i) + \delta_{i-1}(S_i - \mathrm{CPC}_i) > 0$. Thus, we have shown that $\mathrm{CPC}_i < \mathrm{CPC}_{i-1} < S_{i-1}$ holds for any $i = 2, \ldots, n$, which completes our proof.

Property 3. *The unique Bayes Nash equilibrium in the generalized English auction is stable in the sense that the realized bids form a Nash equilibrium of the GSP auction. In other words, no advertiser will regret her bid after she knows others' values per click.*

Proof of Property 3. We follow the notations used in the proof of property 2. As EOS pointed out, to prove that the realized bids form a Nash equilibrium, it is sufficient to show that the vector of equilibrium bids is "locally envy-free." *Locally envy-free* is a refined Nash equilibrium concept defined by EOS that requires that no advertiser has incentive to exchange bids with the advertiser who ranks above. We show that the vector of bids characterized by Equation (A1) satisfies this condition. For any advertiser $i = 2, \ldots, n$, the profit for advertiser i at the ith position is $\pi_i(b_i) = E(Q_i)(S_i - b_{i+1})$, and the profit for advertiser i at the $i-1$th position after exchanging bids with advertiser $i-1$ is $\pi_i(b_{i-1}) = E(Q_{i-1})(S_i - b_i)$. Since $b_i = S_i - \delta_{i-1}(S_i - b_{i+1})$ from Equation (A1), and $\delta_{i-1} = E(Q_i)/E(Q_{i-1})$ by definition, we have $\pi_i(b_i) = (E(Q_i)/\delta_{i-1})(S_i - b_i) = \pi_i(b_{i-1})$. Thus, we have proved that the vector of realized bids is a stable locally envy-free Nash equilibrium.

Appendix B. Expected Revenue Function of Search Advertisers

Based on Equations (2) and (14), we have a revenue function of search:

$$\begin{aligned}\Pi_k(n_k) &= E_{Q,S,\lambda}\big\{Q\big(b_k(n_k), \lambda_k(n_k)\big)\big[S(V_k(n_k)) - \mathrm{CPC}\big(S_k, \lambda_k(n_k)\big)\big]\big\}\\ &= E_{Q,S,\lambda}\Bigg\{\exp(b_k + \varepsilon_{ki})\prod_{l=1}^{j_{ki}-1}\delta_{kl}\big[S_{ki}(V_k) - \mathrm{CPC}_{ki}\big(\{S_{km}\}_{j_{km}>j_{ki}}, \{\delta_{km}\}_{m=j_{ki}}^{n_k-1}\big)\big]\Bigg\}\\ &= \exp(b_k + 0.5\sigma_q^2)E_{S,\lambda}\Bigg\{\prod_{l=1}^{j_{ki}-1}\delta_{kl}\big[S_{ki}(V_k) - \mathrm{CPC}_{ki}\big(\{S_{km}\}_{j_{km}>j_{ki}}, \{\delta_{km}\}_{m=j_{ki}}^{n_k-1}\big)\big]\Bigg\},\end{aligned} \tag{B1}$$

where $\delta_{kl} = \exp(\lambda_{kl})/(1+\exp(\lambda_{kl}))$, and we define $\prod_{l=1}^{0}\delta_{kl} = 1$ with a slight abuse of notation.

Under the assumption of symmetrical advertisers, the post-entry ad position is uniformly distributed; that is, $\Pr(j_{ki} = l) = 1/n_k$, $l = 1, \ldots, n_k$. Thus, the expected revenue function can be rewritten as

$$\Pi_k(n_k) = \frac{\exp(b_k + 0.5\sigma_q^2)}{n_k}E_{S,\lambda}\Bigg\{\sum_{j=1}^{n_k}\Bigg[\prod_{l=1}^{j-1}\delta_{kl}\big(S_{kj}(V_k) - \mathrm{CPC}_{kj}(\{S_{k,m+1}, \delta_{km}\}_{m=j}^{n_k-1})\big)\Bigg]\Bigg\}, \tag{B2}$$

where S_{kj} and CPC_{kj} stand for the value per click and CPC, respectively, of the advertiser who stays at position j at equilibrium.

