A Hierarchical Model for Value Estimation in Sponsored Search*

Eric Sodomka†
Brown University
Providence, RI 02912
sodomka@cs.brown.edu

Sébastien Lahaie
Yahoo! Research
New York, NY 10018
lahaies@yahoo-inc.com

Dustin Hillard
Yahoo! Labs
Santa Clara, CA 95054
dhillard@yahoo-inc.com

ABSTRACT

Sponsored search is a form of online advertising where advertisers bid for placement next to search engine results for specific keywords. As search engines compete for the growing shares of online ad spend, it becomes important for them to understand what keywords advertisers value most, and what characteristics of keywords drive value. In this paper we propose an approach to keyword value prediction that proceeds in two steps. We first estimate values on high-volume keywords based on advertiser bids, assuming rational bidding behavior. We then fit a hierarchical model on top of these estimates, drawing on demographic and textual features of keywords and taking advantage of the hierarchical structure of sponsored search accounts. The predictive quality of our model is evaluated on fifty high-spending advertising accounts on a major search engine. In the process of fitting our model we uncover evidence that advertiser utility is additive across keywords, an implicit assumption in the literature to date. Our evaluation shows that our model outperforms several baselines for value inference, and that improvements are even more pronounced for large accounts. Besides the general evaluation of advertiser welfare, our approach has potential applications to keyword and bid suggestion.

1. INTRODUCTION

As consumer attention continues to shift online, so do the marketing budgets and resources of advertisers. Online advertising spend is set to surpass newspaper advertising in 2011, with a significant fraction going to advertising on search engines, a segment commonly known as sponsored search.[1] Sponsored search refers the practice of displaying ads alongside search results whenever a user issues a query. Advertisers develop campaigns by selecting the keywords they wish to advertise on and setting bids for those keywords. The exact placement and cost of the ads is then determined via an auction process. On today's modern platforms, advertisers can also set targeting criteria to show their ads to users only from specific locales or demographics, and can specify how closely the user query needs to match their keyword.

While it is a simple matter for small and large businesses alike to set up sponsored search accounts, developing a high-performance online ad campaign is a complex task. The process of selecting bids, keywords, and targeting options to optimize returns is one of continuous refinement, so much so that a vibrant *search engine marketing* (SEM) industry has emerged to assist advertisers in managing their online campaigns [11]. The search engines themselves compete on advertiser experience by introducing new targeting capabilities and improving their ad-to-keyword matching technologies [6]. In this competitive environment, evaluating advertiser welfare, identifying the drivers of value behind keywords, and leveraging these insights to enhance online ad campaigns are therefore all questions of interest to search engines.

This paper proposes a hierarchical linear model to infer an advertiser's value per click on search terms. Our choice of model is based on the hedonic hypothesis that keywords are valued for their characteristics, such as the demographic profile of users they attract. The growing trend towards geographic, demographic, and even behavioral targeting is evidence that these characteristics can be closely linked to returns [23]. Our model is developed in two steps. We first apply existing techniques from microeconomics to obtain value per click estimates given advertiser bids. Next, we fit a hierarchical linear model to these value estimates. The result is a model that can provide value inferences for any new search term, or even a collection of terms, as long as the necessary predictive features are available.

The two techniques behind our approach can be traced back to the microeconomic and marketing literatures on sponsored search, respectively. To derive value estimates from bids, we use the methodology outlined in the recent work of Athey and Nekipelov [4], with certain changes to the implementation details. Our hierarchical model, meanwhile, follows in the footsteps of similar models in marketing analyses of individual campaigns from the advertiser's perspective [14, 19]. Building such models on the search engine side can be an insightful exercise because, while the search engine may not have conversion data, it may have much finer-grained information about the user traffic that visits the ads.

Our work makes two contributions that distinguish it from the received literature. From the search engine's perspective, no previous work has examined advertiser values *across* their accounts. The standard model of sponsored search,

*We are submitting this paper for confidential review to ACM EC 2011.

†This work was done while the author was an intern at Yahoo! Research.

[1] www.emarketer.com/Article.aspx?R=1008126

and empirical analyses based on it, all focus on keyword-level bidding [4, 10, 22]. This makes the implicit assumption that advertiser utility is additive across keywords, so that each can be considered in isolation. In our analysis we allow for synergies between keywords by modeling utility within the family of *constant elasticity of substitution* (CES) functions. Nevertheless, we find that our estimated values do respect the basic implications of additive utility: bids rarely exceed estimated values, and values almost always exceed costs per click. Also, our hierarchical model achieves the best fit for utilities that are close to additive. This provides evidence, lacking so far in the literature, that additive utility is indeed a suitable assumption for sponsored search.

Our second and main contribution is the predictive model for advertiser values. From the search engine's perspective, no previous work has developed any value models that can extend beyond head terms. Our model draws on demographic features such as the age, gender, and income profile of users that search on a term, as well as textual features such as the length of ad titles and descriptions. We also include competitive features such as the number of ads shown at the top of the page, which should in principle correlate with bids but not with values. We consider two baselines: the average value in a term's ad group (a collection of related terms within a campaign), and an unpooled linear model fit to ad groups. Using cross-validation to assess predictive quality, we find that our model outperforms both baselines for value prediction, and that the improvement is more pronounced for large accounts. For bid prediction, our results suggest an improvement, but we do not have enough accounts to confirm a significant difference with the baselines.

Besides the general evaluation of advertiser welfare, we see two potential applications of the value estimates provided by our approach: keyword and bid suggestion. Search engines typically provide keyword suggestion tools to help advertisers augment their campaigns. The current state of the art provides keyword suggestions based on statistical and semantic similarities using a campaign's initial set of keywords [9], but we have not found any research on how to filter and rank keyword suggestions according to value to the advertiser. The value estimates from our model provide a principled ranking criterion. The only existing criterion we are aware of that takes into account bid information is the average cost per click of the keyword; in our evaluation we show that, for value prediction, our model significantly outperforms the closely related baseline of average bid.

Beyond ranking suggested keywords, our value estimates also introduce the possibility of generating bids for those keywords. The only work on bid generation we are aware of is the very recent paper by Broder et al. [7], who take a machine learning approach to directly predicting bids based purely on textual features of keywords and ads. While they report good prediction performance, their approach cannot indicate how to adapt bids if competition on a keyword rises or falls, and does not provide a criterion for ranking suggestions—high bids may indicate the most competitive keywords, rather than the most valuable. By uncovering the primitives behind advertiser behavior (i.e., values) it becomes possible to automate the complete process of keyword ranking and bidding.

