Marketing Science

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

# Aggregation Bias in Sponsored Search Data: The Curse and the Cure

Vibhanshu Abhishek, Kartik Hosanagar, Peter S. Fader

To cite this article:
Vibhanshu Abhishek, Kartik Hosanagar, Peter S. Fader (2015) Aggregation Bias in Sponsored Search Data: The Curse and the Cure. Marketing Science 34(1):59-77. http://dx.doi.org/10.1287/mksc.2014.0884

MARKETING SCIENCE
Vol. 34, No. 1, January–February 2015, pp. 59–77
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2014.0884


# Aggregation Bias in Sponsored Search Data: The Curse and the Cure

Vibhanshu Abhishek
Heinz College, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, vibs@cmu.edu

Kartik Hosanagar, Peter S. Fader
Wharton School of Business, University of Pennsylvania, Philadelphia, Pennsylvania 19104
{kartikh@wharton.upenn.edu, faderp@wharton.upenn.edu}

Recently there has been significant interest in studying consumer behavior in sponsored search advertising (SSA). Researchers have typically used daily data from search engines containing measures such as average bid, average ad position, total impressions, clicks, and cost for each keyword in the advertiser's campaign. A variety of random utility models have been estimated using such data and the results have helped researchers explore the factors that drive consumer click and conversion propensities. However, virtually every analysis of this kind has ignored the intraday variation in ad position. We show that estimating random utility models on aggregated (daily) data without accounting for this variation will lead to systematically biased estimates. Specifically, the impact of ad position on click-through rate (CTR) is attenuated and the predicted CTR is higher than the actual CTR. We analytically demonstrate the existence of the bias and show the effect of the bias on the equilibrium of the SSA auction. Using a large data set from a major search engine, we measure the magnitude of bias and quantify the losses suffered by the search engine and an advertiser using aggregate data. The search engine revenue loss can be as high as 11% due to aggregation bias. We also present a few data summarization techniques that can be used by search engines to reduce or eliminate the bias.

*Keywords*: sponsored search; generalized second-price auctions; consumer choice models; Hierarchical Bayesian estimation; latent instrumental variables; aggregation bias
*History*: Received: December 28, 2012; accepted: August 22, 2014; Preyas Desai served as the editor-in-chief and Gary Russell served as associate editor for this article.



## 1. Introduction

Sponsored search advertising (SSA) has not only transformed the way companies conduct their marketing activities, but it has also been a tremendous resource to academic researchers who seek to better understand how consumers respond to such ads. A myriad of researchers have turned to SSA data to uncover new insights about consumer search (Ghose and Yang 2009, Rutz and Bucklin 2011), choice, related purchasing behaviors (Jeziorski and Segal 2009, Yang and Ghose 2010, Agarwal et al. 2011), and advertiser/search engine strategies (Animesh et al. 2010, Yao and Mela 2011, Rutz et al. 2012). Many of these papers have used random utility models to study the effect of ad position, keyword length, presence or absence of brand name, etc., on the click-through and conversion rates of the ads.

Sponsored search refers to ads that are displayed alongside organic search results when a user issues a query at a search engine. The advertisers submit bids for keywords that are relevant to them, along with these ads.[1] When a user enters a query, the search engine identifies the advertisers bidding on keywords closely related to the query and uses data on bids and ad quality/performance to rank order the ads in a list of sponsored results. The most widely used pricing model is the *pay-per-click* model in which the advertiser pays only when a user clicks on his ad. The advertiser's cost per click (CPC) is determined using a generalized second price (GSP) auction, i.e., whenever a user clicks on an ad at a particular position, the advertiser pays an amount equal to the minimum bid needed to secure that position.

Although SSA is a relatively new practice, it already has fairly well established data standards associated with it. Most researchers who have modeled SSA-related issues have worked with a data structure such as the one illustrated in Table 1. Advertisers also

[1] The term *keyword* refers to the term or phrase on which an advertiser bids. *Query* (or *query term*) is the search phrase entered by the consumer when conducting the search.

**Table 1** **Sample Data Set for a Particular Keyword**

| Date | Impressions | Clicks | Avg. pos | Avg. bid | Avg. CPC |
|---|---|---|---|---|---|
| 01/11/09 | 180 | 1 | 19.33 | 1.00 | 0.30 |
| 01/12/09 | 202 | 0 | 18.42 | 1.00 | 0.00 |
| 01/13/09 | 223 | 5 | 8.19 | 2.00 | 1.24 |
| 01/14/09 | 198 | 3 | 7.94 | 2.00 | 0.89 |
| 01/15/09 | 197 | 5 | 8.08 | 2.00 | 1.21 |
| 01/16/09 | 321 | 21 | 2.00 | 3.00 | 2.12 |

obtain similar data sets from search engines and analyze them to design their bidding policies. In almost all cases, these data are aggregated to the daily level and contain summary statistics for the day, such as the number of ad impressions, average position of the ad, number of clicks received, and the average CPC. It should be clear how this kind of data set lends itself to the types of models mentioned above, as well as analysis of a variety of other customer behaviors (and related firm actions). Yet despite the creativity and methodological prowess that has been demonstrated in this growing body of literature, we believe that these modeling efforts are plagued by a major problem, i.e., an aggregation bias due to the way that the raw (search-by-search) data are "rolled" up into Table 1.

In practice, the position of an ad can vary substantially within a day. Aggregated data fail to capture this variation. It is also widely known that the impact of position on CTR is nonlinear. For example, an ad at the topmost position tends to receive a disproportionately large number of clicks compared to ads in other positions. The convexity in the CTR, coupled with the intraday variation in position suggests that the daily aggregation might lead to estimation bias. The goal of this paper is to provide a thorough evaluation of the nature of this bias, to demonstrate its effects, and provide recommendations to eliminate the bias.

The paper makes the following contributions. First, we show that applying a logistic model to aggregated SSA data can lead to biased estimation of the parameters of a random utility model. Because of the bias, the effect of position on CTR is attenuated and the predicted CTR is higher than the actual CTR. Surprisingly, we find that if all of the advertisers use aggregate data, the consequence of the bias is borne entirely by the search engine. Second, we quantify the magnitude of the bias and measure its economic impact. While some researchers studying the use of aggregated consumer data in other marketing settings suggest that aggregation bias is not important (Gupta et al. 1996, Russell and Kamakura 1994), others emphasize the need to explicitly address aggregation bias (Narayanan and Nair 2013, Neslin and Shoemaker 1989). Using a unique and large disaggregate data set from a major search engine, we show that aggregation bias is very significant in this context. We find that the search engine losses can be as high as 11% on average due to aggregation bias. Our findings raise serious concerns for SSA researchers and practitioners and also question the adequacy of the data standards that have become common in SSA. Finally, we present alternative data summarization techniques and modeling approaches that can reduce or eliminate the bias. We find that sharing harmonic means instead of arithmetic means or complementing arithmetic means by other summary information such as the variance can help significantly reduce the bias. These results are robust in our analysis of a large real-world data set as well as simulations that attempt to vary several aspects of the sponsored search market.

The rest of the paper is organized as follows. Section 2 discusses related work and positions our work in the literature. In §3, we analytically prove the existence of the bias and build a game-theoretic model to study the economic impact of the bias. In §4, we analyze a large disaggregate data set from a search engine using the Hierarchical Bayesian (HB) model with latent instrumental variables (LIV). We present the managerial implications of the bias in §5. In §6, we present some cures for the problem of aggregation, including data summarization techniques and modeling approaches that can reduce the bias. Finally, we discuss the implications of the bias on research and practice. Section 7 summarizes this paper and discusses possible directions for future research.

## 2. Related Work

There has been a considerable amount of work on auction design and consumer choice models in SSA (Weber and Zheng 2007, Liu et al. 2009, Goldfarb and Tucker 2010). More specifically, there are two streams of work that are closely related to our study, i.e., empirical research on consumer click and conversion behavior in SSA and work related to aggregation biases in choice models.

### Empirical Research in Sponsored Search

Recently there has been a lot of interest in trying to understand the factors driving keyword performance in SSA. Craswell et al. (2008) and Ali and Scarr (2007) propose individual keyword-level models to study how consumers navigate sponsored links. Other researchers have used logit models to measure the influence of factors such as ad position and keyword characteristics on consumer behavior in SSA (Rutz et al. 2012, Rutz and Bucklin 2011, Ghose and Yang 2009, Agarwal et al. 2011). Rutz et al. (2012) compare the performance of several logit models in predicting the conversions for various keywords. Their results show that keywords are heterogeneous in their conversion rate and that a significant portion of this variation can be explained by the presence of brand or geographical information in the keyword. In another paper, Rutz and Bucklin (2011) measure the spillover effect of generic keywords on branded keywords. Ghose and Yang (2009) use a random effect logit model to understand the relationship between different metrics such as CTR, conversion rates, bid prices, and ad position using the advertiser's aggregate data. They show that keywords containing retailer information have a higher CTR whereas keywords that are more specific or that contain brand information have a lower CTR. Recent work by Agarwal et al. (2011) uses a logit model to show that although the CTR decreases with position, the conversion rate is nonmonotonic in position. They point out that the topmost position is not necessarily the revenue maximizing position.

Most of this stream of research uses aggregate data to estimate the parameters of the model. The aggregate data obfuscate the variations in ad position. To our knowledge, research in this area has overlooked this fact. Ignoring this variation can lead to potential biases in the estimation of parameters and ultimately affect the conclusions from these studies.

**Aggregation Bias in Choice Models**

Though researchers have grappled with the issue of data aggregation for many years, there is still no clear consensus. The problems associated with aggregation have been commonly encountered in spatial and demographic studies and are referred to as the Yule-Simpson effect (Good and Mittal 1987). The drawbacks of aggregation have also been pointed out in various studies in the economics and marketing literature. Neslin and Shoemaker (1989) point out the limitations of aggregate data by refuting the claim that sales promotions undermine the consumer's repeat-purchase propensity. They show that even if the individual purchase propensities do not change before and after promotions, statistical aggregation would lead to lower average repeat probabilities for post promotional purchases. Yatchew and Griliches (1985) discuss the implications of aggregation in the context of probit models. Issues related to data aggregation in the case of logit models have been presented by Kelejian (1995). He discusses why aggregation bias might occur when logit models are estimated on aggregate data and proposes a test for the existence of this bias.

On the other hand, several researchers believe that the effect of aggregation is negligible or absent when the disaggregate model can be approximated by the aggregate model (Gupta et al. 1996, Russell and Kamakura 1994). Using household-level panel data and store-level purchase data, Gupta et al. (1996) show that the price elasticity estimated from the two models differ by a very small amount (4.7%). Allenby and Rossi (1991) present an analytical proof for the nonexistence of aggregation bias in nested logit models of consumer choice when the products are close substitutes of each other, though they assume that the micro-level consumer behavior is approximately linear in the product attributes. More specifically, in the context of sponsored search, Rutz and Trusov (2011) use a latent instrumental variable approach that addresses endogeneity in search auctions. Their model has an added benefit in that it addresses some aspects of aggregation bias, such as aggregation of data from heterogeneous consumers. However, as we show later in §5, the approach does not fully address the aggregation bias resulting from aggregated ad performance data when ads are shown in multiple positions.

The discussion reveals two themes. First, a number of recent studies have applied the logit model on aggregated SSA data to study consumer choice behavior. Second, although aggregation bias has been shown to exist in a number of environments, its nonexistence has also been demonstrated in several other environments. It is unclear which of these arguments is most applicable in the SSA context. Thus, it is unclear whether and to what extent aggregation bias affects SSA research. This paper uses a theoretical model to show why data aggregation might lead to biased estimates in the SSA context and how this bias affects the outcome in a search engine auction. For the first time (to our knowledge), an extensive disaggregate search engine data set is used to empirically measure the extent of aggregation bias in SSA research. Finally, we suggest ways in which the bias can be reduced or eliminated.

