Adaptive Weighing Designs
for Keyword Value Computation

John W. Byers [*]
Computer Science Dept.
Boston University &
Adverplex Inc.
byers@cs.bu.edu

Michael Mitzenmacher [†]
School of Eng. & Appl. Sci.
Harvard University
Cambridge, MA 02138
michaelm@harvard.edu

Georgios Zervas [‡]
Computer Science Dept.
Boston University &
Adverplex Inc.
zg@bu.edu

ABSTRACT

Attributing a dollar value to a keyword is an essential part of running any profitable search engine advertising campaign. When an advertiser has complete control over the interaction with and monetization of each user arriving on a given keyword, the value of that term can be accurately tracked. However, in many instances, the advertiser may monetize arrivals indirectly through one or more third parties. In such cases, it is typical for the third party to provide only coarse-grained reporting: rather than report each monetization event, users are aggregated into larger channels and the third party reports aggregate information such as total daily revenue for each channel. Examples of third parties that use channels include Amazon and Google AdSense.

In such scenarios, the number of channels is generally much smaller than the number of keywords whose value per click (VPC) we wish to learn. However, the advertiser has flexibility as to how to assign keywords to channels over time. We introduce the channelization problem: how do we adaptively assign keywords to channels over the course of multiple days to quickly obtain accurate VPC estimates of all keywords? We relate this problem to classical results in weighing design, devise new adaptive algorithms for this problem, and quantify the performance of these algorithms experimentally. Our results demonstrate that adaptive weighing designs that exploit statistics of term frequency, variability in VPCs across keywords, and flexible channel assignments over time provide the best estimators of keyword VPCs.

Categories and Subject Descriptors

G.1.6 [**Mathematics of Computing**]: Least squares methods

General Terms

Algorithms, Design, Economics, Experimentation

Keywords

Least squares, weighing designs, design of experiments, regression

1. INTRODUCTION

With the advent of web advertising, a profitable new business model has emerged which is based on purchasing traffic through online keyword auctions. The advertiser hopes that in expectation, the arriving users he has paid for will monetize (i.e. take an action of monetary value once they arrive at his website), providing revenue in excess of cost. For example, monetization could come in the form of a purchase, the filling out of a saleable lead form, or the clicking of a revenue-generating ad on the advertiser's website. Any such revenue-producing action is commonly referred to as a *conversion.* With keyword auctions, when the user *clicks* on an ad, the advertiser is charged. The basic mechanisms of keyword auctions have been the subject of a number of works; for more background, see for example [1, 2, 5, 7, 10, 14].

Our problem is motivated by the real-world scenario in which the product or ad inventory is indirectly supplied by a third-party. Consider this funnel: a user performs a query on a search engine and clicks on an ad for Advertiser A. The ad directs the user to the website of Advertiser A, who is billed for the click. The advertiser in turn is affiliated with a third-party supplier such as Amazon.com[1] who provides relevant product inventory. If the user eventually purchases something, the order is – transparently to the user – fulfilled by Amazon.com and the proceeds are split between the advertiser and the supplier.

Another common scenario, known in the industry as ad-arbitrage, involves Advertiser B showing a new set of ads to incoming visitors. When the visitor clicks on an ad at Advertiser B's website, where ads are again usually supplied

---

[1] Information on Amazon's affiliate program can currently be found at `https://affiliate-program.amazon.com/`. Google also has an affiliate program, with information at `http://www.google.com/ads/affiliatenetwork/`.

by a third-party such as the Google and Yahoo Publisher Networks[2], the advertiser receives some compensation. For more background on advertiser networks, see for example [4, 12]. While pure arbitrage of this sort often has negative connotations, there are many examples of sites that offer ads alongside additional beneficial information to the user and monetize traffic in this way.

In both of these cases it is essential for the profit-maximizing advertiser to determine how many visitors arrived on a given keyword, how often they converted, and what the aggregate value of conversions on each keyword was. The advertiser needs this information to calibrate keyword bids to run a profitable campaign.

Unfortunately, third-party associates are often unmotivated (or unable) to provide such complete data. In an extreme case the third-party may only supply the advertiser with an aggregate daily report listing the number of arrivals, the number of conversions, and their total value, with no breakdown by keyword. More commonly, the advertiser is given some flexibility: each visitor can be assigned to one of a limited number of *channels*. Then, the third party supplies the advertiser with a revenue report broken down by channel, typically on a daily basis. The motivation for providing channels is to allow the advertiser to measure the monetization potential of different pages on his site. Consider for example the case where the website is broken down into pages by product. Each page could be associated with a channel allowing the advertiser to know how well it performs. Channels can also be useful for multivariate testing. The website owner may wish to try various designs for his website and measure the performance of each to eventually select the best. This can be achieved by associating each design with a channel.