The remainder of the paper is organized as follows. We next briefly review related work in more depth. In Section 2, we provide the background on sponsored search needed to follow the paper. Section 3 describes our data sets and a short exploratory analysis of advertiser behavior. Section 4 describes our utility model for advertisers and the methodology for estimating values from bids. In Section 5 we develop our hierarchical linear model and evaluate its predictive performance. Section 6 concludes with directions for improvement and future work.

Related work. As mentioned, this work relates to two strands of research from the marketing and economics literatures. It connects most directly with recent marketing research that seeks to identify drivers of conversions across keywords, such as the demographics they attract (e.g., males or females), ad performance (e.g., rank on the page), or intrinsic characteristics (e.g., whether the keyword is branded). Jansen and Sobel [15] examine the ad campaign of a large nationwide retailer and find that brand terms in the keyword or ad copy impact conversions, while Jansen and Solomon [16] find a relationship between conversions and the gender orientation of keywords in this same data set.

The paper of Rutz and Bucklin [19] is most closely related to ours in that it expressly addresses the problem of estimating conversions (and hence value) on low-volume keywords. Using data from the paid search campaign of a hotel chain, they apply several logit models to predict conversions based on features such as the presence of brand or geographic information. Our work has the same aim but we approach the problem from the perspective of the search engine.

Ghose and Yang [14] use a hierarchical logit model to examine the relationship between conversions, click-through-rate, cost per click, and keyword characteristics, using data from the ad campaign of a nationwide retailer. They find that keyword characteristics, such as the presence of brand or retailer terms, as well as rank impact conversions. The relationship between rank and conversions is in fact a recurring theme in the literature [2, 8]. In our work we model advertisers as having a fixed value per click on a keyword. The refinement to rank-dependent value estimation is an important next step in this line of research.

Our work also connects with the small but influential economic literature on equilibrium models of sponsored search. Edelman et al. [10] introduced the solution concept of envy-free equilibrium, while Varian [22] showed how it could be applied to derive bounds on values per click using actual bid data from Google. The limitation of this approach is that it assumes a static auction environment with a fixed competitors and quality scores. Athey and Nekipelov [4] recognize that this is an oversimplification, and develop a model that incorporates uncertainty in competitors and quality scores. With sufficient uncertainty, they find that values per click are point-identified, and otherwise their approach provides bounds on values. Our approach to estimating values on high-volume terms is modeled on theirs, with minor differences in how values are selected from the bounds when they are not point-identified.

2. OVERVIEW OF SPONSORED SEARCH

We now describe the process of sponsored search and the associated SEM terminology used in the paper. A sponsored search account is populated with many **campaigns**, which themselves consist of **ad groups**. Each ad group contains a set of **terms** or keywords, for example *sports shoes*, *stilettos*, *canvas shoes*, etc. A **creative** is associated with an ad group and is composed of a **title**, a **description** and a **dis-**

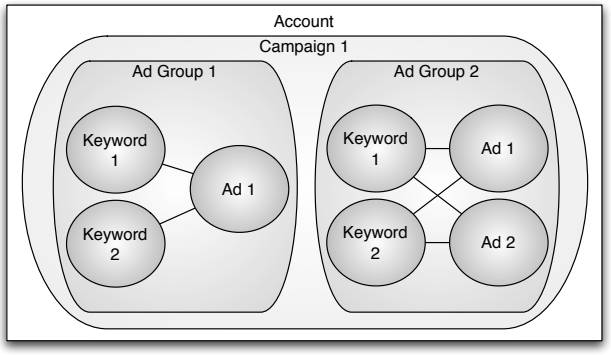


Figure 1: Hierarchical structure of a sponsored search account.

play URL. The title is typically 2–3 words in length and the description has about 10–15 words. An advertiser can in fact assign several creatives to an ad group, in which case they are rotated; for our purposes only the average ad effect matters so this is not an issue. Figure 1 illustrates the hierarchical structure of sponsored search accounts, which is shared among the leading search engines.

An advertiser can choose to use **standard** or **advanced** match for the keywords in an ad group. For example, enabling only standard match for the keyword "sports shoes", will result in the corresponding creative being shown only for that exact query. Whereas, if the keyword is enabled for advanced match, the search engine can show the same ad for the related queries "running shoes" or "track shoes". Advertisers can also set **targeting** settings for their campaigns or ad groups to specify that their ads should only be shown to users from specific locales or demographics. For example, an advertiser might target a campaign to California, with some ad groups specifically targeted to users in their thirties. The combination of matching and targeting means that, for any given query, the set of advertisers competing to be shown can vary widely.

For each of its keywords an advertiser sets a **bid**; it is also possible, and not uncommon, to set bids at the ad group level. When a user issues a query, an auction is run among eligible ads to determine page placement and pricing. The ranking of an ad is determined by a combination of its bid and **quality score**, which is meant to capture the ad's relevance to the keyword; an important ingredient in this score is the search engine's estimate of the ad's **click-through rate (CTR)**. By convention, the payment scheme is *per click*, meaning the advertiser only pays when its ads are clicked, not simply when they are shown; the price of a click is often called the **cost per click (CPC)**. The actual keyword auction mechanics are not needed to follow our work—see [10] for the details. We only note that, as one would expect, clicks and CPCs are monotonic in bid.

Finally, we also note that advertisers can specify **budgets** at the account and perhaps campaign levels. The way advertisers use budgets is hard to interpret: they represent a daily cap on spending, and so may serve more as a means of smoothing spend over time, rather than setting a hard cap on total spend. For this study we did not have access to budgets, and make the simplifying assumption that they do not bind. Previous work from both microeconomics and marketing universally ignores budgets, which suggests that significant progress can be made in modeling the sponsored search environment without taking them into account.

3. DATA DESCRIPTION

Our data set consists of Yahoo! sponsored search logs over one month in the summer of 2010. We focus on a controlled testing bucket that makes up 5% of all queries on Yahoo's search engine during this time. For each query in the data set, we have information about the *auction*, *displayed advertisers*, and *user* (who issued the query). Auction information includes the query and the number of ads displayed at the top and to the right of the page. Advertiser information includes each advertiser's bid, whether the ad was displayed via exact or advanced match, the original keyword each advertiser bid on, the displayed creatives, and the predicted CTRs decomposed into position and advertiser effects. User information includes demographics such as predicted age, gender and zip code.