## 3. Aggregation Bias in Sponsored Search

In this section, we explore the estimation bias due to the aggregation of SSA data. There is a distinction between the complete (disaggregate) and the summary (aggregate) data that have been referred to in the paper. Table 2 is a stylized example of impression level data for an ad, reflecting every search query for a particular term. Each observation contains the date on which the impression occurred, position of the ad, bid placed by the advertiser, whether the consumer clicked on the ad, and finally the CPC. Search engines usually do not provide such granular data to advertisers or researchers. They provide aggregated data at the daily level as shown earlier in Table 1 that mask the intraday variation in position. Discussions with several search engines reveal that they do not provide impression level data because it is expensive for advertisers to store, manage, and analyze such huge amounts of data. In addition, concerns about user privacy and click-fraud further reduce the incentive to provide disaggregate data.

**Table 2 Complete Data Set for a Particular Keyword**

| Impression | Date | Click | Pos | Bid | CPC |
|---|---|---|---|---|---|
| 1 | 01/11/09 | 0 | 16 | 1 | 0.00 |
| 2 | 01/11/09 | 0 | 20 | 1 | 0.00 |
| ... | ... | ... | ... | ... | |
| 8,190 | 02/15/09 | 0 | 6 | 2 | 0.00 |
| ... | ... | ... | ... | ... | |
| 9,145 | 02/23/09 | 1 | 1 | 3 | 2.31 |

The intraday variation arises due to two major factors. First, SSA auctions are extremely dynamic with advertisers entering and exiting the auction or changing their bids continuously. Changes in competitors' behavior lead to changes in the ad position. Second, most of the ads, specifically *broad match* and *phrase match* ads are shown for a number of different queries.[2] As the set of competitors can be different for different queries, the position of the ad also varies across queries.

### 3.1. Analytical Proof of Aggregation Bias

**Logit Utility Model.** The logit utility model has been extensively used in economics and marketing to explain consumer choice behavior. Researchers have primarily focused on keyword-level models to analyze the effect that factors like ad position, specificity of the keyword, and presence of brand name have on the consumer's propensity to click on the ad. The consumer's utility has been modeled as

$$U = \mathbf{X}'\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\mathbf{X}$ is a vector of covariates, $\boldsymbol{\beta}$ is the consumer's sensitivity to these attributes, and $\epsilon \sim \text{Logistic}(0,1)$. In binary choice models, this utility is not observed but constitutes a latent variable. The consumer clicks on an ad when $U > 0$. Variable $Y \in \{0,1\}$ denotes whether a click was made. In conjugation with prior research, we build a keyword-level model that ignores other ad characteristics and focuses on the impact of ad position on CTR, i.e., CTR $= f(\text{position})$. The simple model allows us to clearly identify the existence and direction of the bias. However, this assumption does not impose any restrictions on the model as all keyword characteristics (which typically do not change during the day) are subsumed in the intercept term and then we focus on position, which varies intraday. Although we focus on a keyword-level model in this paper, our findings are applicable for different levels of analysis.

[2] An *exact match* occurs when the user's query term exactly matches the advertiser's keyword. A *phrase match* occurs when the advertiser's keyword appears anywhere within the user's query. Finally, a *broad match* occurs when the user query is determined to be broadly similar to an advertiser's keyword. Broad match is commonly used by advertisers as it maximizes the number of ad impressions.

**Estimation Using Complete Data Set.** We now discuss the estimation of $\boldsymbol{\beta}$ when the model is estimated using the complete data set. Let $V_i$ be a random variable denoting the ad position on the $i$th impression. We assume that $V_i$ is independent and identically distributed and has a cumulative distribution function (c.d.f.) given by $F_V(\cdot)$, which is assumed to be constant during the period of observation.[3] The consumer's utility is given by the following expression:

$$U_i = \beta_0 + \beta_1 V_i + \epsilon_i. \tag{2}$$

As $\epsilon_i$ is extreme value distributed, the probability of clicking on an ad (CTR) is $p_i = 1/(1+\exp-\{\beta_0+\beta_1 v_i\})$. Note that $p_i$ might vary across impressions due to variation in the ad position. Let $\hat{\boldsymbol{\beta}}_c$ denote the maximum likelihood estimate from the complete data set. It can be easily shown that $\hat{\boldsymbol{\beta}}_c$ is a consistent and unbiased estimator of $\boldsymbol{\beta}$ (Hayashi 2000, Proposition 7.6).

**Estimation Using Aggregate Data Set.** Researchers do not observe $V_i$ when aggregate data are used. They only observe the mean daily position $W$ which is given by

$$W = \frac{V_1 + V_2 + \cdots + V_N}{N}, \tag{3}$$

where $N$ is the random number of ad impressions on a particular day and $V_1, \ldots, V_N$ is the ad position during each of those impressions. The distribution of $W$ is $F_W(\cdot)$ and it depends on the distribution of $V$ and $N$. If the effect of position is estimated from aggregate data, the consumer utility from clicking the ad is effectively modeled as

$$U_d = \beta_0 + \beta_1 W_d + \epsilon_d, \tag{4}$$

where $W_d$ is the average position for the day and $\epsilon_d$ is the logistically distributed error term. This formulation causes a misspecification as the consumers do not observe the ad at a position $W_d$ but at position $V$. As the variable $Z = V - W_d$ ($\mathbb{E}[Z \mid W] = 0$), which affects the consumer's click behavior, is not accounted for in the regression, the misspecification is similar to omitted variables bias pointed out by Yatchew and Griliches (1985) and Wooldridge (2001). However, this issue arises primarily due to data aggregation. Our approach is closely related to prior work in marketing by Christen et al. (1997), Steenkamp et al. (2005) and Gupta et al. (1996). Furthermore, as $\text{Var}(Z)$ is not constant (it depends on the number of impressions on a particular day), the findings of Yatchew and Griliches (1985) and Wooldridge (2001) are not directly applicable in this context. Hence, we derive an important relationship between $W$ and $V$ to prove the aggregation bias.

[3] Let $\mu_V = \mathbb{E}[V]$ and $\sigma_V^2 = \text{Var}(V)$.

Lemma 1. *$W$ is less than $V$ in convex order, i.e., $W \leq_{\text{cx}} V$.*[4]

This relationship between $W$ and $V$ is very general and holds for any distribution, $F_V(\cdot)$. Using Lemma 1, we prove an important result of this paper.

Proposition 1. *The estimate for the aggregate data, $\hat{\boldsymbol{\beta}}_s$, is a biased estimator of $\boldsymbol{\beta}$.*

If $\hat{\boldsymbol{\beta}}_c$ is equal to $\hat{\boldsymbol{\beta}}_s$ then the convex order between $W$ and $V$ implies that

$$\mathbb{E}\left[\frac{\exp\{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V\}}{1+\exp\{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V\}}\right] > \mathbb{E}\left[\frac{\exp\{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W\}}{1+\exp\{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W\}}\right].$$

Because both the left-hand side and the right-hand side equal the overall CTR during the observation period (as shown in the appendix), this inequality is incorrect and hence $\hat{\boldsymbol{\beta}}_c$ cannot equal $\hat{\boldsymbol{\beta}}_s$. Since $\hat{\boldsymbol{\beta}}_c$ is a consistent and unbiased estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_s$ is biased. This finding is contrary to earlier work by Allenby and Rossi (1991), Gupta et al. (1996), and Russell and Kamakura (1994), which prove that aggregation bias in market or store-level scanner data is negligible. Aggregation bias is significantly reduced in their context as products are very close substitutes for each other and the consumers (or households) are exposed to very similar marketing activities. However, position has a very strong effect on sponsored search (Craswell et al. 2008) and ads in different positions may be perceived very differently by consumers. Coupled with variation in ad position, aggregation bias can be quite substantial, which is formalized in the following proposition.

Proposition 2. *The direction of aggregation bias is such that* (i) *the CTR estimated from the summary data is greater than or equal to the actual CTR at any position,* (ii) *$\hat{\beta}_{1,s} > \hat{\beta}_{1,c}$, and under certain conditions,* (iii) *$\hat{\beta}_{1,s} \xrightarrow{p} \beta_1/\sqrt{\beta_1^2\{\sigma_V^2+(3\sigma_V^2+\mu_V^2)\phi\}+1}$, where $\phi = \mathbb{E}[1/N]$.*[5]

The estimate for the aggregate data, $\hat{\boldsymbol{\beta}}_s$ is biased and predicts a CTR that is higher than the actual CTR at any position. Incorrect estimation of CTR might lead advertisers to make suboptimal choices in sponsored search auctions. The second part of Proposition 2 is consistent with Wooldridge (2001) which shows that the estimates are scaled towards zero. Furthermore, the omitted variable $Z$ increases the disturbance term in the regression. The estimated error can be computed as a convolution of $\epsilon$ and $Z$, but this is analytically intractable as $\epsilon$ is logistically distributed. However, Proposition 2(iii) holds when $\beta_1 Z + \epsilon$ can be approximated closely by a logistic distribution. Proposition 2 shows that, if the variation in the intraday position is known, the actual $\beta_1$ can be approximated by multiplying $\hat{\beta}_{1,s}$ by the scaling factor. However, this approach suffers from several simplifying assumptions that are required for analytical tractability. We provide more general and robust empirical methods to remove the bias in §5. Although we prove that the bias exists for a logistic model, it is easy to show that these results would be applicable in any binary choice model where the choice behavior is convex in position.

### 3.2. Effect on Equilibrium Behavior

As advertisers use estimates from historical data to bid in SSA auctions, incorrect estimation of the CTR might have a negative impact on their revenues. In this section, we build a game-theoretic model to analyze the impact of aggregation on the advertisers' and search engine's revenues.

For our analysis, we make the following assumptions. (i) There are $K$ advertising slots and $K+1$ advertisers; (ii) The advertisers' valuation for a click, $s_k$ are drawn from a continuous distribution with support on $[0,\infty)$; (iii) Advertisers know their own valuation and the distribution of competing bids and finally; (iv) The CTR, $\alpha_i$, decreases with the position $i$. In addition, we assume that advertisers estimate $\alpha_i$ from historical data and are unaware of the aggregation bias. The advertisers are indexed in decreasing order of their valuations, i.e., $s_1 > s_2 > \cdots s_{K+1}$ and their bids are $b_1,\ldots,b_{K+1}$, respectively.[6] In addition, let $h=(b_i,\ldots,b_{K+1})$ refer to the history of bids prior to assignment of position $i$.[7] The case wherein complete data are used is analyzed first. Here the advertisers correctly estimate $\alpha_i$. Edelman et al. (2007) show that under the above assumptions, there exists a unique *envy-free* perfect Bayesian equilibrium and the optimal strategy for advertiser $k$ under this equilibrium is to bid as follows:

$$b_k(s_k,i,h) = s_k - \frac{\alpha_i}{\alpha_{i-1}}(s_k - b_{i+1}). \tag{5}$$

This is the maximum CPC that the advertiser is willing to pay to move to position $i-1$ and receive more clicks. At this point, advertiser $k$ is indifferent between getting position $i-1$ at a CPC of $b_k(s_k,i,h)$ and position $i$ at $b_{i+1}$. This is an ex-post equilibrium, i.e., it is optimal for advertiser $k$ to follow the equilibrium strategy for any realization of other

[4] A random variable $V$ is less than $Y$ in *convex order* if $E[f(X)] \leq E[f(Y)]$ for all real convex functions $f$ such that the expectation exists. All proofs appear in the appendix.

[5] The expectation of the inverse of the number of daily impressions, $\mathbb{E}[1/N]$, is bounded by $(1-(1-e^{-\lambda}))/\lambda \leq \phi \leq (1-[1-F_\lambda(1)+3(1-F_\lambda(2))/m])$ where $\lambda = \mathbb{E}[N]$ and $F_\lambda(\cdot)$ is the CDF of $N$. When there are a large number of daily impressions, i.e., $\lambda \to \infty$, $\phi \to 1$.

[6] The bid for the last advertiser is normalized to zero, $b_{K+1} = 0$.