The number of channels available to an advertiser is commonly in the range of tens to hundreds. This is usually enough to support the motivations above, but not enough to measure the VPC of each keyword in a straightforward manner by mapping each keyword to its own channel on a daily basis, as the number of keywords in even a moderately sized ad campaign can dwarf the number of available channels.

Still, given the availability of channels, it is natural to ask how they can be used to efficiently obtain keyword-level estimates of VPC. An obvious solution is to use the channels over multiple days, assigning (or "weighing") one keyword per channel per day, until we get a sufficient number of measurements per keyword to make a suitable estimate of its VPC (or "weight").

We demonstrate that one can do much better than this naïve approach. A starting point can be found in classical results in experimental design, notably that initiated by Hotelling [9], that demonstrate the power of conducting carefully constructed weighings of groups of objects, or *weighing designs*, leading to systems of equations that can be solved for lower-variance estimates of individual object weights. Our work extends this to the notion of an *adaptive* design, where the results from previous weighings inform the subsequent weighing (channel assignment). We demonstrate that the power of adaptivity can be most effectively

<hr>

[2]Information on Yahoo's publisher network can currently be found at `http://publisher.yahoo.com/`, and information on Google's Adsense network can be found at `https://www.google.com/adsense/login/en_US/`.

harnessed in improving overall estimates when used in conjunction with keywords whose values and search frequencies are themselves highly variable, as is the case in the search advertising context.

2. PROBLEM STATEMENT, BACKGROUND, AND INTUITION

2.1 Definitions and Notation

We associate each keyword with two key attributes for the advertiser, a value per click (VPC), and a rate at which clicks arrive on this keyword. We model each of these attributes by a distribution: V_i is the distribution of keyword i's daily VPC, while C_i is the distribution of the number of daily received clicks on keyword i (where "day" should be thought to represent any appropriate time period). On day t keyword i receives $c_{it} = C_i(t)$ clicks with a VPC of $v_{it} = V_i(t)$, where v_{it} is distributed as V_i and c_{it} is distributed as C_i. In this work we generally assume the distributions are stationary and the realizations of the random variables are independent across days and among all v_{it} and c_{it} values, although one can imagine more complex distribution models.

Generally we take $V_i \sim N(\mu_i, \sigma_i^2)$ and $C_i \sim N(\nu_i, \tau_i^2)$; that is, the distributions are taken to be normal. Again one could consider other models; we offer some justification below. Here the μ_i, ν_i, σ_i, and τ_i can themselves be random variables, taken from some appropriate distribution.

In our examples we take μ_i and ν_i to be drawn from a heavy-tailed distribution. The motivation for using skewed distributions of this sort in advertising campaigns is that there are often a small number of keywords responsible for most of the traffic, and a small number of keywords that have quite large VPC. (Moreover, it is certainly not always the case that these are the same words!)

We assume that we have a set of n keywords whose VPCs we wish to estimate over d days using h available channels. We shall use the term *measurement* to denote the total value of a specific channel on a specific day. Therefore, after d days we will have conducted $m = dh$ measurements. If $j = th + k$ we define r_j to be the value of the measurement corresponding to channel k on day t. Furthermore, let H_j be the set of keywords assigned to channel k on day t. Recall that the assignment of keywords to channels can vary daily. It follows that $r_j = \sum_i \mathbf{1}_{H_j(i)} c_{it} v_{it}$. Here $\mathbf{1}_{H_j(i)}$ is the indicator function whose value is one if keyword i is in measurement j and zero otherwise. Finally, let $\hat{v}_{it}$ be our estimate of μ_i on day t. Our objective is to come up with algorithms assigning keywords to channels over time to minimize the error in the estimates.

We emphasize that in practice we may have the added restriction that each keyword must appear in exactly one channel per day, so the keywords are partitioned over channels each day. The reason behind this is that partners attribute revenue-producing events to channels, and thus unassigned keywords are incapable of producing revenue. We assume that the advertiser wishes to maximize global total revenue over all keywords. We enforce this restriction for our proposed regression-based algorithms, but do not enforce this restriction on other methods, as it allows for stronger comparisons. (Note that we could always keep one spare channel and assign any otherwise unassigned keywords on that day to that channel; hence the loss in allowing this assump-

tion is small.) Similarly, in some settings it could be possible to randomly split the traffic associated with a keyword over multiple channels, or to channelize traffic on attributes other than the keyword; we do not explore these possibilities in this paper, except again for some simple strategies we use as comparison points to our regression-based strategies. These variations in the model seem potentially interesting for future work.