Equation (7) suggests that the equilibrium CPC at each position can be expressed as a weighted summation of values per click of advertisers who ranked below, which is

$$\mathrm{CPC}_{kj} = \sum_{l=j+2}^{n_k}\Bigg[(1-\delta_{k,l-1})S_{kl}\prod_{m=l-2}^{j}\delta_{km}\Bigg] + \prod_{l=j}^{n_k-1}\delta_{kl}\,\mathrm{CPC}_0 + (1-\delta_{kj})S_{k,j+1}, \tag{B3}$$

where CPC_0 stands for the CPC at the bottom position, which is treated as an exogenous variable. Equations (B2) and (B3) fully characterize the expected revenue function of advertisers. Note that we are unable to analytically integrate out two sets of error terms, which are $\{\delta_{kj}\}_{j=1}^{n_k-1}$ and $\{S_{kj}\}_{j=1}^{n_k}$. We therefore rely on a simulation-based approach. First, we create two sets of random samples, $\mathbf{e_1}$ and $\mathbf{e_2}$, generated from an i.i.d. standard normal distribution, where $\mathbf{e_1}$ is a (n_k-1) by R matrix, and $\mathbf{e_2}$ is a n_k by R matrix. Here, R stands for the number of random draws used for integration. For each random draw r, we first use $e_1^{(r)}$ to generate the vector of decay factors:

$$\delta_{kj}^{(r)} = \frac{\exp(\lambda_k + \sigma_d e_{1j}^{(r)})}{1+\exp(\lambda_k + \sigma_d e_{1j}^{(r)})}. \tag{B4}$$

As for the vector of value per click, we first sort the vector of $e_2^{(r)}$ in descending order to create $enew_2^{(r)}$. Then we construct the value per click of the jth advertiser as follows:

$$S_{kj}^{(r)} = \exp\big(V_k + \sigma_{\text{vpc}}\, enew_{2j}^{(r)}\big) = \exp(V_k)\big[\exp(enew_{2j}^{(r)})\big]^{\sigma_{\text{vpc}}}. \tag{B5}$$

Combining Equations (B2)–(B5), we can fully specify the expected revenue function of an advertiser as

$$\Pi_k(n_k) = \frac{\exp(b_k+0.5\sigma_q^2)}{n_k} \cdot \sum_{r=1}^{R}\left\{\sum_{j=1}^{n_k}\left[\prod_{l=1}^{j-1}\delta_{kl}^{(r)}\left(S_{kj}^{(r)} - \sum_{l=j+2}^{n_k}\left[(1-\delta_{k,l-1}^{(r)})S_{kl}^{(r)}\prod_{m=l-2}^{j}\delta_{km}^{(r)}\right] + \prod_{l=j}^{n_k-1}\delta_{kl}^{(r)}\text{CPC}_0 + (1-\delta_{kj}^{(r)})S_{k,j+1}^{(r)}\right)\right]\right\}. \tag{B6}$$

Note that the keyword-specific parameters including log baseline click volume (b_k), mean transformed decay factor (λ_k), and mean log value per click (V_k) are all functions of number of entrants. Thus, the expected revenue of advertisers is ultimately a function of the following parameters:

$$\Pi_k = f\big[n_k, b_k(n_k), \lambda_k(n_k), V_k(n_k), \sigma_q^2, \sigma_d^2, \sigma_{\text{vpc}}^2\big]. \tag{B7}$$