We focus on accounts in the data set that are in the top 5% for both total spend and query volume. This relatively small fraction of accounts make up a substantial fraction of overall revenue, making them desirable candidates to model from the search engine's standpoint. The fact that these advertisers are high spenders may provide them with greater incentives to place optimal bids, and the fact that their ads are shown frequently may provide them with enough data to optimally bid. We take from this subset the 200 accounts with the highest bid variation within ad groups in order to remove accounts that do not exhibit interesting behavior, such as nearly constant bids across all terms. Finally, we subsample 50 of these accounts to provide more anonymity to the advertisers being modeled.

Metric	Min	Median	Mean	Max
Terms in account	13	1112	7183	80010
Ad groups in account	1	125	208	1213
Campaigns in account	1	8	14	77
Terms in ad group	1	10	35	953
Ad group bid variation	0.00	0.50	0.51	2.73
Mean distinct daily bids	1.00	3.37	4.49	28.00

Table 1: Summary statistics for the subsampled accounts in our data set.

Summary statistics for the 50 subsampled accounts are presented in Table 1. The first four rows in the table provide information on the size of accounts; accounts in our subsample tend to be larger than average. The last two rows show bid variation within an ad group (for a given snapshot in time) as well as the number of distinct bids across time. We find that bids within an ad group usually do not vary substantially. Yahoo! provides an interface for advertisers to easily place identical bids for all terms at the ad group level, so some advertisers might be reasoning about terms in an ad group as if they were a single entity, perhaps adjusting these bids periodically as more information about term-level performance is obtained. We also observe that bids are usually updated sporadically during the month. This makes estimation by revealed preference techniques problematic, since an advertiser that updates bids infrequently is not actively choosing a bundle of clicks to receive each day. In such cases, advertisers may be making decisions with respect to longer time windows, for example by setting a bid with some

expectation of clicks and costs over the course of multiple weeks. We take this frequency of bid updates into consideration when generating our models of advertiser beliefs, which are described in the following section.

4. UTILITY MODEL

We set the stage for our value estimation approach on high-volume terms by describing our utility model for advertisers. Let $i = 1, \ldots, n$ index the terms in an ad group. When bidding on a term i, an advertiser implicitly selects an expected quantity of clicks x_i (over a certain time period such as a day, week, or month). Let $p_i(x_i)$ denote the expected total cost of obtaining these clicks. We postulate that advertiser utility for the clicks obtained over terms in an ad group takes the form

$$u(x) = \left[\sum_{i=1}^{n} v_i x_i^{1-\rho}\right]^{\frac{1}{1-\rho}} - \sum_{i=1}^{n} p_i(x_i) \qquad (1)$$

The first component takes the familiar form of the *constant elasticity of substitution* (CES) utility function. It has a parameter v_i for each term i that in essence captures the relative value of clicks from the term, and a global parameter ρ that modulates the elasticity of substitution among clicks on different terms. At $\rho = 0$ the component is additive, and utility decomposes among terms so that they can be considered in isolation; as $\rho \to +\infty$ we reach pure complements.

In the usual formulation of an economic demand problem there is an explicit budget constraint. However, if we assume that utility is quasi-linear in money and that budgets do not bind (i.e., part of the advertising budget is left unspent at an optimal choice of clicks), then the budget constraint can be eliminated and replaced with a cost component in the objective as in (1); see for instance [21, p. 147]. We operate under these assumptions for the remainder of the paper.

If an advertiser bids so high on a term that it always achieves the top slot in every resulting auction, then it has obtained the maximum amount of clicks possible from that term. Letting X_i denote this maximum, the advertiser's consumption set is therefore constrained by

$$x_i \le X_i \qquad (i = 1, \ldots, n) \qquad (2)$$

The objective (1) together with the constraints (2) encode the advertiser's demand problem.

PROPOSITION 1. *Assume each price function p_i is differentiable. Then at an optimal solution $x^* > 0$ of the advertiser's demand problem we have*

$$\log v_i \ge \log p_i'(x_i^*) + \rho \log x_i^* + \rho\, C, \qquad (3)$$

for $i = 1, \ldots, n$, with equality *if $x_i^* < X_i$. Here C is a constant that is independent of the terms.*

The assumptions in the proposition imply that the objective is differentiable, and the result follows from straightforward inspection of the Karush-Kuhn-Tucker necessary conditions for optimality. The restriction $x^* > 0$ that the optimal choice of clicks be strictly positive (the inequality is understood component-wise) is without loss of generality for our purposes, because in our empirical analysis we will only consider terms for which an advertiser meets the reserve price.

Note that (3) is an equality for all terms except those where the advertiser reaches the maximum amount of clicks. It is fundamental to our approach as it relates the value parameter v_i for each term i to observables: the quantity of clicks demanded and the price derivative at that quantity. At $\rho = 0$ (additive utility), we recover the familiar condition that marginal value equals marginal price at an optimal choice of clicks in (3). If the capacity constraint (2) binds for a term—which is observable, as we simply check whether the bidder always reaches the top slot—then marginal price is only a lower bound on value.

We have only modeled advertiser utility at the ad group level. For higher levels, we assume that utility is additive across ad groups and campaigns, so that ad groups can be considered in isolation. We considered the possibility of modeling utility across the entire account as a *nested CES* function, with value and substitution parameters at all levels. However, the resulting formula (3) relating the value parameters to observables becomes nonlinear in the higher-level substitution parameters, which makes it unsuitable to incorporate into our later regressions. In the interest of simplicity, we leave this possible extension to future work.

We mention that our value estimation methodology is still relevant if an advertiser does not care for money[2], or more generally budgets bind, but our results need to be reinterpreted with care. With a binding budget constraint, $p_i(x_i^*)$ in (3) needs to be replaced with $\lambda p_i(x_i^*)$ where $\lambda \ge 0$ is a multiplier. As a result, the value estimates v_i are no longer denominated in the same units as money even for additive utility. Nevertheless, the estimates are enough to quantify the marginal rate of substitution between clicks on any two terms, even with λ unknown.

In the next Section 4.1 we elaborate our approach to estimating the two observables that enter into Proposition 1: the quantity of clicks demanded on a term given an advertiser's bid, and the price derivative at that demanded quantity. In Section 4.2 we perform some diagnostic checks and a sensitivity analysis of our empirical estimates.