[7] Positions $K$ through $i+1$ are assigned before bidding for position $i$ starts.

advertisers' valuations. The search-engine revenue is $\Pi_S^C = \sum_{i=1}^{K} \alpha_i b_{i+1}$ and the payoffs for advertiser $i$ is $\Pi_i^C = \alpha_i(s_i - b_{i+1})$. Note that this equilibrium ensures an assortive match, i.e., if $s_i > s_j$ then advertiser $i$ bids higher than advertiser $j$ and occupies a slot above advertiser $j$ in equilibrium. This case serves as a reference for the ensuing discussion.

Next, we consider the case in which aggregate data are used. Let the CTR estimated from aggregate data be denoted as $\alpha'_i$.

Remark 1. When aggregate data are used to estimate CTR, the ratio $\alpha_i/\alpha_{i-1}$ is overestimated due to the presence of aggregation bias.[8]

Because of this overestimation, advertisers might bid incorrectly. As the equilibrium considered here is ex-post, the advertisers' bidding strategies depend neither on their beliefs about each others' valuations nor on the fact that some advertisers might be using aggregate data. The bidding strategies continue to be similar to the one outlined in Equation (5), but the bids in this case, $b'_1, \ldots, b'_K$, might be different.[9] We consider two extreme cases to study the impact of aggregation bias. (i) All advertisers except one use complete data and; (ii) All advertisers use aggregate data. Let the search-engine revenue in Case I(II) be denoted by $\Pi_S^{\mathrm{AI(II)}}$ and advertisers' profit by $\Pi_i^{\mathrm{AI(II)}}$.

*Case* I. Suppose advertisers other than advertiser $j$ have access to complete data and can compute $\alpha_i$ correctly. Only advertiser $j$ uses aggregate data and overestimates $\beta_1$. This leads him to overestimate the ratio $\alpha_i/\alpha_{i-1}$ and he bids in the following manner:

$$b'_j(s_j, i, h) = s_j - \frac{\alpha'_i}{\alpha'_{i-1}}(s_j - b'_{i+1}). \qquad (6)$$

As advertiser $j$ bids lower in equilibrium, he occupies a position $j' \geq j$. The following proposition characterizes the equilibrium in this case (detailed analysis and proofs are provided in the appendix).

Proposition 3. (i) *If only advertiser $j$ uses aggregate data, the top advertisers ($i \leq j$) bid lower, advertisers in between ($j < i \leq j'$) bid higher, and the remaining advertisers ($i > j'$) bid the same as they would have when everyone had complete data*. (ii) *The payoffs of the search engine and advertiser $j$ decrease ($\Pi_S^{\mathrm{AI}} < \Pi_S^C$, $\Pi_j^{\mathrm{AI}} < \Pi_j^C$) while all other advertisers receive payoffs that are either the same or higher than payoffs they would have received if all advertisers were using complete data ($\Pi_i^{\mathrm{AI}} \geq \Pi_i^C$, $i \neq j$).*

Advertiser $j$ underestimates the impact of position and incorrectly bids less, which might move him to a lower position. In turn, some advertisers who were below him move up one position. The ordering of these advertisers does not change, which is a consequence of the bidding policy (Edelman et al. 2007). As $b'_j < b_{j'}$, bids required to acquire all positions above $j'$ decrease. As bids for all positions are (weakly) lower, the search engine loses revenue. Clearly, advertiser $j$'s payoff is lower because he deviates from the optimal policy. However, the loss in revenue for the search engine is substantially higher than the loss in revenue for the advertiser using aggregate data. Interestingly, all of these losses are transferred to the other advertisers ($\neq j$) as excess surplus since GSP is a zero-sum game. Hence, the search engine suffers the most due to aggregation bias and all advertisers, apart from $j$, are better off due to aggregation. In the subsequent case, we observe that the search engine internalizes all of the negative impact of aggregation.

*Case* II. When all advertisers use aggregate data, their estimates of the CTR, $\alpha'_i$ are greater than the actual CTR as shown in Proposition 3. For simplicity, we assume that all advertisers arrive at the same estimates for $\alpha'_i$.[10] As a result, advertiser $k$ adopts the following bidding strategy:

$$b'_k(s_j, i, h) = s_k - \frac{\alpha'_i}{\alpha'_{i-1}}(s_k - b'_{i+1}).$$

It is easy to see that the bid placed by advertiser $K$ is less than the bid he would have placed had he estimated CTR from complete data. Proceeding in an iterative fashion we show that all advertisers place a lower bid. The equilibrium in this case is specified in the following proposition:

Proposition 4. (i) *When all advertisers use aggregate data, the advertisers are arranged in assortive order. The resulting bids are lower than the bids when complete data are used* ($b'_i < b_i$, $i = 1, \ldots, K$). (ii) *Search-engine revenue is lower* ($\Pi_S^{A2} < \Pi_S^C$) *and advertisers' payoffs are higher* ($\Pi_i^{A2} > \Pi_i^C$) *as compared to the complete case.*

In the appendix, we show that the advertisers bid less than what they would have had they known the actual CTR. As all of the advertisers use the same incorrect CTR estimate, the eventual ranking remains the same as in the complete case. They receive the same number of clicks but at a lower CPC; hence, their payoffs are higher. Surprisingly, the search-engine revenue suffers the most when all advertisers use aggregate data even though the advertisers make the wrong decisions. These results question the

[8] Writing the CTR in terms of the logit model we get, $\alpha_i/\alpha_{i-1} = \exp\{\beta_0 + \beta_1 i\}/(1 + \exp\{\beta_0 + \beta_1 i\}) \times (1 + \exp\{\beta_0 + \beta_1(i-1)\})/\exp\{\beta_0 + \beta_1(i-1)\} \approx \exp\{\beta_1\}$ when CTRs are small. Since $\hat{\beta}_{1,s} > \hat{\beta}_{1,c} \Rightarrow \alpha_i/\alpha_{i-1} < \alpha'_i/\alpha'_{i-1}$.

[9] We continue to assume that advertiser $K+1$ still bids 0.

[10] This result continues to hold even if the advertisers arrive at different estimates of $\alpha'_i$ as long as $\alpha_i/\alpha_{i-1} < \alpha'_i/\alpha'_{i-1}$, which always hold true due to aggregation bias as shown earlier.

data standards that have become common in SSA and underscore the need to provide better data to advertisers. We also show that it is incentive compatible for the search engine to provide richer/better data to advertisers.

Note that an advertiser always receives a higher payoff when he uses complete data as compared to aggregate data, irrespective of the fraction of advertisers using aggregate data. This intuition is formalized in the following proposition.

Proposition 5. *An advertiser can always increase his payoff from SSA by unilaterally using complete data instead of aggregate data.*

The difference in payoff between the two cases (complete versus aggregate) can be considered as the value of complete data or alternately the disutility from aggregate data. Although advertisers cannot get impression level data, they can periodically crawl the search engine and estimate the empirical distribution of the ad position. Moreover, with recent improvements in ad tagging and user tracking techniques, advertisers might collect impression level data for a few customers.

## 4. Empirical Analysis

In the previous section, we analytically show that aggregation at a daily level leads to a bias. However, as stated earlier, several papers show that aggregation bias is negligible in various marketing data (Allenby and Rossi 1991, Gupta et al. 1996, Russell and Kamakura 1994). They argue that aggregation bias is significantly reduced because the products considered in their analysis are very close substitutes. These findings might not hold in SSA as Craswell et al. (2008), Ghose and Yang (2009), and Abhishek and Hosanagar (2013) show that ad position has a very strong effect on SSA. We perform the following empirical analysis on large representative search engine data to examine and conclusively prove that ad position has a strong influence on consumer click behavior, which leads to significant aggregation bias. Furthermore, we measure the extent of the bias and observe its economical significance.

### 4.1. Data Description

We analyze a large disaggregate data set from a major search engine, which is extremely representative of consumer behavior in SSA. The data set contains around 8 million unique impressions chosen randomly from all user queries between August 10, 2007 and September 25, 2007. These are very unique data as search engines rarely provide impression level data to advertisers or researchers. For every impression, the data set contains the user query, ads shown on the page, and number of ads on the preceding pages. Each ad is identified by a unique ad identifier, though the data set does not contain any ad-specific information. The data set also contains information about clicks during this period of observation. The summary statistics for this data is presented in Table 3. We construct an ad-level data set that contains information about the keyword the advertiser was bidding on and all of the impressions of the ad associated with the keyword, which is similar to the one presented in Table 2.

**Table 3** Summary Statistics

| | |
|---|---|
| Total impressions | 8,142,210 |
| Unique queries | 24,235 |
| Unique ads | 229,960 |
| Ads with more than one impression | 184,481 |
| Mean impressions for every ad | 64.4 |
| Median impressions for every ad | 7.0 |

There is evidence of substantial variation in position. Reporting average position alone results in the loss of information on actual position as shown in Figure 1. We next investigate the impact of data aggregation.

The ad level data described earlier are summarized at a daily level to create aggregate data. The data thus generated are similar to the campaign summaries that search engines make available to the advertisers (as presented in Table 1).

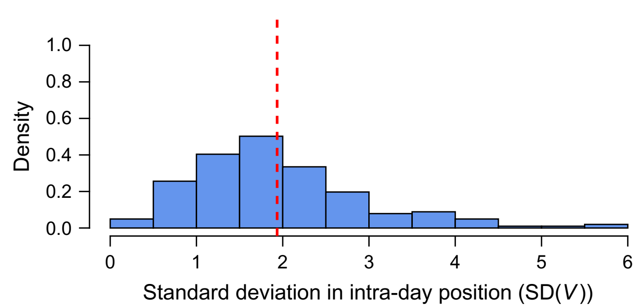
### 4.2. Hierarchical Bayesian Model

We estimate a random-effect logit model using Hierarchical Bayesian (HB) techniques that are commonly used in SSA. As our data do not contain any ad-specific attributes, the only covariate included in our models is position. The effect of ad characteristics is captured in the ad-specific intercept term. We demonstrate the aggregation bias for the HB model.

We extend the binary choice logit model proposed earlier in §3 to account for multiple keywords. Under this specification, the consumer's utility from clicking on ad $k$ during impression $i$ is given by

$$U_{ik} = \beta_{0k} + \beta_{1k} V_{ik} + \epsilon_{ik}, \tag{7}$$

**Figure 1** (Color online) Mean Intra-Day Variation in Position



where $\epsilon_{ik}$ is the idiosyncratic, logistically distributed error term. The keyword specific parameters $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k})'$ are assumed to be random and heterogeneous across ads. They are drawn from a multivariate normal distribution in the following manner:

$$\boldsymbol{\beta}_k \sim \mathrm{N}_2(\boldsymbol{\mu}_\beta, V_\beta) \quad \text{where } V_\beta = \begin{pmatrix} \sigma^2_{\beta_0} & \sigma^2_{\beta_0\beta_1} \\ \sigma^2_{\beta_0\beta_1} & \sigma^2_{\beta_1} \end{pmatrix}.$$

Similar models have been extensively used in prior SSA research (Ghose and Yang 2009, Yang and Ghose 2010). Note that although this random coefficient model captures heterogeneity across ads, it still fails to account for the intraday variation in $V_{ik}$ if the model is estimated on aggregate data. As a result, we expect the aggregation bias to extend to the random-coefficient model as well. To test this hypothesis, we take a random sample of ads from our data and apply the model in Equation (7) to both the disaggregate and aggregate data sets and compare the estimates.