Appendix C. The MCMC Algorithm and Simulation Studies

We estimate the model parameters using the Bayesian inference method. We run the Markov chains for 100,000 iterations, where the first 50,000 draws are discarded as the initial burn-in. We kept every 100th draw from the last 50,000 draws for inference. We use 10 draws of the vector of value per click and the vector of decay factors to simulate advertisers' expected revenue in a keyword. Before we describe the prior and posterior distribution for each parameter, we first present the posterior distribution for all parameters:

$$\Pr(\Theta \mid Q, \text{CPC}, n) \propto \prod_{k=1}^{K}\Big[L_k^q(Q\mid\lambda,\gamma_1,\eta_{k1},\sigma_q^2)L_k^{cpc}(\text{CPC}\mid\lambda,\gamma_3,\eta_{k3},\sigma_{vpc}^2) \cdot L_k^n(n\mid\eta,\gamma,\sigma^2,\beta)\Pr(\lambda\mid\gamma_2,\eta_{k2},\sigma_d^2)\Pr(\eta\mid\Omega)\Big], \tag{C1}$$

where $\Theta=\{\eta,\lambda,\gamma,\beta,\sigma^2\}$ is the set of all model parameters; $L_k^q(\cdot)$, $L_k^{cpc}(\cdot)$, and $L_k^n(\cdot)$ are likelihood functions of the click volume vector, vector of observed CPCs, and number of entrants in keyword k, respectively; $\Pr(\lambda)$ is the prior probability for the vector of decay factors $\vec{\lambda}_k = (\lambda_{k1},\ldots,\lambda_{k,n_k-1})$; and $\Pr(\eta)$ is the marginal distribution of $(\eta_{k1},\eta_{k2},\eta_{k3})$.

Next we characterize each of these functions. The likelihood function of viewing the vector of click volume in keyword k is $L_k^q(\cdot)$:

$$L_k^q = (2\pi\sigma_q^2)^{-(n_k/2)} \cdot \exp\left[-\sum_{j=1}^{n_k}\left(\ln(Q_{kj}) - \sum_{l=1}^{j-1}\ln(\delta_{kl}) - \gamma_2 X_k - \eta_{k2}\right)^2 (2\sigma_q^2)^{-1}\right], \tag{C2}$$

where

$$\delta_{kj} = \frac{\exp(\lambda_{kj})}{1+\exp(\lambda_{kj})},$$

and we define $\sum_{l=1}^{0}\ln(\delta_{kl}) = 0$ with a slight abuse of notation.

For expositional convenience, we index advertisers in a keyword by their equilibrium positions (i.e., S_{kj} stands for the value per click of advertiser i at the jth position). Then $L_k^{\text{cpc}}(\cdot)$ can be described as

$$L_k^{\text{cpc}} = \prod_{j=2}^{n_k}\left\{\frac{\exp[-(\ln S_{kj} - \gamma_3 X_k - \eta_{k3})^2(2\sigma_{\text{vpc}}^2)^{-1}]}{(2\pi\sigma_{\text{vpc}}^2)^{1/2}}\right\} \cdot I(S_{k2} > S_{k3} > \cdots > S_{kn_k})\left|\frac{\partial S_k}{\partial\,\text{CPC}_k}\right|, \tag{C3}$$

where $S_{kj} = (\text{CPC}_{k,j-1} - \delta_{k,j-1}\text{CPC}_{kj})/(1-\delta_{k,j-1})$ is derived from Equation (7), and $I(S_{k2} > S_{k3} > \cdots > S_{kn_k})$ is an indicator function that equals one if the vector of value per click is in a descending order; $\partial S_k/\partial\,\text{CPC}_k$ is a (n_k-1) by (n_k-1) Jacobian matrix, which is an upper-triangle matrix with $\{1/(1-\delta_{k,j-1})\}_{j=2}^{n_k}$ on the diagonal. Thus, the likelihood function of the CPC vector can be further expressed as

$$L_k^{\text{cpc}} = \prod_{j=2}^{n_k}\left\{\frac{\exp[-(\ln S_{kj} - \gamma_3 X_k - \eta_{k3})^2(2\sigma_{\text{vpc}}^2)^{-1}]}{(2\pi\sigma_{\text{vpc}}^2)^{1/2}(1-\delta_{k,j-1})}\right\} \cdot I(S_{k2} > S_{k3} > \cdots > S_{kn_k}). \tag{C4}$$