We also require an objective function in order to evaluate various possible algorithms. The advertiser will use the VPC estimates he has obtained to calibrate his bids. Without loss of generality we will assume that he wishes to run his campaign at a 0% profit margin so ideally his cost per click for each keyword should be equal the keyword's true VPC. Moreover, we assume the advertiser wishes to maximize traffic at this break-even operating point. Consider an advertiser bidding on two keywords, K_1 and K_2. On a given day, assume that the advertiser has overestimated the true VPC for K_1 and underestimated the true VPC for K_2. The following day his bid on K_1 will be too high and he will lose money in expectation. Conversely, on K_2, the advertiser is underbidding, and therefore could and should bid up to the true VPC of the keyword. Bidding up will allow the advertiser to acquire additional clicks at break-even or better. So we view underbidding as incurring an opportunity cost, overbidding as incurring a monetary cost, and both of these situations as undesirable.

An error metric which captures the adverse bottom-line effects of both over- and under-estimated VPCs is the root mean square error (RMSE) of the click-weighted values. The RMSE metric is commonly used in similar weighing problems [3, 8, 13], but our ideas could also be applied to other error metrics as well. We emphasize that here we require a *weighted* RMSE, to take into account the effect of the number of clicks on the value proposition to the advertiser. For our purposes, we choose the following form for the weighted version of the RMSE, which generalizes the standard unweighted variation:

DEFINITION 1 (RMSE). *The* root mean square error *(RMSE) of a set of n weighted estimates $v_1, \ldots v_n$ with respect to the ground truth $\mu_1, \ldots, \mu_n$ and with weights $w_1, \ldots, w_n$ is defined to be $\sqrt{1/n \sum_i w_i^2 (\mu_i - v_i)^2}$*

Using our notation, the relevant RMSE relates the estimated values of keywords at day t to the actual underlying values, weighted by the true click frequency of the keyword:

$$e_t = \sqrt{1/n \sum_i \nu_i^2 (\mu_i - \hat{v}_{it})^2}. \tag{1}$$

This quantity is in units of dollars, and in our context, offers a suitable proxy for optimizing bid values. Therefore, in the remainder of the paper, we focus on strategies to achieve the lowest click-weighted RMSE.

2.2 Related Work: Design of Experiments

Before continuing, it is worthwhile to put our problem in a larger context. Our problem is similar to others found in the statistical subarea of the design of experiments, and in particular the class of weighing problems, which have a long and rich history. (For extensive background, see for example [3, 13]). Our problem is most closely related to what is known as the spring balance problem, where foundational work was done by Yates [15], Hotelling [9], and Mood

[11]. In the standard version of this problem, a number of items are to be weighed using a scale, where each weighing will have some associated error according to a fixed distribution over all weighings, and one wants to estimate the weight of each item. The key insight is that by grouping measurements together and solving the corresponding set of equations to estimate each variable, one can reduce the variance in the estimates over the simple technique of weighing each item separately. Readers of a more combinatorial bent will naturally associate these problems with the theory of block designs, and in particular balanced incomplete block designs (introduced by Yates [15]; see Chapter 1 of [13]). It is interesting to note that another class of problems, called chemical balance problems, which allow two sets of items to be weighed and their difference found, are closely connected to Hadamard matrices [3, 13, 11].

Despite this rich history of work connected to our problem, our formulation (and our approach) appears significantly different than previous formulations. Specifically, our problem has the following characteristics:

- Our measurements necessarily conflate two different random variables: the clicks over a time period and the VPC over that time period.

- In our setting, there is significant skew among the random variables, both among clicks and VPCs.

- The errors in our measurements are associated with the performance of each keyword, not with the overall measurement itself.

- We seek to minimize the error in total value, so the VPC of individual keywords in necessarily weighted by clicks.

- The scale of our problems appears larger than previous conventional problems.

- Most importantly, we have the power to adaptively change the choice of measurements on a regular basis; we do not have to set up a single design at the beginning of the process.

Because of these novel features, we view our work both as opening interesting mathematical directions for work in the design of experiments, while also connecting some ideas from this rich field to computational advertising.