The log-likelihood function for the complete data is as follows:

$$\mathrm{LL}(\boldsymbol{\beta} \mid \text{complete data}) \propto \sum_{k=1}^{K}\sum_{i=1}^{I_K} Y_{ik}\log p_{ik} + (1-Y_{ik})\log(1-p_{ik}), \tag{8}$$

where $Y_{ik}$ is the indicator variable that denotes whether the $i$th impression of keyword $k$ received a click and $p_{ik}$, the click-through probability is given by

$$p_{ik} = \frac{\exp\{\beta_{0k} + \beta_{1k} v_{ik}\}}{1+\exp\{\beta_{0k}+\beta_{1k} v_{ik}\}}. \tag{9}$$

The log-likelihood function for the aggregate data is as follows:

$$\mathrm{LL}(\boldsymbol{\beta} \mid \text{aggregate data}) \propto \sum_{k=1}^{K}\sum_{d=1}^{D} c_{dk}\log p_{dk} + (n_{dk}-c_{dk})\log(1-p_{dk}), \tag{10}$$

where $n_{dk}$ and $c_{dk}$ denote the number of impressions and clicks on day $d$, respectively, and $p_{dk}$, the click-through probability is given by

$$p_{dk} = \frac{\exp\{\beta_{0k} + \beta_{1k} w_{dk}\}}{1+\exp\{\beta_{0k}+\beta_{1k} w_{dk}\}}. \tag{11}$$

As the data on clicks are often sparse for most keywords in sponsored search, the SSA literature primarily uses HB models. We use a similar approach and assume that the mean and variance-covariance matrix for $\boldsymbol{\beta}_k$ have the following priors

$$\boldsymbol{\mu}_\beta \sim \mathrm{N}_2(\boldsymbol{\mu}, \Sigma), \tag{12}$$

$$V_\beta^{-1} \sim \mathrm{Wishart}(\nu, \Delta). \tag{13}$$

The parameters $\mu$, $\Sigma$, $\nu$, $\Delta$, $\boldsymbol{\mu}_\beta$, and $V_\beta^{-1}$ are estimated separately from the complete and aggregate data sets using a Markov Chain Monte Carlo (MCMC) approach. Before discussing the details of the MCMC estimation procedure, we discuss some identification issues associated with the model presented here.

### 4.3. Identification

The ad position in the previous exposition has been assumed to be exogenous. However, the position is decided by the bids placed by the advertiser. In addition, we know that past performance affects the quality score of the ad, which in turn affects the position. The auction process and historical performance jointly determine the position, which is one of the most important strategic variables that advertisers focus on in SSA. This indicates that the position is endogenous (i.e., $\mathbb{E}[\mathrm{pos}_t \epsilon_t] \neq 0$) and that the endogeneity should be explicitly incorporated in the HB model presented earlier.[11]

Endogeneity has been a major concern in the SSA literature. Researchers have proposed several techniques to address this issue. Ghose and Yang (2009) and Yang and Ghose (2010) use a simultaneous equation model to address this problem. Their simultaneous model forms a triangular system of equations that can be identified without any further identification constraints. Agarwal et al. (2011) use a series of random bids to address the endogenous nature of position. In their specification, position is completely determined by the random bids and quality score, which are exogenous. Recent econometric advances have led to the development of the latent instrument variable (LIV) framework (Ebbes et al. 2005), which has been used by Rutz and Trusov (2011) and Rutz et al. (2012) to account for position endogeneity. The LIV framework uses a likelihood based approach, which can be easily integrated with the HB model proposed earlier, and can be estimated using the MCMC estimator.

In an LIV formulation, the endogenous covariate is decomposed into a stochastic term that is uncorrelated with the error and another one that is possibly correlated with the error, i.e., $X = \theta + \eta$ where $\theta$ is the uncorrelated part of $X$ such that $\mathbb{E}[\theta\epsilon] = 0$ and $\mathbb{E}[\eta\epsilon] = \sigma_{\eta\epsilon}$. Because $\theta$ varies in the data set, it is possible to identify the correlation between $\eta$ and $\epsilon$, denoted by $\sigma_{\eta\epsilon}$. For the sake of simplicity, we modify the model presented in Equation (7) such that, $\epsilon_{ik} = \zeta_{ik} + \vartheta_{ik}$, where $\zeta_{ik}$ is correlated with ad position and $\vartheta_{ik}$ is orthogonal to position and logistically distributed. Clearly, $\mathbb{E}[\eta\zeta] = \mathbb{E}[\eta\epsilon] = \sigma_{\eta\epsilon}$, which we denote by $\sigma_{\eta\zeta}$ for the sake of exposition. The LIV approach is extended to binary choice models by

[11] We thank the anonymous reviewer for this suggestion.

**Table 4** **Estimates of Parameters**

| | HB | | HB-LIV | |
|---|---|---|---|---|
| Parameters | Complete | Aggregate | Complete | Aggregate |
| $\boldsymbol{\mu}_\beta$: | | | | |
| $\mu_{0\beta}$ | −1.495 (0.000) | −1.459 (0.001) | −1.672 (0.301) | −1.612 (0.189) |
| $\mu_{1\beta}$ | −0.793 (0.085) | −0.727 (0.098) | −0.642 (0.120) | −0.558 (0.074) |
| $V_\beta$: | | | | |
| $\sigma_{\beta_0}$ | 0.654 (0.128) | 0.753 (0.170) | 0.678 (0.107) | 0.689 (0.192) |
| $\sigma_{\beta_1}$ | 0.153 (0.037) | 0.155 (0.038) | 0.147 (0.033) | 0.162 (0.042) |
| $\sigma_{\beta_1\beta_2}$ | 0.025 (0.067) | −0.085 (0.076) | 0.019 (0.052) | −0.090 (0.076) |
| $V_\xi$: | | | | |
| $\sigma_\zeta$ | 0.263 (0.067) | 0.389 (0.082) | 0.142 (0.047) | 0.196 (0.065) |
| $\sigma_\eta$ | | | 1.733 (0.238) | 2.113 (0.414) |
| $\sigma_{\eta\zeta}$ | | | 0.136 (0.039) | 0.176 (0.052) |
| $\boldsymbol{\Theta}$: | | | | |
| $\theta_1$ | | | 0.962 (0.083) | 0.748 (0.091) |
| $\theta_2$ | | | 2.838 (0.029) | 2.364 (0.037) |
| $\theta_3$ | | | 4.552 (0.012) | 5.927 (0.045) |
| $\boldsymbol{\Pi}$: | | | | |
| $\pi_1$ | | | 0.523 (0.255) | 0.322 (0.181) |
| $\pi_2$ | | | 0.238 (0.173) | 0.193 (0.127) |
| Instrument variable: | | | | |
| $\text{pos}_{t-1}$ | | | 0.885 (0.096) | 0.766 (0.127) |

*Note.* The values reported for parameters $V_\xi$, $\Theta$, and $\Pi$ in the complete data case are means of the daily estimates.

introducing a latent categorical variable with $M$ categories (Rutz et al. 2012). Position can assume any of these $M$ categorical values with a probability $\Pi = \{\pi_1, \pi_2, \ldots, \pi_M\}$, $\sum_{m=1}^{M}$. More specifically

$$\text{pos}_{tk} = \boldsymbol{\Theta}'_{kt}\gamma + \mathbf{Z}'_{kt}\delta + \eta_{kt}, \tag{14}$$

where $Z_{kt}$ captures the observed instruments. The stochastic part of $\text{pos}_{kt}$ is captured by $\Theta \sim$ Multinomial($\Pi$), which is exogenous and $\eta$, which is endogenous.[12] The errors $(\eta_{kt}, \epsilon_{kt})'$ are MVN distributed in the following manner:

$$\begin{pmatrix}\eta_{tk}\\ \zeta_{tk}\end{pmatrix} = \text{MVN}\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_\eta & \zeta_{\eta\xi}\\ \sigma_{\eta\zeta} & \sigma_\zeta\end{pmatrix}\right]. \tag{15}$$

In our analysis we use lagged position as an IV, which is similar to the approach adopted by Rutz et al. (2012). Similar to Rutz et al. (2012), we estimate the parameters $\Pi, \Theta$, and $V_\xi$ jointly for the entire observation period when the estimation is performed on aggregate data. In the case of complete data, we exploit the richness of the data to estimate the parameters, $\Pi, \Theta$, and $V_\xi$, on a daily basis. Such an approach can account for variations over time, due to changes in advertisers' bids or other auction related factors, and might lead to better performance. The exact difference between estimation on the complete versus aggregate data is explained in detail in the online appendix (available as supplemental material at http://dx.doi.org/10.1287/mksc.2014.0884).

### 4.4. Estimation Results

We estimate both the HB model and HB model with LIV (HB-LIV) to draw comparisons between the two methods. A sample size of 200 ads is chosen for estimating the parameters. We make this choice primarily for computational convenience as estimating the model on disaggregate data takes a long time. The disaggregate data set contains a large number of observations, hence the estimation on the disaggregate data set is really slow.[13]

We begin with diffused priors ($\mu = 0$, $\Sigma = 100I$, $\nu = 5$, $\Delta = \nu I$) and refine them as the estimation proceeds. The exact estimation procedure is outlined in the online appendix. We run the MCMC simulation for 100,000 draws; the first 50,000 samples are discarded. The MCMC chains are stationary after the burn-in period. The MCMC chains are thinned to remove autocorrelation between draws. Every tenth draw in the stationary period is used for the subsequent analysis.

The estimation results are presented in Table 4. We observe that there are significant differences in the $\boldsymbol{\mu}_\beta$ estimated on the complete and aggregate data, for both the HB and HB-LIV models. The effect of position on CTR $\mu_1$ is overestimated by 10% when the estimation is performed on aggregate data, indicating that aggregation bias exists when a HB model is used. The bias in $\mu_1$ increases to 12.9% when an HB-LIV model is used. This result strongly indicates that

[12] The position $\text{pos}_{tk} = v_{ik}$ or $w_{dk}$ depending on the context, where $t$ represents the unit of time.

[13] We estimate the HB model on 25 different samples; the qualitative findings remain the same.

aggregation bias is significant in the context of SSA. This finding is of concern as the extant literature on SSA (Agarwal et al. 2011, Ghose and Yang 2009, Yang and Ghose 2010) is (to our knowledge) silent about this issue. Only very recently, Rutz and Trusov (2011) acknowledge its existence and account for it in their model. Rutz and Trusov (2011) significantly advance the SSA literature by proposing a novel methodology to address endogeneity, which also addresses aggregation to some extent. However, as demonstrated in Table 4, the aggregation bias is not completely eliminated, which points to the value of complete information. Aggregate data is prone to two primary disadvantages. (i) It does not capture intraday variations, and (ii) it is difficult to identify temporal patterns due to the limited amount of data (one observation per day). When the estimation is performed on complete data, these variations can be explicitly captured and accurately estimated underscoring the importance of richer data in SSA. Our findings demonstrate that this bias is nontrivial in practice and future SSA research should be informed about the pitfalls of aggregate data. In conjunction with prior literature, we also observe a statistically significant difference in the estimates from the HB and the HB-LIV model, confirming the need to control for endogeneity of ad position in a sponsored search. Aggregation bias acts in addition to the bias due to endogeneity, and it is just as important to correct.

We use the differences between the estimates from summary and complete data for a random sample of 5,000 exact-match keywords to compute the empirical distribution of the error ($\boldsymbol{\varepsilon} = \hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_c$) due to aggregation. This empirical distribution is used in §5 to quantify the impact of aggregation bias on search engine and advertiser revenues.