Combining Equations (17) and (19), we have

$$L_k^n = \Pr(F_k \mid \eta_{k1},\eta_{k2},\eta_{k3})\left|\frac{\partial F_k}{\partial n_k}\right| = \Pr(\eta_{k4}\mid\eta_{k1},\eta_{k2},\eta_{k3})\left|\frac{\partial F_k}{\partial n_k}\right|, \tag{C5}$$

where the prior distribution of F_k is $F_k = \gamma_4 X_k^F + \eta_{k4}$ from Equation (13), and the description of F_k is $F_k = \Pi_k + (1/\beta)[\ln(N_k-n_k) - \ln(n_k)]$ from Equation (17). As shown in Equation (B7), advertiser's expected revenue Π_k depends on all parameters except for β. However, because F_k is a function of both Π_k and β, all model parameters enter into the likelihood function of number of entrants. Combining Equation (C5) with the marginal distribution function $\Pr(\eta_{k1},\eta_{k2},\eta_{k3}\mid\Omega)$, we have

$$L_k^n(n_k)\Pr(\eta\mid\Omega) = \Pr(\eta_{k4}\mid\eta_{k1},\eta_{k2},\eta_{k3})\Pr(\eta_{k1},\eta_{k2},\eta_{k3}\mid\Omega)\left|\frac{\partial F_k}{\partial n_k}\right| \propto \exp\left[-\frac{1}{2}\eta_k'\Omega^{-1}\eta_k\right]\left|\frac{\partial F_k}{\partial n_k}\right|, \tag{C6}$$

where $\eta_k = (\eta_{k1}, \eta_{k2}, \eta_{k3}, \eta_{k4})'$. Finally, the prior of decay factors can be expressed as follows:

$$\Pr(\lambda) = \prod_{j=1}^{n_k-1} \left\{(2\pi\sigma_d^2)^{-(1/2)} \exp\left[-(\lambda_{kj} - \gamma_2 X_{k2} - \eta_{k2})^2 (2\sigma_d^2)^{-1}\right]\right\}. \tag{C7}$$

Step 1 (Draw λ_{kj} and update S_{kj}). We use the Metropolis–Hastings algorithm with a random walk chain to generate draws (see Chib and Greenberg 1995). Since the error terms in Equations (2) and (7) are independent across keywords, we draw the vector of λ_{kj} for each keyword separately. The λ_{kj} in the same keyword k is updated simultaneously during each iteration. Recall that $\vec{\lambda}_k$ is referred to a $n_k - 1$ by 1 vector of $(\lambda_{k1}, \ldots, \lambda_{k, n_k-1})$. Let $\vec{\lambda}_k^{(p)}$ denote the previous draw; then the next draw, $\vec{\lambda}_k^{(n)}$, is given by the following:

$$\vec{\lambda}_k^{(n)} = \vec{\lambda}_k^{(p)} + \Delta_1$$

with acceptance probability

$$\min\left\{\frac{L_k^q(\vec{\lambda}_k^{(n)}) L_k^{cpc}(\vec{\lambda}_k^{(n)}) \Pr(\vec{\lambda}_k^{(n)})}{L_k^q(\vec{\lambda}_k^{(p)}) L_k^{cpc}(\vec{\lambda}_k^{(p)}) \Pr(\vec{\lambda}_k^{(p)})}, 1\right\},$$

where Δ_1 is a draw from the density $N(0, 0.0025I)$, and I is the identity matrix. To ensure the ordered structure of values per click, we consider a candidate draw of decay factors only if the vector of S_{kj} defined by $S_{kj} = (\text{CPC}_{k,j-1} - \delta_{k,j-1}\text{CPC}_{kj})/(1 - \delta_{k,j-1})$ satisfies $S_{k2} > S_{k3} > \cdots > S_{kn_k}$. Finally, given an accepted vector of λ_{kj}, we update the vector of S_{kj} accordingly.