4.1 Value Estimation

Our approach to modeling advertiser beliefs, as to the total clicks and cost they can expect at different bids, follows Athey and Nekipelov [4] conceptually but differs in the estimation details. Let $\hat{x}_i(b_i)$ denote the advertiser's expected total clicks when bidding b_i on term i, and similarly let $\hat{p}_i(b_i)$ denote the expected total cost. We stress that these are totals over the relevant time period (e.g., week, month), not CTR and CPC. Given how infrequent bid updates tend to be, we chose to model beliefs over a *month*.

We estimate beliefs by using log data of actual searches that occurred in the chosen month and the observed advertiser bids in each auction. We assume that the advertiser makes a decision to place a single bid for each term over the entire month. The advertiser's position on a keyword can still vary substantially across time because opponents change (due to targeting and matching) and quality scores are constantly updated.

For each realized auction instance, we run through the bids that the advertiser could have hypothetically placed

[2]This may occur with SEM firms, who charge a fee in proportion to the marketing budget; this can result in distorted incentives that lead firms to spend remnant budget on terms that run a loss. For instance, see the first example in [12], where the optimal portfolio has a term with negative ROI.

within some relevant range $[b_{min}, b_{max}]$[3], and simulate the counterfactual CTR and CPC it would have obtained at that bid. We can simulate changes in position because we have the opponent's bids and quality scores, and we obtain predicted CTR from the search engine's estimates of ad- and position-specific CTR effects. Aggregating CTRs and CPCs over all auction instances gives the expected total clicks and cost over the month for each bid setting. We should mention that accurately simulating the auction is quite challenging, as there are many added complexities like reserve prices, policies on placing ads in the north or east, etc.

Our approach to estimating expected clicks and costs results in step functions of the bids. By the rules of sponsored search auctions, each function is necessarily non-decreasing in bid. We define an inverse to the expected clicks function by

$$\hat{x}_i^{-1}(x_i) = \inf\{b_i : \hat{x}(b_i) \geq x_i\}.$$

The expected total cost as a function of clicks, which is one of our two key observables, is then reconstructed from $\hat{x}_i$ and $\hat{p}_i$ by

$$p_i(x_i) = \hat{p}_i(\hat{x}_i^{-1}(x_i)).$$

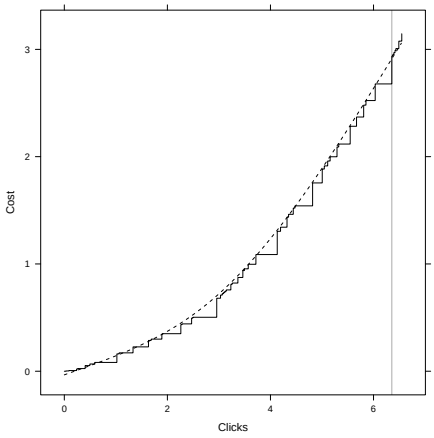
We refer to this function as an estimated *supply curve*. It inherits the properties of $\hat{x}_i$ and $\hat{p}_i$ in that it is a non-decreasing step function of clicks, and for high-volume terms it is close to continuous. Finally, we approximate p_i at the points where it changes value using cubic spline interpolation, which results in a curve with smooth derivatives [18, chap. 3.3]. Figure 2(a) exhibits an estimated supply curve for a term together with its differentiable approximation.

From the supply curve we can obtain the marginal cost, corresponding to our value estimates (for additive utility), as well as the average cost per click. Figure 2(b) exhibits the estimates derived from the supply curve. Given the rules of sponsored search auctions, we expect average CPC to increase with clicks. It is a basic result in microeconomics that increasing average cost implies marginal cost lies above average cost (e.g., [21]), which we see in the figure.

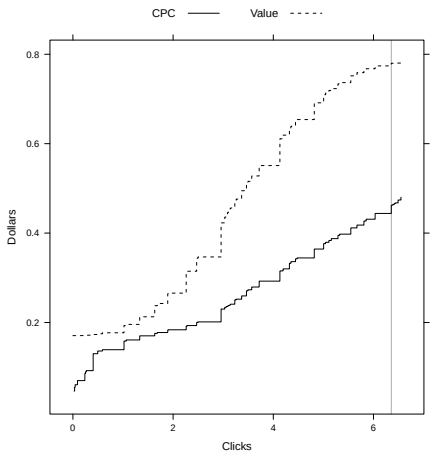
Our model of an advertiser's beliefs is close to ideal, as it is based on data from the same time period during which the advertiser bids, and draws on offer- and click-stream data as well as the search engine's own CTR estimates. In practice, advertiser beliefs are often based on daily aggregate reports of clicks and cost at best, and an advertiser needs to explore different bids to get just a partial sense of the supply curve. The estimation of supply curves from the advertiser's perspective is complicated by aggregation bias in the reports, which can arise because CTR is convex in bid and rank, so that average CTRs at average bids will overestimate true CTRs–see [1]. However, as noted by Varian [22], the informational requirements of bidding optimally are not as stringent as having an ideal estimate of the supply curve. Advertisers can also arrive at the right bid using a tatonnement process that equates value per click with marginal cost.

Our estimate of the selected quantity of clicks for the given time period is simply $x_i = \hat{x}_i(\bar{b}_i)$, where $\bar{b}_i$ is the advertiser's mean bid over offers on term i. Because our estimation approach assumes a fixed bid, we only consider terms where the relative standard deviation of the bid over the month is no more than 10%. Our sensitivity analysis in the next section indicates that value estimation is robust within this band. Note that our use of the average bid does not run into the aggregation bias just mentioned, because we are fully aware of the convex shape of the supply curve.

[3] b_{min} was set to \$0, and b_{max} was set to a bid that allowed the advertiser to achieve the top position in every auction in the month. The interval was discretized by \$0.01.



(a) Supply curve



(b) Estimated values

Figure 2: An instance of value estimation. Reference lines indicate the advertiser's chosen (expected) quantity of clicks over the month via its bid.

4.2 Diagnostic Checks

The simplest utility function within the CES family we consider is additive utility ($\rho = 0$). This structure has been implicit in all the equilibrium models of sponsored search to date since they all assume that advertisers can reason about bidding on different keywords in isolation [4, 10, 22]. We now examine whether the value estimates we obtained for high-volume terms using the approach just outlined respect some basic implications of additive utility.