## 5. Managerial Implications

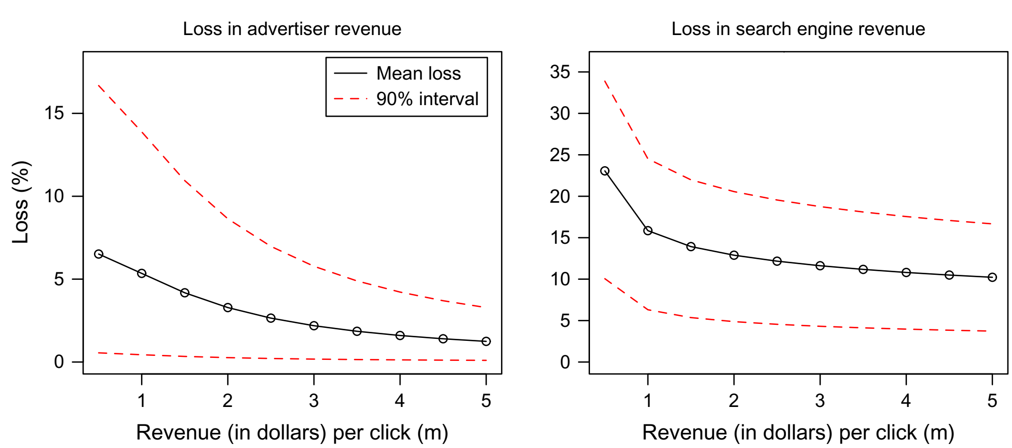
In the previous section, we have shown the existence of aggregation bias using the HB and HB-LIV models. In this section, we discuss the managerial implications of the bias. The analysis presented in §3.2 provides evidence that aggregation bias can lead to a decrease in the search engine and possibly an advertiser's revenue. Here, we extend the analysis to characterize the loss, exploiting the large data set available to us. Suppose $m$ is the expected revenue *per-click*, then the advertiser's profit per-impression is given by

$$\Pi = \mathrm{CTR}(\mathrm{CPC}) \times \{m - \mathrm{CPC}\}. \tag{16}$$

It is easy to see that the trade-off between bidding high to get more clicks (CTR increases with CPC) and bidding low to earn greater profit per click. The optimal bids can be computed by substituting the expression for CTR (as a function of advertisers CPC) into the profit function. Unfortunately, our data do not contain bidding information (CPC), and as a result, we use estimates commonly found in the extant literature to illustrate the magnitude of the loss. We assume that the relationship between position and CPC is given by $\mathrm{pos} = e^{2(1-\mathrm{CPC})}$ (Ghose and Yang 2009) and the CTR for the average keyword is a logit with $\hat{\boldsymbol{\beta}}_c = (-1.672, -0.642)'$ (Table 4). The optimal bid can be derived by substituting these relationships in Equation (16). When the advertiser uses aggregate data, his estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_c + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the estimation error computed in the previous section. We sample $\hat{\boldsymbol{\beta}}_s$ from this empirical distribution, and for each $\hat{\boldsymbol{\beta}}_s$, compute the optimal bid that maximizes the advertiser's profit in Equation (16). The bids computed using $\hat{\boldsymbol{\beta}}_c$ are greater than almost all of the bids computed using samples of $\hat{\boldsymbol{\beta}}_s$, which supports our earlier claim that aggregation bias results in an advertiser placing a lower-than-optimal bid.

Next, we use the computed CPCs to estimate the effect of aggregation bias on the advertiser's and search engine's revenues. Figure 2 shows the percent loss suffered by the advertiser due to aggregation bias. There is a great deal of variation in this loss as

**Figure 2** (Color online) Loss in Revenues Caused Due to Aggregation Bias



the valuation changes. The impact of aggregation bias is more pronounced when the advertiser's valuation for the click is low. In this situation, he bids lower. Yet even small deviations from the optimal bid can lead to significant changes in the position, leading to significantly lower payoffs. On the other hand, when the valuation is high, the bid is correspondingly higher and deviations from the optimal bid have little impact on position, and hence the loss is relatively smaller.

The effect of aggregation is considerably higher for a search engine. On average, it loses more that 11% of its payments from the advertiser as a result of aggregation bias. Lower bids, as a consequence of aggregation bias, negatively impact search engine revenues in two ways. First, lower bids imply that the search engine generates less revenue *per-click*. Second, the ad appears at a lower position due to the lower bid, which in turn leads to fewer clicks. The advertiser pays the search engine for fewer clicks and pays less for each click. In this example, we estimate that the search engine loses 1.4¢ for every impression of the representative ad.

In the preceding analysis we consider how payments from *one advertiser* to the search engine are affected by aggregation bias. Naturally, the dynamics are more complicated when we try to quantify the effect of the bias on payments by all advertisers and the overall profitability of the search engine. For example, when an advertiser bids suboptimally and moves to a lower position, another advertiser moves up to occupy the vacant position. Though the search engine loses revenue from the advertiser that moves down, it earns more from the advertiser that moves up. This reduces the overall loss suffered by the search engine. It is difficult to accurately estimate this loss as we do not have data from multiple advertisers. Therefore, we leave this as a direction for future research.

# 6. Suggested Cures

In the previous discussion, we outlined the problem of aggregation bias and some of its implications. As mentioned earlier, aggregation bias arises due to inadequate data. It might be infeasible for search engines to store and report impression level data due to the size of such data sets and potential privacy concerns. However, a search engine can provide different data to reduce the effect of aggregation. Kendall and Stuart (1977) show that information about a few moments of a distribution can be used to create a good approximation of the distribution. Accordingly, we explore various summary statistics and measure the improvements they offer over the standard aggregate data provided by search engines. We also consider a few modeling approaches that explicitly account for the variation in position and estimate the improvements offered by them.[14] We first outline these approaches and subsequently draw comparisons between them. These approaches are compared in two ways: (i) using the large representative search engine data available to us, and (ii) an extensive evaluation using simulated data.

## 6.1. Proposed Summarization Techniques

Here, we present a few approaches that might address the problem of aggregation bias.

### 6.1.1. Sample Mean.

**Different Ways of Aggregation.** An important reason for bias is the nonlinearity of the position-CTR curve. As a result, linear aggregation of the position does not yield the correct underlying response parameters. Christen et al. (1997) and Danaher et al. (2008) show that when the response is multiplicative, i.e., of the form $\alpha x_1^{\beta_1} x_2^{\beta_2}, \ldots,$ where $x_i$ are marketing mix variables, an aggregate model should use geometric means to correctly estimate the coefficients. Unfortunately, there is no analytical analog of this result when the underlying model is logit. We use both the geometric and harmonic means and empirically compare them to determine which method of aggregation works better in SSA. As the position-CTR curve is convex in nature, both of these aggregation methods might perform better than linear aggregation.

**Modeling Position Variation Using a Poisson.** We also consider a model where $V_i$ is drawn from a Poisson distribution with mean equal to the daily (arithmetic) mean $u$. The log-likelihood of observing the data in this case is given by

$$\mathrm{LL}(\boldsymbol{\beta} \mid \text{data}, \lambda) \propto \sum_{k=1}^{K}\sum_{d=1}^{D} c_{dk} \log\left\{\sum_{i=0}^{\infty} P(V_{idk}=v)p_{idk}\right\} + (n_{dk}-c_{dk})\log\left\{1-\sum_{i=0}^{\infty} P(V_{idk}=v)p_{idk}\right\}, \quad (17)$$

where $\lambda$ is a $K \times D$ matrix, and every column of $\lambda$ contains the scale parameter of the Poisson distribution for every day. The position of every impression $V_{idk}$ for keyword $k$ on the $d$th day is drawn from a Poisson distribution with $\lambda_{kd} = u_{kd}$ and the probability $P(V_{idk}=v) = \lambda_{dk}^{v} e^{\lambda_{dk}}/v!$.

**Pooling Positions Across Days.** In the previous approach, we try to model the data generating process (DGP) for position. Here, we extend this analysis to model the DGP by pooling data across multiple days.

[14] We thank the associate editor and anonymous reviewers for recommending this extension.

Assume that the $\mathbb{E}[V_i] = \mu$ and $\text{Var}(V_i) = \sigma_V^2$.[15] Using the law of large numbers, it is easy to show that the mean daily position, $W \sim \text{N}(\mu, \sigma_v^2/n)$, where $n$ is the number of impressions on day $d$. To estimate $\boldsymbol{\beta}$, we follow a two-step process. First, we estimate $\mu$ and $\sigma_V^2$ for each keyword by maximizing the following likelihood:

$$\text{LL}(\mu, \sigma_V^2 \mid \text{data}) \propto \prod_{d=1}^{D} \phi\left(\frac{w_d - \mu}{\sigma_V^2/n_d}\right),$$

where $\phi(\cdot)$ represents the p.d.f. of the standard normal distribution. Second, we approximate $V_i$ by a normal distribution such that

$$P(V_{idk} = v) = \Phi\left(v - \frac{\mu_k}{\sigma_{Vk}^2}\right) - \Phi\left(v - 1 - \frac{\mu_k}{\sigma_{Vk}^2}\right),$$

where $\Phi(\cdot)$ represents the c.d.f. of a standard normal distribution. Substituting this p.d.f. in Equation (17) gives us the overall log-likelihood. Note that this approach relies on the implicit assumption that $F_V(\cdot)$ does not change over time.

**6.1.2. Higher Order Statistics.** If a search engine provides higher order moments in addition to the mean, the aggregation bias may be significantly reduced. We discuss three approaches with increasing data requirements.

**Mean and Variance.** When the mean ($\mu_{dk}$) and variance ($\sigma_{dk}^2$) of position are provided, we assume that the position has a negative binomial distribution (NBD) with probability of success $p_{dk}$ and (real) number of trails $r_{dk}$. The NBD makes intuitive sense as the success of probability $p_{dk}$ of an NBD can be thought of as the probability of a competing advertiser placing a higher bid. The log-likelihood function is similar to Equation (17), but the distribution of ad position in this model is given by

$$P(V_{idk} = v) = \binom{v + r_{dk} - 1}{v} (1 - p_{dk})^{r_{dk}} p_{dk}^{v}, \quad \text{where}$$

$$p_{dk} = 1 - \frac{\mu_{dk}}{\sigma_{dk}^2} \quad \text{and} \quad r_{dk} = \frac{\mu_{dk}^2}{\sigma_{dk}^2 - \mu_{dk}}.$$

**Empirical Distribution.** In the preceding approaches, the variation in position is modeled in a parametric manner due to the limitation in the numbers of moments reported. A search engine can provide further moments, e.g., skewness and kurtosis, which might help the research model more accurately determine the randomness in position. For the sake of brevity, we adopt the extreme case and assume that the search engine provides the empirical distribution of the ad position as shown in Table 5, in addition to the daily summary. In this case, the variation in position can be modeled nonparametrically. The log-likelihood is given by Equation (17), where $P(V_{idk} = v)$ is provided by the empirical distribution, e.g., $P(V_{idk} = 5) = 0.29$. As pointed out earlier, this data can also be independently collected by an advertiser by periodically crawling ads from a search engine.

**Table 5 Empirical Distribution of Ad Position**

| Position | Frequency | $P(V = v)$ |
|---|---|---|
| 2 | 20 | 0.13 |
| 3 | 10 | 0.07 |
| 4 | 50 | 0.32 |
| 5 | 45 | 0.29 |
| 6 | 30 | 0.19 |

**Position-Level Summary.** Although the complete data set entirely eliminates aggregation bias, it might be difficult for search engines to provide this data due to privacy or technical concerns. Instead, the search engine can provide a position-level summary that gives *sufficient statistic* for the logit model, as we show below. The position-level summary reports the keyword performance measures separately for every position (where the ad appeared). It is easy to show that Equation (8) can be simplified to

$$\text{LL}(\boldsymbol{\beta} \mid \text{complete data}) \propto \sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{v=1}^{\infty} c_{vdk} \log p_{vdk} + (n_{vdk} - c_{vdk}) \log(1 - p_{vdk}), \quad (18)$$

where $n_{vdk}$ and $c_{vdk}$ are the number of impressions and clicks on the $k$th ad at position $v$ on the $d$th day, respectively. Because Equation (18) is identical to the likelihood function of the position-level data, a position-level summary provides sufficient statistics to correctly estimate the parameters of the model.

## 6.2. Application to Search Engine Data

We apply the summarization and modeling techniques presented earlier to the search engine data available to us. Because the data used in §4 are quite extensive and represent typical SSA data, the results presented here aim to provide real world validation for the suggested summarization techniques. Results demonstrating the performance of these techniques vis-à-vis the search engine data are presented in Table 6. Comparisons between the different summarization techniques are performed using the mean average percentage error (MAPE) in the estimates of $\beta_0$ and $\beta_1$, measured as $|\hat{\beta}_{0c} - \hat{\beta}_{0s}|/\hat{\beta}_{0c}$ and $|\hat{\beta}_{1c} - \hat{\beta}_{1s}|/\hat{\beta}_{1c}$, respectively. Note that the HB-LIV model is used for this analysis.