Step 2 (Draw $\theta = (\gamma_1, \gamma_2, \gamma_3, \sigma_q^2, \sigma_d^2, \sigma_{\text{vpc}}^2, \beta)$). We use the Metropolis–Hastings algorithm with a random walk chain to generate draws of θ. We transform the variance parameters $\sigma_q^2, \sigma_d^2, \sigma_{\text{vpc}}^2$ to their log forms in the implementation of the random walk chain. Let $\theta^{(p)}$ denote the previous draw; then the next draw $\theta^{(n)}$ is given by the following:

$$\theta^{(n)} = \theta^{(n)} + \Delta_2$$

with the acceptance probability

$$\min\left\{\frac{\prod_{k=1}^K L_k^q(\theta^{(n)}) L_k^{cpc}(\theta^{(n)}) L_k^n(\theta^{(n)}) \Pr(\vec{\lambda}_k \mid \theta^{(n)})}{\prod_{k=1}^K L_k^q(\theta^{(p)}) L_k^{cpc}(\theta^{(p)}) L_k^n(\theta^{(p)}) \Pr(\vec{\lambda}_k \mid \theta^{(p)})}, 1\right\},$$

where Δ_2 is a draw from the density $N(0, 0.0001I)$.

Step 3 (Draw γ_4). We first derive the conditional distribution of η_{k4}, given η_{k1}, η_{k2}, and η_{k3}.

Denote

$$(\eta_{k4}, \eta_{k1}, \eta_{k2}, \eta_{k3})' \sim \text{MVN}\left(0, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

where

$$\Sigma_{11} = \Omega_{44}, \ \Sigma_{12} = (\Omega_{14}, \Omega_{24}, \Omega_{34}),$$

$$\Sigma_{21} = \Sigma_{12}', \ \Sigma_{22} = \begin{pmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{pmatrix}.$$

Then we have

$$\eta_{k4} \mid \eta_{k1}, \eta_{k2}, \eta_{k3} \sim N(\bar{\mu}_{k4}, \bar{\sigma}_4^2),$$

where $\bar{\mu}_{k4} = \Sigma_{12}\Sigma_{22}^{-1}(\eta_{k1}, \eta_{k2}, \eta_{k3})'$, $\bar{\sigma}_4^2 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. We assume the prior distribution of γ_4 is

$$\gamma_4 \sim \text{MVN}(\gamma_0, A_0^{-1}).$$

Since $(F_k - \gamma_4 X_k^F) \sim N(\bar{\mu}_{k4}, \bar{\sigma}_4^2)$, the posterior of γ_4 can be given by the following:

$$\gamma_4 \mid F_k, X_k^F, \bar{\mu}_{k4}, \bar{\sigma}_4^2 \sim \text{MVN}[\tilde{\gamma}_4, (X_k^{F\prime} X_k^F \bar{\sigma}_4^{-2} + A_0)^{-1}],$$

where $\tilde{\gamma}_4 = (X_k^{F\prime} X_k^F \bar{\sigma}_4^{-2} + A_0)^{-1}[X_k^{F\prime}(F_k - \bar{\mu}_{k4})\bar{\sigma}_4^{-2} + A_0\gamma_0]$, $A_0 = 0.01I$, and $\gamma_0 = 0$. The entry cost F_k is calculated based on Equation (17).