Non-negative values. Perhaps the simplest check is that estimated values should be non-negative, otherwise it would be nonsensical to bid on the respective terms. Our cubic spline interpolation does not enforce monotonicity, so estimated price derivatives can in principle turn out nega-

tive. Monotone spline interpolation methods do exist, but it turns out that monotonicity was not a serious issue. The proportion of terms where a negative value was estimated, among all the eligible terms in an account, had a maximum of 0.3%, with a median and mean of 0% and 0.03% respectively over all accounts.

Overbids. Though our model of advertiser behavior is not game-theoretic, it is instructive to check our estimates from that perspective. An immediate insight from equilibrium analyses of sponsored search is that bidding more than one's value is a weakly dominated strategy [17], so we should expect this to be relatively rare. We refer to an instance where the bid exceeds the estimated value as an *overbid*. Overbids corresponding to bids that achieve the maximum amount of clicks on a term are not a concern, because recall that in this case our value estimates are only lower bounds on actual values. Figure 3 exhibits an account with one of the largest proportions of overbids. As in this instance, we typically find that the vast majority of overbids are accounted for by the fact that the bidder always reached the top slot. The proportion of terms in an account where bids exceeded exact value estimates, as opposed to lower bounds, ranged from 0% to 24%, with median and mean both at 12%, over all accounts.

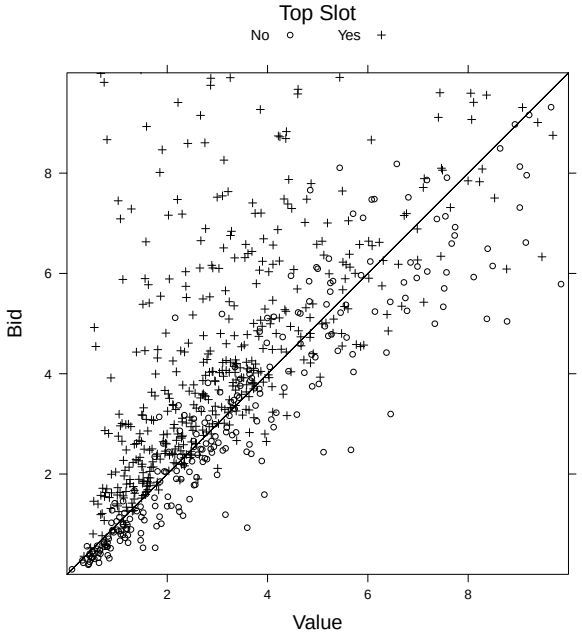
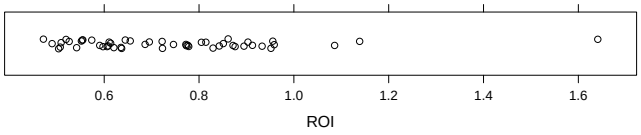


Figure 3: Bids against estimated values for a sample account with a large proportion of overbids. Ranges for value and bid have been truncated for clarity.

Individual rationality. Of greater concern are terms where the estimated value falls below the estimated CPC, leading to negative profit per click, a violation of individual rationality. The proportion of terms in an account where this occurred ranged from 0% to 5%, with median and mean both at 1%, over all accounts. Because this is much lower than the incidence of bids exceeding values, it suggests that in many cases the latter may occur simply because the advertiser has not found it costly to bid higher than necessary. Other reasons may include vindictive bidding, where an advertiser raises its bid to make clicks more costly for competitors [24], or poor/noisy/sparse value estimation by the advertiser.

The following plot shows the median ROI for all fifty accounts according to estimated values and CPCs. These numbers seem generally in line with other estimates in the literature. Varian [22] reports a median ROI of around 1.5 over eight bidders in a single auction snapshot, using his value estimation method, while Athey and Nekipelov [4, Table 8] report median ROIs of 0.6, 1.1, and 1.2 on the three search phrases they consider.



Sensitivity analysis. We performed a sensitivity analysis to examine how our value (price derivative) and click estimates behave under perturbations to the advertisers' bids. Robustness to bid perturbations is crucial because these observables are fundamental inputs to our hierarchical model, and it is too stringent to assume that advertisers are always bidding optimally on a term. For each term in an account we perturbed the bid uniformly at random between $\pm 20\%$, meaning that the expected absolute perturbation was of 10%, and recorded the resulting absolute perturbation in estimated value, click, and cost averaged over 50 trials. Our findings for all fifty accounts are summarized in Figure 4. Note that in this figure cost refers to the total cost of the clicks rather than CPC.

We see that the median value perturbation is no more than 10% for all but one account, indicating that perturbations in bid are not amplified in value estimates. There is some amplification in the resulting clicks and cost. The cubic spline interpolation implicitly penalizes the second derivative of the estimated supply curve (equivalently, the derivative of estimated value), which could in principle be a source of the robustness in value estimates, though in cases where the interpolation is excellent—see Figure 2(a)—it is an inherent property of the supply curve.

5. HIERARCHICAL MODEL

The hypothesis behind our predictive model is that advertiser value for clicks from a keyword can be decomposed according to value-bearing features. In this section we introduce a hierarchical linear model to determine the relationship between bid or value to some observable, measurable features of keywords. A hierarchical model is natural in this domain given the hierarchical structure of advertiser accounts and campaigns.

We first describe our model for predicting bids. We index the terms in a campaign by i, and the ad groups by j; we use the bracket notation $j[i]$ to denote the ad group that term i belongs to [13]. For each term i we denote its (column) vector of features by x_i. The term-level model takes the form

$$y_i = \alpha_{j[i]} + \beta' x_i + \epsilon_i \tag{4}$$

where $\epsilon_i \sim N(0, \sigma_y^2)$. For now $y_i = \log b_i$, the natural log of the bid on term i. We denote the vector of ad-group-level features by u_j for ad group j. The ad-group-level intercept