First, we observe that both geometric and harmonic means perform better than the arithmetic mean. This

[15] We drop the day and keyword subscripts for simplicity.

**Table 6** **Comparative Performance of Various Data Summarization Approaches on Search Engine Data**

| Method | Data requirement | MAPE($\varepsilon_0$) (%) | MAPE($\varepsilon_1$) (%) |
|---|---|---|---|
| Different ways of aggregation | | | |
| Arithmetic mean | O(X) | 53.4 | 28.9 |
| Geometric mean | O(X) | 42.1 | 17.5 |
| Harmonic mean | O(X) | 40.3 | 10.2 |
| Poisson model | O(X) | 43.3 | 22.8 |
| Pooling data across days | O(X) | 21.8 | 10.8 |
| Mean and variance | O(2X) | 15.2 | 7.8 |
| Empirical distribution | O(NX) | 9.8 | 4.2 |
| Position-level summary | o(NX) | 0 | 0 |

reduction in bias is due to the convexity of both of these aggregation techniques, which match the convexity of the *position*, i.e., the CTR curve. We also observe that the harmonic mean performs better than the geometric mean. This implies that if a search engine wants to provide only the mean position in the campaign reports, it should provide the harmonic mean of the position. This result also suggests that researchers who aggregate sponsored search data at a weekly or monthly level for lack of sufficient data or for computational reasons should use the harmonic mean for aggregation. Second, we observe that modeling techniques have mixed performance. Modeling the variation in position as a Poisson random variable does not work well as shown in Table 6. This approach does not perform well because the ad position does not follow a Poisson distribution, a hypothesis we confirm using the Neyman-Scott test. However, although pooling data across days leads to a significant decrease in the bias, the effect is heterogeneous across keywords. This approach works best when the distribution of the ad position does not change across days, e.g., when the bids are held constant in the observation period. If the bids change, then the performance of this technique drops considerably. Third, we observe that richer data that provide higher order moments lead to significant improvement in the parameter estimation. From Table 6, we observe that using both the mean and the variance to model the variation in the ad position significantly improves the estimates. Not only do we observe a reduction in the aggregation bias but there is also a considerable decrease in the error in $\hat{\beta}_{0s}$. On average, there is a more than 70% reduction in the estimation error associated with $\beta_0$ and $\beta_1$. Using the empirical distribution marginally improves the estimation performance. This summarization technique performs better than all of the preceding techniques, and the MAPE is considerably lower for $\beta_0$ and $\beta_1$. When the position-level summary is used, the aggregation bias is completely eliminated. However, this technique requires substantially more data as compared to the other techniques.

### 6.3. Application to Simulated Data

To perform better characterization of these techniques and ascertain their performance under different conditions, we turn to simulated data.[16]

**6.3.1. A Model of Sponsored Search Auctions.** We model a generalized second price auction to determine the position of the ad across different impressions. We adopt the approach presented by Abhishek and Hosanagar (2013) and assume that the competing advertisers draw their bids from a Weibull distribution, which is given by

$$F(x;\psi,c)=1-\exp\left\{-\left(\frac{x}{c}\right)^{\psi}\right\},$$

where $\psi$ is the shape parameter and $c$ is the scale parameter. A Weibull distribution is quite flexible and has been shown to capture the bid distribution better than other commonly used distributions (Abhishek and Hosanagar 2013). Variations in $\psi$ can give rise to different types of competing bid distributions, whereas $c$ changes the magnitude of the bids. Low values of $\psi$ lead to heterogeneous bids, whereas higher values of $\psi$ lead to relatively homogeneous bids. These competing bids are stochastically changed during the course of the day, leading to intraday variations in position.[17] The probability that the bids are changed before an impression is denoted by $\kappa$. Intuitively, the intraday variation in position increases in $\kappa$. In §3, we had assumed that the distribution of the ad position remains constant during the period of observation for analytical tractability. Here, we relax this assumption such that the advertiser can change his bid several times during the observation period (but not within a day), changing $F_V(\cdot)$. To model consumer choice, a logit data generating process is used and the inputs to the simulation are the coefficients of the logit model $(\beta_0,\beta_1)$ and $\kappa$. The consumer's decision to click on the ad (of the focal advertiser) conditional on its position is simulated for each query using the random utility model specified in Equation (2). The complete data record the position and the binary click decision for every impression. In addition, the daily total number of impressions, clicks, and the mean position for the ad are computed and recorded in the aggregate data. One hundred different runs are generated for every tuple $\beta_0$, $\beta_1$, and $\kappa$, where $\beta_0\in[-2,-1]$, $\beta_1\in[-1,-0.25]$, and $\kappa\in[0,1]$.

For the sake of brevity, the exact characterization of the bias is presented in the online appendix. There are three main insights from this analysis: (i) The bias (and the intraday variation) increases with $\kappa$ as shown

[16] We thank the associate editor and the anonymous reviewers for suggesting this extension.

[17] The findings are qualitatively similar for different values of $\psi$.

**Table 7** **Comparative Performance of Various Data Summarization Approaches on Simulated Data**

| Method | MAPE($\varepsilon_0$) (%) | MAPE($\varepsilon_1$) (%) |
|---|---|---|
| Different ways of aggregation | | |
| Arithmetic mean | 10.3 | 20.1 |
| Geometric mean | 8.1 | 10.6 |
| Harmonic mean | 4.4 | 5.8 |
| Poisson model | 8.1 | 9.7 |
| Pooling data across days | | |
| Not accounting for bid changes | 6.2 | 7.7 |
| Accounting for bid changes | 3.4 | 4.2 |
| Mean and variance | 2.8 | 3.6 |
| Empirical distribution | 2.1 | 2.5 |
| Position-level summary | 0 | 0 |

earlier in Proposition 2; (ii) When $F_V(\cdot)$ is not constant during the simulation period, the bias increases super-linearly in $|\beta_1|$; and (iii) The error in the estimate of intercept term increases in $|\beta_1|$ and $\kappa$.[18] We now apply the techniques proposed in §6.1 on the simulated data to obtain a greater understanding of their effectiveness.

**6.3.2. Analysis on Simulated Data.** Similar to §6.2, we use MAPE to compare the various techniques. Instead of measuring the error with respect to $\hat{\boldsymbol{\beta}}_c$, we use the simulation parameter $\boldsymbol{\beta}$.[19] We begin by presenting the overall performance of the aforementioned techniques and subsequently discuss how their performance varies across different values of $\boldsymbol{\beta}$.

The comparisons between these methodologies are presented in Table 7. We continue to observe that the harmonic mean performs considerably well as compared to the arithmetic and geometric means. The magnitude of the errors reported in Table 7 is lower in comparison to the errors reported in Table 6. This is due to the fact that we include a larger range of $\beta_0$ and $\beta_1$ in our simulation analysis than that observed in practice. Pooling data across days, to derive the underlying distribution of position, significantly reduces the bias. In the absence of adequate data, researchers and academics can use this approach to reduce aggregation bias. This approach is more effective when changes in the position distribution (e.g., due to bid changes) are explicitly accounted for in the estimation. If there are significant differences in the position distribution across days, then this approach might further increase the bias.

Daily summaries that report more data, e.g., the daily mean and variance significantly decrease the estimation error in the simulations. The bias when the daily mean and variance are reported is quite similar to the bias when data is pooled across days. Remember, however, that pooling data across days is effective only when changes in $F_V(\cdot)$ are explicitly accounted for, which can be challenging in a real world setting. Providing the daily empirical distribution of position further decreases the estimation error, but the incremental benefit is small. Not surprisingly, the position-level summary has no bias.

Figure 3 demonstrates the bias for different techniques. Clearly, the magnitude of the bias increases nonlinearly for all techniques as $\beta_1$ increases. Furthermore, we observe that using the empirical distribution outperforms all of the other techniques, while using the harmonic mean leads to the least amount of bias when only the daily means are reported.

### 6.4. Summary of Results

The empirical analysis presented here has two distinct themes. The first suggests that the provision of better data by a search engine can lead to a significant reduction in the aggregation bias. As we show in the preceding sections, it is incentive compatible for the search engine to provide better data to advertisers. The appropriate data set should be determined as a trade-off between the loss due to aggregation and the costs associated with providing richer data to advertisers. If a search engine can provide only daily means, it should report the harmonic mean. On the other hand, if a search engine is at liberty to provide any data, it should provide the position-level summary. The second theme points to steps that can be taken by an advertiser or a researcher to explicitly account for the variation in position using modeling techniques (e.g., pooling data across days) or collecting additional information (e.g., crawling the search engine to generate the empirical distribution). A unilateral reduction in the aggregation bias can be profitable for an advertiser as shown in Corollary 1.

## 7. Conclusions

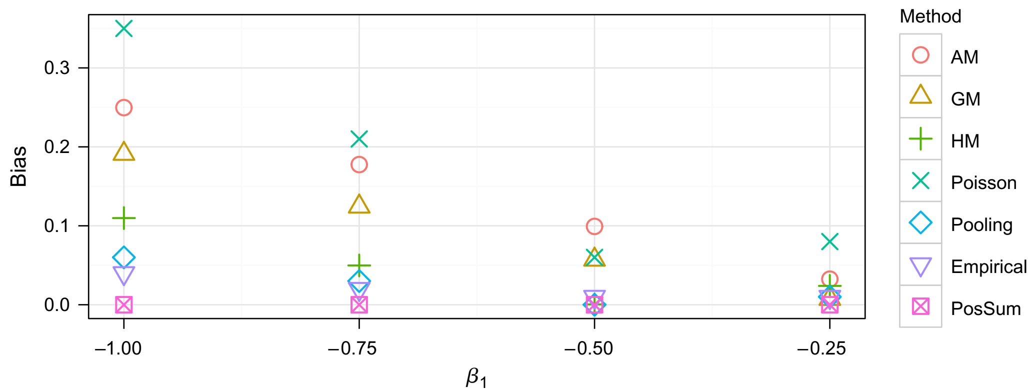
Search engine advertising is fast emerging as an important and popular medium of advertising for several firms. The medium offers rich data for advertisers on consumer click and conversion behavior. As a result, there has been considerable interest among practitioners and researchers in analyzing SSA data. Several models have been proposed to study consumer behavior and inform advertiser strategies.

This paper makes three main contributions. First, we demonstrate the existence of aggregation bias and its effect on the equilibrium of the SSA auction. We show that equilibrium bids are lower when advertisers use aggregate data. As a result, the search engine's revenues are always lower due to the bias. Second, we use a large search engine data set, quantify the magnitude of the bias, and measure its economic impact. Third, we present various summarization techniques

[18] These results are presented in Tables 1 and 2 in the online appendix.

[19] There is no statistical difference between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_c$.

**Figure 3** (Color online) Magnitude of the Bias for Different Values of $\beta_1$ When $\beta_0 = 2.0$ and $\kappa = 0.5$




that can be used by search engines to provide better data sets to advertisers.

These findings have important managerial and economic implications. Advertisers commonly use aggregate data provided by search engines to guide their bidding strategies. Our results suggest that advertisers might not be bidding optimally in these auctions because they overestimate the clicks obtainable at a given position. This not only impacts the advertisers negatively but also leads to a reduction in the revenue of the advertiser. Given the size of the SSA industry, these losses can translate into several million dollars of lost revenue for the search engines. Our study points out that the current format of the data provided to advertisers is not adequate, and that search engines should take steps to address this problem. We recognize that it might be infeasible for search engines to store and report impression level data due to the size of such data sets and potential privacy concerns. However, these constraints do not imply that it is infeasible to provide adequate data to advertisers. We provide guidelines to search engines about the nature of data sets that can be provided to researchers and quantify the reduction in the bias that each of these techniques offer.