Step 4 (Draw Ω). We assume the prior of Ω follows an inverted Wishart distribution:

$$\Omega \sim \text{IW}(w_0, W_0)$$

Then, the posterior can be given by the following:

$$\Omega \mid \eta_{k1}, \eta_{k2}, \eta_{k3}, \eta_{k4} \sim \text{IW}\left(w_0 + K, W_0 + \sum_{k=1}^K \eta_k \eta_k'\right),$$

where $\eta_k = (\eta_{k1}, \eta_{k2}, \eta_{k3}, \eta_{k4})'$, $w_0 = 10$, and $W_0 = 10I$.

Step 5 (Draw η_{k1}). Since the error terms in our model are assumed to be independent across keywords, we can draw η_{k1} for each keyword k separately. In our model, η_{k1} or b_k affects the vector of click volume and the number of entrants. Let $\eta_{k1}^{(p)}$ denote the previous draw; then the next draw $\eta_{k1}^{(n)}$ is given by the following:

$$\eta_{k1}^{(n)} = \eta_{k1}^{(p)} + \Delta_3$$

with acceptance probability

$$\min\left\{\frac{L_k^q(\eta_{k1}^{(n)}) L_k^n(\eta_{k1}^{(n)}) \Pr(\eta_{k1}^{(n)}, \eta_{k2}, \eta_{k3} \mid \Omega)}{L_k^q(\eta_{k1}^{(p)}) L_k^n(\eta_{k1}^{(p)}) \Pr(\eta_{k1}^{(p)}, \eta_{k2}, \eta_{k3} \mid \Omega)}, 1\right\},$$

where Δ_3 is a draw from the density $N(0, 0.25I)$.

Step 6. Draw η_{k2} similar to Step 5.

Step 7. Draw η_{k3} similar to Step 5.

We conduct a simulation study to test whether the proposed algorithm can recover the true parameters in the full model of click volume, CPC, and the number of entrants. We jointly simulate the data of click volume, CPC, and the number of entrants for 200 keywords. For convenience, we fix the CPC at the lowest position to be zero. The keyword characteristics, vector X_k, includes one intercept, one keyword covariate generated from a standard normal distribution, and the number of entrants (n_k) generated by searching for the unique solution of Equation (16). The explanatory variables (X_k^F) of the management cost in Equation (13) use the same keyword covariates in X_k except for the number of entrants. The true model parameters $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \sigma_q^2, \sigma_d^2, \sigma_{\text{vpc}}^2, \beta, \Omega)$ are reported in Table C.1.

We ran the Markov chains for 100,000 iterations and saved every 100th draw for the parameters of interest. We set the starting values for $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ as (0, 0, 0) and for the four diagonal elements in Ω as 0.1. The starting values for $\sigma_q^2, \sigma_d^2, \sigma_{\text{vpc}}^2, \beta$ are also set to be 0.1. We discard the first 50,000 draws and used the last 50,000 to calculate the posterior means and standard deviations of the parameters.

The estimation results are reported in Table C.1. It shows that all parameters of interest are identifiable and that their

Table C.1 Simulation Results of the Proposed Structural Model

Variable	True value	Estimate
γ_1 (baseline click)	2, 1, −3	1.907, 0.930, −3.046 (0.086) (0.061) (0.114)
γ_2 (decay factor)	2, 1, −1	2.138, 1.012, −1.031 (0.113) (0.064) (0.100)
γ_3 (value per click)	2, 1, −2	1.962, 0.965, −1.887 (0.107) (0.068) (0.105)
γ_4 (management cost)	2, 1	2.297, 0.961 (0.206) (0.106)
σ_q^2	0.5	0.498 (0.015)
σ_d^2	0.5	0.489 (0.014)
σ_{vpc}^2	0.5	0.504 (0.018)
β	0.5	0.451 (0.050)
Ω	0.5, 0.5, 0.5, 0.5 1, 1, 1 1.5, 1.5 2	0.592, 0.404, 0.384, 0.454 (0.073) (0.081) (0.097) (0.146) 1.079, 1.058, 1.245 (0.147) (0.159) (0.234) 1.547, 1.706 (0.198) (0.269) 2.397 (0.499)

true values lie in 95% confidence intervals of the posterior mean estimates. We also conducted additional simulations and found that the estimation results are not sensitive to the choice of starting values. Therefore, our simulation studies provide evidence of the validity of the proposed method.