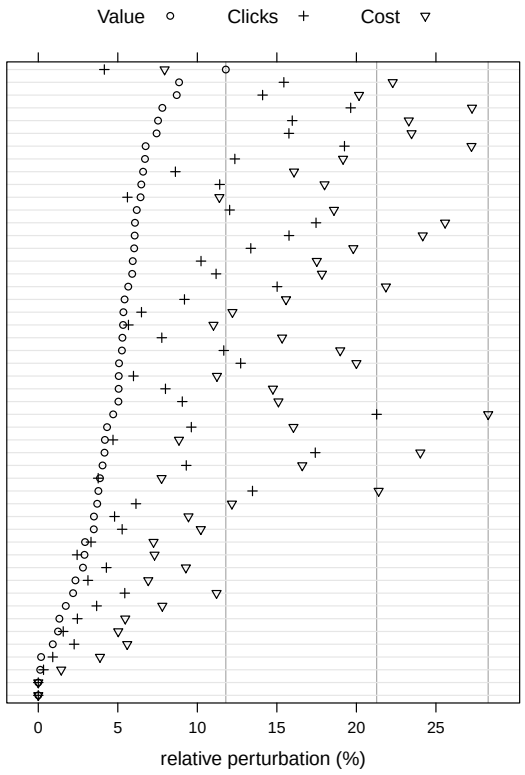


Figure 4: Median expected perturbation, over terms in an account, in estimated value, clicks and cost due to an expected 10% perturbation in bid.

itself follows the model

$$\alpha_j = \nu + \delta' u_j + \epsilon_j \tag{5}$$

where $\epsilon_j \sim N(0, \sigma_\alpha^2)$. We could extend the model further up the hierarchy, for instance by making ν a random effect among campaigns in an account; we refrained from this because of a dearth of meaningful predictors at the campaign level. Note that only the term-level intercepts are random effects; the predictor coefficients are fixed.

To develop a model for value rather than bid, we appeal to the relationship (3) and take $y_i = \log p_i'(x_i^*)$, introducing $\log x_i^*$ as an additional predictor on the right-hand side. The parameter ρ in (3) then corresponds to the coefficient on log clicks—we defer the interpretation of this coefficient to the next section—while the constant term in (3) gets incorporated into the constant term of the regression.

Recall that for bids that reach the maximum amount of clicks, the observables only provide a lower bound on value. If we were to assume additive utility and enforce $\rho = 0$ in the model, it could make sense to take our estimate of v_i to be $\max\{p_i'(x_i^*), b_i\}$, given the game-theoretic analyses that suggest bid should never exceed value. However, this could result in overestimates of value, which could prove costly if we were to later use them for bid suggestion. Performing the regression just on $p_i'(x_i^*)$ is the cautious choice, even though it can be an inexact estimate.

We now proceed to introduce our predictors for bid and value. The following predictors (or categories of predictors) were used at the term level. As indicated, some predictors are logged to account for skew in their distributions.

age five separate predictors indicating the percent of offers to users in their twenties, thirties, etc.

gender two separate predictors indicating the percent of offers to male and female users.[4]

income (log) mean income of users viewing the offer according to their zip code.

users percent of offers that were viewed by distinct users, according to browser cookie.

exact match percent of offers presented due to exact rather than advanced match.

user click propensity (log) mean of a proprietary metric quantifying, for each keyword the offer was matched to, the propensity of users who search on the keyword to click on ads.

query length mean word length of the user queries leading to the offer.

north state four separate predictors indicating the percent of offers shown on a page with one, two, three, or four ads at the top of the page.

competitors (log) mean number of competitors on auctions the offer was matched to.

The last two predictors merit further discussion. While the other predictors all capture some aspects of the population of users the term reaches, such as demographic profile or precision of user intent, the last two should have no bearing on conversions or profit per conversion. On the other hand, they should clearly bear on the bid since they speak to the level of competition on the keyword. We next turn to the predictors at the ad group level.

creatives number of creatives in the ad group.

linead length word length of the keyword bid on.

title length word length of the ad title.

copy length word length of the ad content.

These predictors all give some indication of the specificity of the advertisement, which may correlate with value.

In the next Section 5.1 we compare regressions on bid versus value, to better understand the advantage of backing out values from bids, and in Section 5.2 we report on the predictive performance of our hierarchical linear model.

5.1 Regression Results

As a bid represents an advertiser's reported willingness to pay per click on a term, one could be tempted to use bid as a proxy for value in fitting the model. However, it is well-known that sponsored search auctions are not truthful, and that advertisers have an incentive to shade their

[4]There is also a third category for unknown gender, which was omitted to avoid collinearity with the intercept. We also omitted 'age unknown' and '0-north' predictors for this same reason.

bids to increase profits [3, 17]. While bids naturally correlate with values, they also reflect the amount of competition on a keyword. Looking towards applications such as keyword suggestion, we would like to identify the keywords that would be the most valuable for an advertiser, not the most competitive.

To illustrate this point in more depth, we consider the regression on a single advertiser account as a case study. We stress that this case turned out particularly clean, and that there are no clear observations on regression features that systematically hold across all accounts—the most predictive features vary substantially across accounts, as one would expect, since the accounts do not necessarily belong to the same verticals.

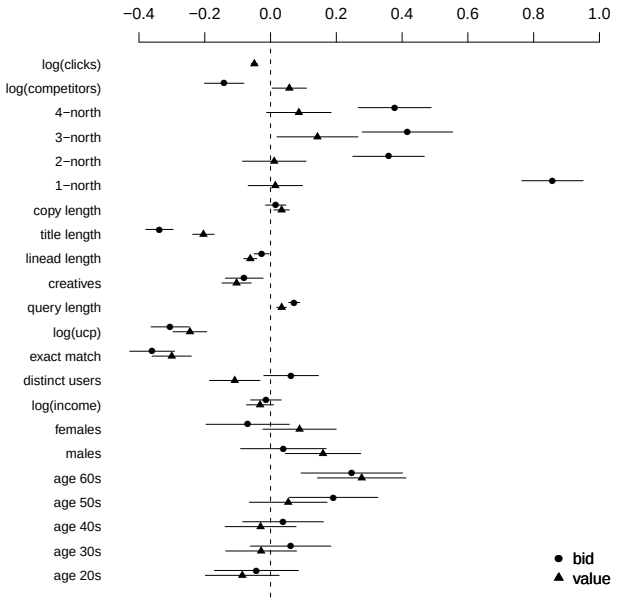
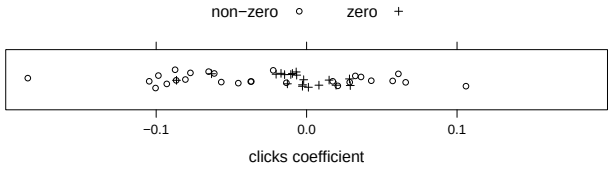


Figure 5: Regression coefficients for log(bid) and log(value). Intervals denote two standard errors.