We also find that, as a result of aggregation bias, consumer response to other ad attributes, such as ad text or branding, may also have been incorrectly estimated. Thus advertisers must be cautious in applying the biased estimates to guide key managerial decisions such as ad design and keyword selection. In the absence of adequate data from search engines, advertisers and researchers must take into account the variation in ad position within a day. This can be determined by examining whether multiple queries are matched to a single keyword (match type is broad) and whether competitors' bids change considerably in a given day. If the ad position for a keyword is somewhat stable across impressions in a day, the bias is likely to be low and existing random utility models can be applied on aggregate data.

Our study primarily demonstrates the existence and direction of aggregation bias in the coefficient of position and identifies some economic consequences of this bias. An interesting and related issue is how aggregation affects ad attributes such as wordographics, the presence of brand information, ad creativity, etc., and whether their coefficients also suffer from aggregation bias. In this paper, the effect of ad attributes is subsumed in the intercept term as we do not have data on ad attributes. A richer data set that contains ad characteristics might help in a more extensive analysis of this issue. Another direction for future research is building models that endogenize the variation in position. This variation in position can be modeled using probabilistic models or structural methods.

SSA presents an exciting opportunity to understand consumer behavior and drivers of firms' advertising strategy. Through this paper we hope to inform the practitioners about the inadequacies of the data standards commonly used in SSA so that they can take steps to address these problems. We also identify issues with some common SSA modeling techniques so that subsequent research in this emerging area can be thus informed.

## Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2014.0884.

## Acknowledgments

The authors thank the major search engine for providing the data used in this research and the reviewers for providing helpful suggestions to improve the paper. The authors are also grateful to Ashish Agarwal, Alessandro Arlotto, Eric Bardlow, and Raghuram Iyengar for insightful discussions. This research was partly supported by the Jay H. Baker Retailing Center.

## Appendix. Proofs of Propositions

### Proof of Lemma 1

We use the following result from Muller and Stoyan (2002, p. 27) to prove this result. Let $V_1, \ldots, V_n$ be iid random variables and $f_1, \ldots, f_n$ measurable real functions. Define the function $\bar{f}$ by

$$\bar{f}(v) = \frac{1}{n}\sum_{i=1}^{n} f_i(v).$$

Then,

$$\sum_{i=1}^{n} \bar{f}(V_i) \leq_{\text{cx}} \sum_{i=1}^{n} f_i(V_i).$$

Using this result we now prove that $W_n \leq_{\text{cx}} V$, where $W_n$ is the average position when there are exactly $n$ impressions on the day. Let $f_i(v) = v/(n-1)$ for all $i = 1, \ldots, n-1$ and $f_n(v) \equiv 0$.

$$\bar{f}(v) = \frac{1}{n}\sum_{i=1}^{n-1}\frac{v}{n-1} = \frac{v}{n}.$$

Since

$$\sum_{i=1}^{n}\bar{f}(V_i) = \frac{1}{n}\sum_{i=1}^{n} V_i, \quad \text{and}$$

$$\sum_{i=1}^{n} f_i(V_i) = \frac{1}{n-1}\sum_{i=1}^{n-1} V_i \;\Rightarrow\; \frac{1}{n}\sum_{i=1}^{n} V_i \leq_{\text{cx}} \frac{1}{n-1}\sum_{i=1}^{n-1} V_i.$$

Proceeding in a recursive manner

$$\frac{1}{n-1}\sum_{i=1}^{n-1} V_i \leq_{\text{cx}} \frac{1}{n-2}\sum_{i=1}^{n-2} V_i,$$

$$\cdots$$

$$\frac{V_1+V_2}{2} \leq_{\text{cx}} V$$

$$\Rightarrow\; W_n = \frac{1}{n}\sum_{i=1}^{n} V_i \leq_{\text{cx}} \frac{1}{n-1}\sum_{i=1}^{n-1} V_i \leq_{\text{cx}} \cdots \leq_{\text{cx}} \frac{V_1+V_2}{2} \leq_{\text{cx}} V. \quad \square$$

Let $g$ be any convex function

$$\text{As} \quad \mathbb{E}[g(W)] = \sum_{i=1}^{\infty} g(W_n)P(n), \quad \text{we have}$$

$$\text{or} \quad \mathbb{E}[\mathbb{E}[g(W)]] = \mathbb{E}\left[\sum_{i=1}^{\infty} g(W_n)P(n)\right],$$

$$\text{or} \quad \mathbb{E}[\mathbb{E}[g(W)]] = \sum_{i=1}^{\infty}\mathbb{E}[g(W_n)]P(n) \leq \sum_{i=1}^{\infty}\mathbb{E}[g(V)]P(n)$$

$$= \mathbb{E}[g(V)]\sum_{i=1}^{\infty} P(n) = \mathbb{E}[g(V)],$$

as $P(n)$ is a probability measure. Therefore $W \leq_{\text{cx}} V$. $\square$

**Proof of Proposition 1**

Let $Y_i$ denote an indicator variable that equals 1 if the $i$th impression resulted in a click and zero otherwise. We assume that the clicks are independent of each other and hence $Y_i$s are independent. The log likelihood of observing the data set with a total of $I$ ad impressions is given by

$$\text{LL}(\boldsymbol{\beta} \mid \text{complete data}) = \sum_{i=1}^{I} y_i \log p_i + (1-y_i)\log(1-p_i). \tag{19}$$

The first-order condition (FOC) for Equation (19) is as follows:

$$\frac{\partial\,\text{LL}}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{I}\{y_i(1-p_i) - (1-y_i)p_i\}\mathrm{x}_i' = 0,$$

$$= \sum_{i=1}^{I}\{y_i - p_i\}\mathrm{x}_i' = 0,$$

where $\mathrm{x}_i = (1 \quad v_i)'$. Since we know that $\text{LL}(\beta \mid \text{data})$ is a convex function in $\beta$ (Hayashi 2000) this FOC gives us the following two equations:

$$C = \sum_{i=1}^{I} p_i, \tag{20}$$

$$\sum_{i=1}^{I} y_i v_i = \sum_{i=1}^{I} v_i p_i. \tag{21}$$

Dividing Equation (20) by $I$ we get

$$\text{obsctr} = \frac{C}{I} = \frac{1}{I}\sum_{i=1}^{I} p_i. \tag{22}$$

If the number of impressions on day $d$ is $n_d$ and the number of clicks is $c_d$. The log-likelihood of observing the aggregate data for $D$ days is given by

$$\text{LL}(\boldsymbol{\beta} \mid \text{aggregate data}) = \sum_{d=1}^{D} c_d \log p_d + (n_d - c_d)\log(1-p_d). \tag{23}$$

Evaluating the first-order condition for Equation (23)

$$\frac{\partial\,\text{LL}}{\partial\boldsymbol{\beta}} = \sum_{d=1}^{D}\{c_d(1-p_d) - (n_d - c_d)p_d\}\bar{\mathrm{x}}_d' = 0,$$

$$= \sum_{d=1}^{D}\{c_d - n_d p_d\}\bar{\mathrm{x}}_d' = 0,$$

where $\bar{\mathrm{x}}_d = (1\; w_d)'$, which in turn gives us

$$\sum_{d=1}^{D} c_d = C = \sum_{d=1}^{D} n_d p_d, \tag{24}$$

$$\sum_{d=1}^{D} w_d c_d = \sum_{d=1}^{D} n_d w_d p_d. \tag{25}$$

Dividing Equation (24) by $I$ we get

$$\text{obsctr} = \frac{C}{I} = \sum_{d=1}^{D}\frac{n_d p_d}{I}. \tag{26}$$

Note that the obsctr is the same in both cases.

Assuming $I$ is large we can apply Chevychev's law of large numbers to rewrite Equation (22) as

$$\text{obsctr} = \mathbb{E}\left[\frac{e^{\hat{\beta}_{0,\text{c}}+\hat{\beta}_{1,\text{c}}V}}{1+e^{\hat{\beta}_{0,\text{c}}+\hat{\beta}_{1,\text{c}}V}}\right]. \tag{27}$$

If we have a large enough observation period, Equation (26) can be simplified as

$$\text{obsctr} = \sum_{d=1}^{D}\frac{n_d e^{\hat{\beta}_{0,\text{s}}+\hat{\beta}_{1,\text{s}}w_n}}{I(1+e^{\hat{\beta}_{0,\text{s}}+\hat{\beta}_{1,\text{s}}w_n})}$$

$$= \mathbb{E}\left[\frac{e^{\hat{\beta}_{0,\text{s}}+\hat{\beta}_{1,\text{s}}W}}{1+e^{\hat{\beta}_{0,\text{s}}+\hat{\beta}_{1,\text{s}}W}}\right].$$

As the observed ctr, obsctr is the same in both cases

$$\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}\right]=\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}{1+e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}\right]. \tag{28}$$

As the convex ordering in Lemma 1 holds and logit is a convex in position for $\beta_0 < 0$ (which is a reasonable assumption in SSA as the CTR on the topmost position is less that 0.2 in all cases), it follows from the definition of convex ordering that

$$\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}\right]\geq\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}{1+e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}\right], \tag{29}$$

if $\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}}_s$. The equality holds only when $F_V(\cdot) = F_W(\cdot)$, i.e., which hold only under a few special cases (e.g., when there is exactly one impression every day or there is no intraday variation in position). Since Equations (28) and (29) cannot simultaneously be true and Equation (28) always holds we prove by contradiction that $\hat{\boldsymbol{\beta}}_c \neq \hat{\boldsymbol{\beta}}_s$. □

**Proof of Propositions 2(i) and 2(ii)**

(i) Since we know that

$$\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}\right]=\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}{1+e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}\right],$$

and by definition of convex order

$$\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}V}}\right]\geq\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}W}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}W}}\right],$$

we can say that

$$\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}{1+e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}W}}\right]\geq\mathbb{E}\left[\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}W}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}W}}\right].$$

As this result holds for any distribution of $W$, this relation should hold pointwise for the two functions. Hence,

$$\frac{e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}x}}{1+e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}x}}\geq\frac{e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}x}}{1+e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}x}}. \quad \square$$

(ii) The preceding relationship implies that

$$e^{\hat{\beta}_{0,s}+\hat{\beta}_{1,s}x} \geq e^{\hat{\beta}_{0,c}+\hat{\beta}_{1,c}x} \quad \forall x \geq 0$$

$$\Rightarrow \quad \hat{\beta}_{1,s} > \hat{\beta}_{1,c}. \quad \square$$

**Proof of Proposition 2(iii)**

If the true model is $U = \beta_{0k} + \beta_{1k}V_i + \epsilon$, but we use the aggregate data for estimation, then we have

$$\begin{aligned} U &= \beta_0 + \beta_1(W_d + Z_i) + \epsilon, \\ &= \beta_0 + \beta_1 W_d + \beta_1 Z_i + \epsilon. \end{aligned}$$

The disturbance term in this case is $\beta_1 Z_i + \epsilon$. Assuming that $\beta_1 Z_i + \epsilon$ is approximately logistically distributed, $\hat{\beta}_{1,s} \to \beta_1/\sqrt{\mathrm{Var}(\beta_1 Z_i)+1}$, where $\sigma_\epsilon = 1$ for the purpose of identification. The variance $\mathrm{Var}(\beta_1 Z_i)$ can be computed in the following manner:

$$\begin{aligned} \mathrm{Var}(Z_i) &= \mathrm{Var}(V_i - W_d), \\ &= \mathrm{Var}(V_i) + \mathrm{Var}(W_d) - 2\,\mathrm{Cov}(V_i, -W_d), \\ &= \sigma_V^2 + \{\sigma_V^2 + \mu_V^2\}\mathbb{E}\left[\frac{1}{N}\right] + 2\sigma_V^2\,\mathbb{E}\left[\frac{1}{N}\right], \\ &= \sigma_V^2 + \{3\sigma_V^2 + \mu_V^2\}\mathbb{E}\left[\frac{1}{N}\right]. \end{aligned}$$

Hence,

$$\hat{\beta}_{1,s} \to \frac{\beta_1}{\sqrt{\beta_1^2\{\sigma_V^2 + (3\sigma_V^2 + \mu^2)\phi\} + 1}},$$

where $\phi = \mathbb{E}[1/N]$.