References

Anselin L (1988) *Spatial Econometrics: Methods and Models*, Vol. 4 (Kluwer Academic Publishers, Dordrecht, The Netherlands).

Athey S, Nekipelov D (2012) A structural model of sponsored search advertising auctions. Working paper, University of California, Berkeley, Berkeley.

Bajari P, Benkard CL, Levin J (2007) Estimating dynamic models of imperfect competition. *Econometrica* 75(5):1331–1370.

Bajari P, Hong H, Krainer J, Nekipelov D (2010) Estimating static models of strategic interactions. *J. Bus. Econom. Statist.* 28(4):469–482.

Berry ST (1992) Estimation of a model of entry in the airline industry. *Econometrica* 60(4):889–917.

Bresnahan TF, Reiss PC (1990) Entry in monopoly markets. *Rev. Econom. Stud.* 57(4):531–553.

Bresnahan TF, Reiss PC (1991) Entry and competition in concentrated markets. *J. Political Econom.* 99(5):977–1009.

Chan TY, Park Y-H (2013) The value of consumer search activities for sponsored search advertisers. Johnson School Research Working Paper 45-09, Cornell University, Ithaca, NY.

Datta S, Sudhir K (2011) The agglomeration-differentiation tradeoff in spatial location choice. Working paper, Purdue University, West Lafayette, IN.

Datta S, Sudhir K (2013) Does reducing spatial differentiation increase product differentiation? Effects of zoning on retail entry and format variety. *Quant. Marketing Econom.* 11(1): 83–116.

Edelman B, Ostrovsky M, Schwarz M (2007) Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *Amer. Econom. Rev.* 97(1):242–259.

Feng J, Bhargava HK, Pennock DM (2007) Implementing sponsored search in Web search engines: Computational evaluation of alternative mechanisms. *INFORMS J. Comput.* 19(1):137–148.

Ghose A, Yang S (2009) An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Sci.* 55(10):1605–1622.

Goldfarb A, Tucker C (2010) Search engine advertising: Channel substitution when pricing ads to context. *Management Sci.* 57(3):458–470.

Jia P (2008) What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica* 76(6):1263–1316.

Katona Z, Sarvary M (2010) The race for sponsored links: Bidding patterns for search advertising. *Marketing Sci.* 29(2):199–215.

Mazzeo MJ (2002) Product choice and oligopoly market structure. *RAND J. Econom.* 33(2):221–242.

Rutz OJ, Bucklin RE (2011) From generic to branded. A model of spillover dynamics in paid search advertising. *J. Marketing Res.* 48(1):87–102.

Seim K (2006) An empirical model of firm entry with endogenous product-type choices. *RAND J. Econom.* 37(3):619–640.

Varian HR (2007) Position auctions. *Internat. J. Indust. Organ.* 25(6): 1163–1178.

Villas-Boas JM, Winer RS (1999) Endogeneity in brand choice models. *Management Sci.* 45(10):1324–1338.

Vitorino MA (2012) Empirical entry games with complementarities: An application to the shopping center industry. *J. Marketing Res.* 49(2):175–191.

Yang S, Ghose A (2010) Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Sci.* 29(4):602–623.

Yang S, Chen Y, Allenby GM (2003) Bayesian analysis of simultaneous demand and supply. *Quant. Marketing Econom.* 1(3): 251–275.

Yao S, Mela CF (2011) A dynamic model of sponsored search advertising. *Marketing Sci.* 30(3):447–468.

Zhu T, Singh V (2009) Spatial competition with endogenous location choices: An application to discount retailing. *Quant. Marketing Econom.* 7(1):1–35.