The regression coefficients for our case study account are presented in Figure 5. We see here the desired effect of regressing on value rather than bid. The competitive features (number of competitors, ads in the north) lose in significance when moving from bid to value, while certain demographic or ad group features (e.g., males) gain in significance. To get a sense of the practical significance of these coefficients, we see that a 1% increase in offers to males with a corresponding 1% decrease in offers to females is associated with an 11% increase for the bid and an 8% increase for the value (though for the bid neither feature is significant at the 5% level). An increase of 1 to the average word length of user queries (the 'query length' feature) is associated with a 7% increase in bid and a 3% increase in value, while adding 1 word to the keyword actually bid on (the 'linead length' feature) is associated with a 3% decrease in bid and a 6% decrease in value. This exercise is illustrative—we do not expect any of these features to change with others held fixed.

Note that the logged clicks feature is present for the value but not the bid regression. Clicks are endogenous to bids and not available at prediction time, so are not appropriate as features. For value prediction the clicks coefficient does not come into play either, because we are in effect predicting the right-hand side of (3)—modulo the constant term, which does not affect relative value within an ad group—which includes logged clicks. This raises the question of the purpose and interpretation of the coefficient on clicks.

The original purpose of the clicks feature was to introduce ρ as an additional free parameter in the regression and see if it consistently clustered around a non-zero value over the accounts. (Note that the coefficient on logged clicks corresponds to $-\rho$.) If so, this could have provided suggestive evidence that some other member of the CES family is generally a better fit than additive utility. For instance, consistently positive estimated ρ coefficients may have provided evidence for the presence of complementarities among terms in ad groups. The following plot summarizes the estimated ρ coefficients over the accounts.



We find that in fact ρ is not consistently positive or negative, and that it does not stray too far from zero; with the exception of one account, it falls mainly within ± 0.1. Over the 50 accounts, 11 coefficients are positive, 18 are negative, and 21 are zero. Our conclusion is that additive utility is a good modeling choice, and that the coefficient on clicks can serve a different purpose than situating utility within the CES family. Because constant value per click cannot correlate with the number of clicks, by definition, a non-zero clicks coefficient indicates that there remain predictive features to be found (that correlate with clicks).

5.2 Prediction Performance

We now evaluate the performance of our hierarchical model in predicting values and bids. Combining our data set and the value estimates from Section 4.1, we have for each term an observed bid, an inferred value, a set of features for that term, and the term's corresponding ad group features. For each account, after training on some subset of the terms in that account, our goal is to predict values and bids for the remaining terms using only the corresponding term-level and ad-group-level features.

The features we use in the hierarchical model are the same as those described at the start of Section 5. Given the analysis of the ρ parameter in the preceding section, we focus on the case of additive utility where $\rho = 0$. The values and bids used in the hierarchical model are logged and mean-centered around the account's mean value and bid.

For each account, we evaluate value and bid prediction accuracy by performing k-fold cross-validation on the terms in that account, where terms are randomly split into each of the k folds. Note that two distinct types of prediction arise in the cross-validation: when all terms within an ad group are in the test set, there is no training data available for that ad group. This is analogous to the real-world problem of predicting the advertiser's value for a term in a newly created ad group that contains creatives but not yet terms; this type of prediction could therefore apply to initial keyword recommendation for a new ad group. When an ad group's terms are split between the training and testing sets, we face

a prediction problem that could come about as an advertiser seeks to evolve its campaign over time by adding terms to already existing ad groups. The median percentage of terms that were predicted as new ad groups for each account was 18%, 9%, and 8% for 2-, 5-, and 10-fold cross-validation, respectively.

We evaluate the performance of our hierarchical model using the estimated values from Section 4.1 and the observed term bids as ground truth. Performance is with respect to the following baselines:

- The `Unpooled` baseline is a classical linear regression model using the same predictors as the hierarchical model, except that a separate regression model is estimated for each ad group. This baseline does not handle predictions for new ad groups since it does not have any observed terms in the ad group to fit its model.
- The `Grouping` baseline predicts the value or bid of a given term to be the advertiser's average value or bid for other terms in the same ad group. Use of this baseline is motivated by the observation from Section 3 that variation of bids within an ad group is relatively small. If the training data contains no terms in the same ad group, the baseline reverts to the average value or bid of terms in the campaign.
- The `Opponent` baseline predicts the bid of a given term to be the average bid of opponents that appeared in the same auctions as that term. Note that this baseline only predicts bids and not values, but could be extended to values if we had value estimates for more than the 50 accounts from our data set.

The results of k-fold cross-validation averaged across all 50 accounts are shown in Table 2, expressed in percent error. Results are shown for 2, 5, and 10 folds, and performance is further broken down by showing prediction error for terms in both existing and new ad groups. The hierarchical model has a smaller amount of error than any of the baselines in 17 of 18 instances. This difference was statistically significant using a two-sided paired-means t-test (confidence level .95) when predicting values on 5 or 10 folds, but not on 2 folds or when predicting bids.

All models (except the `Opponent` baseline) improve their prediction accuracy as the number of folds increases, which is to be expected, since more folds result in more training data for making predictions, and fewer folds meant more predicting on new ad groups. For both value and bid prediction, the hierarchical model improves in performance at a faster rate than the `Grouping` and `Opponent` baselines as the number of folds increases. The gain in performance of the hierarchical model over the `Grouping` baseline tends to come from its ability to more accurately predict values and bids on terms in new ad groups.

Disaggregating these results, Figure 6 shows individual account performance of the hierarchical model compared to the `Grouping` baseline. Error for the hierarchical model is greatest on small accounts, suggesting that it did not have enough training data in these situations. Creating a hierarchical model across accounts or otherwise pooling across accounts would provide more training data in these situations, but this would likely not help unless the model was trained using data from similar accounts, which itself presents an additional challenge.