**Proof of Proposition 3**

(i) We assume that advertiser $j$ moves to position $j'$ if he uses aggregate data. Since $j'$ ends up higher than $j$ in equilibrium, $b_j'(s_j, j', h) < b_{j'}(s_{j'}, j', h)$ or

$$\begin{aligned} &\frac{1}{\alpha_{j'-1}}\big((\alpha_{j'-1} - \alpha_{j'})s_{j'} + \alpha_{j'} b_{j'+1}\big) \\ &\qquad > \frac{1}{\alpha'_{j'-1}}\big((\alpha'_{j'-1} - \alpha'_{j'})s_j + \alpha'_{j'} b_{j'+1}\big), \end{aligned} \tag{30}$$

which also implies that his bid in the equilibrium decreases, i.e., $b_j' < b_{j'}$. In addition, as $b_j'$ is lower than $b_j$, $j' \geq j$. The bids for all of the advertisers in this case are as follows:

$$\begin{aligned} b_i' &= \frac{1}{\alpha_{i-1}}\left(\sum_{k=i}^{K}(\alpha_k - \alpha_{k+1})s_k\right) \quad \text{for } i > j', \\ b_j' &= \frac{1}{\alpha'_{j'-1}}\left((\alpha'_{j'-1} - \alpha'_{j'})s_j + \frac{\alpha'_{j'}}{\alpha_{j'}}\sum_{k=i}^{j'+1}(\alpha_k - \alpha_{k+1})s_k\right), \\ b_i' &= \frac{1}{\alpha'_{i-2}}\left(\sum_{k=i}^{j'}(\alpha_{i-2} - \alpha_{i-1})s_i + \alpha_{j'+1} b_j'\right) \quad \text{for } j' \geq i > j, \\ b_i' &= \frac{1}{\alpha'_{i-1}}\left(\sum_{k=i}^{j-1}(\alpha_{i-1} - \alpha_i)s_i + \alpha_{j-1} b'_{j+1}\right) \quad \text{for } i < j. \end{aligned}$$

As $h$ do not change for advertisers below $j'$, therefore their bids remain the same. It is easy to see that advertisers $j+1$ to $j'$ end up bidding higher, i.e., $b_i' > b_i$ for $j < i \leq j'$ though they move up by one position.

We now show that the bid associated with every position $\geq j'$ is lower than when complete data is used. Consider the bid $b'_{j'}$, placed by advertiser $j'$ who occupies position $j'-1$. We start off by showing that $b'_{j'} < b_{j'-1}$.

$$\begin{aligned} &b'_{j'} - b_{j'-1} \\ &= \frac{1}{\alpha_{j'-2}}\Big((\alpha_{j'-2} - \alpha_{j'-1})\underbrace{(s_{j'} - s_{j'-1})}_{<0 \text{ by construction}} + \alpha_{j'-1}\underbrace{(b_j' - b_{j'})}_{<0 \text{ by assumption}}\Big) < 0. \end{aligned}$$

So the bid for position $j'-1$ is lower than the bid in the complete case. Proceeding in a similar manner it is easy to show that bids for all positions above $j'-1$ will also be lower. This implies that $b_i' < b_i$ for $i < j$ (these ads do not change position). To summarize—$b_j' < b_j$, $b_i' < b_i$ for $i < j$, $b_i' > b_i$ for $j < i \geq j'$ and $b_i' = b_i$ for $i > j'$.

(ii) As all of the bids are either the same or lower in this case, search-engine revenue is lower ($\Pi_S^{A2} < \Pi_S^{C}$). The payoff of advertiser $j$ is lower as any deviation from the optimal bidding policy results in a strictly lower payoff. Advertisers

$j'+1$ onwards receive the same payoff and all other advertisers are better off due to suboptimal bid by advertiser $j$. Starting off with advertiser $j'$,

$$\begin{aligned}\Pi_{j'}^{A2}-\Pi_{j'}^{C} &= (\alpha_{j'-1}-\alpha_{j'})s_{j'}-\alpha_{j'-1}b'_{j}+\alpha_{j'}b'_{j'+1}\\ &= (\alpha_{j'-1}-\alpha_{j'})s_{j'}+\alpha_{j'}b_{j'+1}\\ &\quad -\frac{\alpha_{j'-1}}{\alpha'_{j'-1}}((\alpha'_{j'-1}-\alpha'_{j'})s_j+\alpha'_{j'}b_{j'+1})\\ &> 0 \quad \text{(by Equation 30).}\end{aligned}$$

Similarly, using induction we can show that $\Pi_i^{A2}>\Pi_i^{C}$ for advertiser $i$, s.t. $j<i\le j'$. For $i<j$, the revenues remain the same but the payment to the search engines is lower, hence their payoff are higher for these advertisers too ($\Pi_i^{A2}>\Pi_i^{C}$, $i<j$).

**Proof of Proposition 4**

(i) If all advertisers use the same incorrect estimate of $\alpha_i$, the optimal bidding policy is the one proposed by Edelman et al. (2007). They just use $\alpha'_i$ instead of $\alpha_i$ to compute the optimal bids. We can show by induction that $b'_j\le b_j$ for all advertisers.

*Step* 0. Let $b'_{K+1}=b_{K+1}=0$.
*Step* 1. $b'_K=s_K(1-\alpha'_K/\alpha'_{K-1})<s_K(1-\alpha_K/\alpha_{K-1})=b_K$.
Assuming $b'_{j+1}<b_{j+1}$,
*Step* $j$.

$$\begin{aligned}b'_j &= s_j\left(1-\frac{\alpha'_j}{\alpha'_{j-1}}\right)+\frac{\alpha'_j}{\alpha'_{j-1}}b'_{j+1}\\ &< s_j\left(1-\frac{\alpha'_j}{\alpha'_{j-1}}\right)+\frac{\alpha'_j}{\alpha'_{j-1}}b_{j+1}\\ &< s_j\left(1-\frac{\alpha_j}{\alpha_{j-1}}\right)+\frac{\alpha_j}{\alpha_{j-1}}b_{j+1}\\ &< b_j.\end{aligned}$$

Hence, $b'_j<b_j\,\forall\, j\le K$.

(ii) Because all advertisers occupy the same position as they did earlier and pay less, search engine profits are lower ($\Pi_S^{A1}<\Pi_S^{C}$). The advertisers' payoff in the case are higher: $\Pi_i^{A1}=\alpha_i(s_i-b'_{i+1})>\alpha_i(s_i-b_{i+1})=\Pi_i^{C}$ from Proposition 5(i).

**Proof of Proposition 5**

Assume that advertiser $j$ uses aggregate data and appears at a position $j'$. Let the equilibrium bids be denoted by $b'_1,\ldots,b'_K,0$. In the equilibrium, $\alpha_{j'}(s_j-b_{j'})>\alpha_i(s_j-b_{i+1})$ $\forall\, i\ne j'$. Now suppose that advertiser $j$ does not have access to aggregate data and overestimates $\alpha_i/\alpha_{i-1}$. As a result, he bids lower and moves to position $j''\ge j'$. Following the argument in the Proof of Proposition 4, the bids for all positions $i$, $i\le j''$ decrease and all *other* advertisers are better off. As the bids are (weakly) lower, the search-engine revenue is lower. The payoff to advertiser $j$ is $\alpha_{j''}(s_j-b_{j''})<\alpha_{j'}(s_j-b_{j'})$ as he found it optimal to bid for position $j'$ when he could correctly estimate the CTR. This implies that he is worse off using aggregate data.

## References

Abhishek V, Hosanagar K (2013) Optimal bidding in multi-item multislot sponsored search auctions. *Oper. Res.* 61(4):855–873.

Agarwal A, Hosanagar K, Smith MD (2011) Location, location, location: An analysis of profitability and position in online advertising markets. *J. Marketing Res.* 48(6):1057–1073.

Ali K, Scarr M (2007) Robust methodologies for modeling web click distributions. *Proc. 16th Internat. Conf. World Wide Web, WWW '07* (ACM, New York), 511–520.

Allenby GM, Rossi PE (1991) There is no aggregation bias: Why macro logit models work. *J. Bus. Econom. Statist.* 9(1):1–14.

Animesh A, Ramachandran V, Viswanathan S (2010) Research note–Quality uncertainty and the performance of online sponsored search markets: An empirical investigation. *Inform. Systems Res.* 21(1):190–201.

Christen M, Gupta S, Porter JC, Staelin R, Wittink DR (1997) Using market-level data to understand promotion effects in a nonlinear model. *J. Marketing Res.* 34(3):322–334.

Craswell N, Zoeter O, Taylor M, Ramsey B (2008) An experimental comparison of click position-bias models. *Proc. Internat. Conf. Web Search Web Data Mining, WSDM '08* (ACM, New York), 87–94.

Danaher PJ, Bonfrer A, Dhar S (2008) The effect of competitive advertising interference on sales for packaged goods. *J. Marketing Res.* 45(2):211–225.

Ebbes P, Wedel M, Steerneman T, Boeckenholt U (2005) Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quant. Marketing Econom.* 3(4):365–392.

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.

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: Pricing ads to context. SSRN eLibrary. http://papers.ssrn.com/abstract_id=10214151.

Good IJ, Mittal Y (1987) The amalgamation and geometry of two-by-two contingency tables. *Ann. Statist.* 15(2):694–711.

Gupta S, Chintagunta P, Kaul A, Wittink DR (1996) Do household scanner data provide representative inferences from brand choices: A comparison with store data. *J. Marketing Res.* 33(4):383–398.

Hayashi F (2000) *Econometrics* (Princeton University Press, Princeton, NJ).

Jeziorski P, Segal I (2009) What makes them click: Empirical analysis of consumer demand for search advertising. Working paper, Stanford University, Palo Alto, CA. http://web.stanford.edu/~isegat/ads.pdf.

Kelejian HH (1995) Aggregated heterogeneous dependent data and the logit model: A suggested approach. *Econom. Lett.* 47(3–4):243–248.

Kendall M, Stuart A (1977) *The Advanced Theory of Statistics*, Vol. 1. *Distribution Theory* (Macmillan Publishing Co., New York).

Liu D, Chen J, Whinston AB (2009) Competing keyword auctions. *J. Marketing* 73(4):125–141.

Muller A, Stoyan D (2002) *Comparison Methods for Stochastic Models and Risks* (John Wiley, Chichester, UK).

Narayanan S, Nair HS (2013) Estimating causal installed-base effects: A bias-correction approach. *J. Marketing Res.* 50(1):70–94.

Neslin SA, Shoemaker RW (1989) An alternative explanation for lower repeat rates after promotion purchases. *J. Marketing Res.* 26(2):205–213.

Russell GJ, Kamakura WA (1994) Understanding brand competition using micro and macro scanner data. *J. Marketing Res.* 31(2):289–303.

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.

Rutz OJ, Trusov M (2011) Zooming in on paid search ads—A consumer-level model calibrated on aggregated data. *Marketing Sci.* 30(5):789–800.

Rutz OJ, Bucklin RE, Sonnier GP (2012) A latent instrumental variables approach to modeling keyword conversion in paid search advertising. *J. Marketing Res.* 49(3):306–319.

Steenkamp J-BEM, Nijs VR, Hanssens DM, Dekimpe MG (2005) Competitive reactions to advertising and promotion attacks. *Marketing Sci.* 24(1):35–54.

Weber TA, Zheng ZE (2007) A model of search intermediaries and paid referrals. *Inform. Systems Res.* 18(4):414–436.

Wooldridge JM (2001) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

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.

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

Yatchew A, Griliches Z (1985) Specification error in probit models. *Rev. Econom. Statist.* 67(1):134–139.