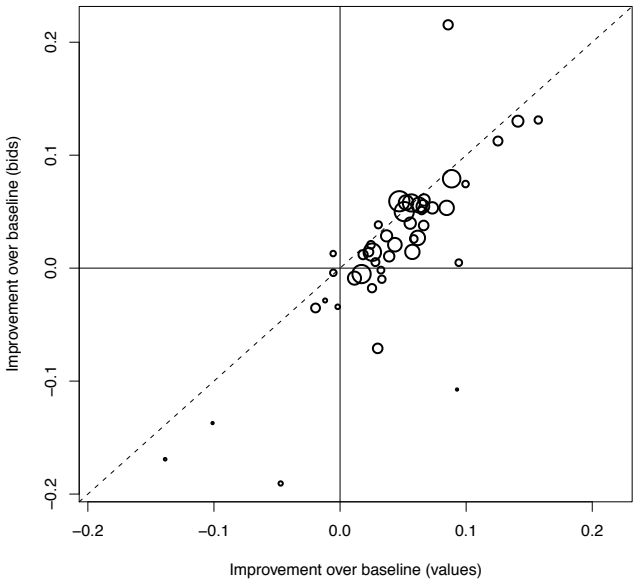


Figure 6: Hierarchical model performance against the `Grouping` baseline when predicting both bid and value using 10-fold cross-validation. Point size is proportional to log account size in terms.

6. CONCLUSIONS

In this paper we proposed a two-step approach to predicting values in sponsored search: we first estimate values on high-volume keywords based on advertiser bids, then fit a hierarchical model on top of the estimates using keyword features at several levels of the advertiser campaigns. In the process of developing our model we found evidence that advertiser utility is indeed additive across terms: bids rarely exceed estimated values, values exceed costs per click, and our hierarchical model achieves the best fit for utility that is close to additive. In terms of predictive quality, our hierarchical model overall outperformed the baseline by ten percent for value inference. We saw that the improvement was even more pronounced for large accounts and new ad groups, making the model particularly useful for augmenting campaigns, one of our principal use cases.

We see several avenues for improvement and future work. The approach taken in this paper assumes that advertiser beliefs are ideal: supply curves are estimated from the finest-grained auction and click stream data, using data from the same month that bids are placed. It would be an informative exercise to perform supply curve estimation using only the aggregate, partial data available to advertisers, to better understand the uncertainty they face when setting bids. Another line of research, to better understand when advertisers are bidding optimally, would be to develop ways to distinguish between learning behavior and actual structural breaks in preferences when advertisers update their bids.

There is also room for improvement in the prediction performance of our hierarchical model. We chose this kind of model with a view towards interpretation as well as prediction, which can be important if advertisers demand explanations for keyword or bid suggestions. We believe good improvements could be obtained using machine learning algorithms specialized for prediction (e.g, boosting [20]) if that

Output	Model	2 Folds			5 Folds			10 Folds		
		Seen	New	Total	Seen	New	Total	Seen	New	Total
Value	`Hierarchical`	0.51	0.56	0.51	0.47	0.54	0.48	0.47	0.55	0.47
	`Grouping`	0.53	0.61	0.53	0.52	0.59	0.52	0.51	0.62	0.52
	`Unpooled`	0.58			0.51			0.49		
Bid	`Hierarchical`	0.49	0.57	0.50	0.46	0.57	0.47	0.45	0.56	0.46
	`Grouping`	0.48	0.61	0.50	0.48	0.57	0.48	0.47	0.61	0.48
	`Opponent`	0.91	0.94	0.91	0.91	0.99	0.91	0.91	0.98	0.91
	`Unpooled`	0.56			0.49			0.47		

Table 2: Percent error of different models using k-fold cross-validation. The hierarchical model is the most accurate predictor in all cases except for 2-fold bid prediction on terms whose ad group is in the training set.

were the sole concern. We also see the need to move beyond ad-hoc feature selection. To this end, we intend to apply techniques such as topic models [5] to uncover conceptual and semantic regularities among campaign terms.

Acknowledgments

We thank Amy Greenwald and David Pennock for initiating this project, and Eliot Li for bringing the collaborators together. We received valuable comments and suggestions from Amy Greenwald, Patrick Jordan, Ashvin Kannan, Prabhakhar Krishnamurthy, Eren Manavoglu, David Pennock, and Michael Schwarz.

7. REFERENCES

[1] V. Abhishek, K. Hosanagar, and P. Fader. On aggregation bias in sponsored search data: Existence and implications. *SSRN eLibrary*, 2009.

[2] A. Agarwal, K. Hosanagar, and M. D. Smith. Location, location, location: An analysis of profitability and position in online advertising markets. *SSRN eLibrary*, 2008.

[3] G. Aggarwal, A. Goel, and R. Motwani. Truthful auctions for pricing search keywords. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 1–7, 2006.

[4] S. Athey and D. Nekipelov. A structural model of sponsored search advertising auctions. Technical report, Microsoft Research, May 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. In *Journal of Machine Learning Research*, volume 3, pages 993–1022, March 2003.

[6] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th International World Wide Web Conference*, pages 511–520, Madrid, Spain, 2009.

[7] A. Broder, E. Gabrilovich, V. Josifovski, G. Mavromatis, and A. Smola. Bid generation for advanced match in sponsored search. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 515–524, 2011.

[8] N. Brooks. The Atlas rank report: How search engine rank impacts conversions. Technical report, Atlas Institute, 2004.

[9] Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *Proceedings of the International Conference on Web Search and Web Data Mining*, pages 251–260, 2008.

[10] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the Generalized Second-Price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1), March 2007.

[11] Efficient Frontier, Inc. The algorithmic solution for search marketing optimization. White paper, November 2007.

[12] Efficient Frontier, Inc. Algorithms and optimization. White paper, December 2008.

[13] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[14] A. Ghose and S. Yang. An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55:1605–1622, October 2009.

[15] B. J. Jansen and K. Sobel. Investigating the brand effect in search engine marketing. Working paper, 2010.

[16] B. J. Jansen and L. Solomon. Gender demographic targeting in sponsored search. Working paper, 2010.

[17] S. Lahaie. An analysis of alternative slot auction designs for sponsored search. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 218–227, 2006.

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.

[19] O. J. Rutz and R. E. Bucklin. A model of individual keyword performance in paid search advertising. *SSRN eLibrary*, 2007.

[20] R. E. Schapire. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, pages 1401–1406, 1999.

[21] H. R. Varian. *Microeconomic Analysis*. Norton, 1992.

[22] H. R. Varian. Position auctions. *International Journal of Industrial Organization*, 25:1163–1178, 27.

[23] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International World Wide Web Conference*, pages 261–270, Madrid, Spain, 2009.

[24] Y. Zhou and R. Lukose. Vindictive bidding in keyword auctions. In *Proceedings of the 9th International Conference on Electronic Commerce*, pages 141–146, 2007.