2.3 Simple Examples and Intuition

Before considering our specific problem setting, it is worth gaining some intuition by looking at some specific examples, starting with the historical literature. We suggest the following variation, described for example in [3]: suppose we have a spring scale, with seven objects a, b, c, d, e, f, g to weigh. The scale is such that the variance in the estimate is σ^2 (regardless of the mean). One could weigh each object individually; using seven measurements, this gives a variance of σ^2 in the estimate for each item. Alternatively, one could use seven weighings, of the following form: $w_1 = a+c+e+g$, $w_2 = b+c+f+g$, $w_3 = d+e+f+g$, $w_4 = a+b+e+f$, $w_5 = b+c+d+e$, $w_6 = a+c+d+f$, and $w_7 = a+b+d+g$. This is actually an optimal system of weighings [11], and corresponds to a symmetric balanced incomplete block design, where each pair of elements are together in exactly two

weighings. The natural estimate $\hat{a}$ for the weight of item a is then

$$\hat{a} = (w_1 - w_2 - w_3 + w_4 - w_5 + w_6 - w_7)/4;$$

that is, we add the weighings that include item a and subtract the others. The estimates are similar for the other items. If the measurements are independent with the same variance, then the estimate for each item has variance $\frac{7\sigma^2}{16}$ instead of σ^2 (because the variance of the sum is the sum of the variances, divided by the scale factor 4^2). By combining items in a single measurement, the variance in the estimators is substantially reduced.

Notice that if the variance in the measurement depended on the number of items in the measurement, as is the case in our problem, there would be no gain in this approach. (Indeed, in this example, the estimates would be worse.) However, we still expect that proper combining of measurements should lead to improved estimates in our setting. The reason here is the skew in the values and the clicks. Relative errors matter much more for keywords with high VPC, and absolute errors matter much more for keywords with high click volumes. Because of this, the payoff for having a larger number of measurements involving such keywords, even at the expense of additional noise from mixing them with keywords that contribute less significantly to the error, is likely to pay off over performing independent measurements for each keyword. Moreover, this intuition suggests that if we can adaptively mix keywords which are negatively impacting our error the most with keywords that are accurately estimated by our measurements, we should be able to improve overall estimates much faster.

The question that remains is under what settings we realize the benefits of combining measurements of keywords within a channel and using regression methods. The rest of the paper will focus on answering this question, including developing novel adaptive weighing methods that take into account previous measurements in deciding what keyword split to use on any given day.

2.4 Additional Modeling Decisions

In testing our approach, we had to make some key decisions regarding the model. We believe the problem of developing standard, appropriate models for keyword performance under general settings is interesting and worthy of further attention, beyond the scope of this work. We explain some of our choices here.

We chose for our test to use the normal distribution for the VPCs per day. To justify this choice, consider a specific keyword i that gets a large number of clicks over a day. The value corresponding to those clicks will depend on the actual number of conversion events that produce revenue.[3] If the conversion rate for that keyword is treated as a fixed probability p, then the number of conversions per day will be a

---

[3]In practice, it is often possible for the advertiser to segment the inbound users into two categories: those users who definitely did not monetize via a third-party, having never clicked on an outbound link; and those who did click out and may have monetized. In such a setting, all inbound clicks known not to have monetized can be assigned a value of zero, and thus weighings can focus on valuing the remaining clicks. For simplicity, we ignore this consideration; it would not change our resulting analyses, assuming that daily VPCs could be still be modeled as normal random variables as we do in our experiments.

binomial random variable $B(c_{it}, p)$. When $c_{it}p$ is sufficiently large, the number of conversions is approximately normal; if the value per conversion is fixed, then the value per click will be approximately normal as well. More generally, even if the value per conversion varies (say, according to a fixed probability distribution over values), the value per click will be approximately normal as well.

We acknowledge that in many regimes the value may be better represented by other distributions; for example, when p is small (on the order of $1/c_{it}$), then the number of conversions may be better modeled by a Poisson distribution. Additional experiments we have performed suggest that at a high level our results do not depend on the assumption of the normal distribution, but our suggested algorithms would have to be further tested for specific cases.

A further design decision to consider regards the distribution of clicks by day. While in our experiments, we draw from a real-valued distribution, technically, the number of clicks per keyword should be non-negative integers. At first blush, using a real-valued distribution may seem to make little difference, but in fact, keywords that receive a fractional click, as opposed to no clicks, are highly significant. For example, if we employ a strategy that gives a keyword its own dedicated channel, and there are 0 clicks for that keyword on a day, we have learned nothing about the value for that keyword. If we instead simulate a fraction of a click as simply a weight within the formulas, we do obtain information. Therefore, we have adopted the following randomized rounding approach to ensure that x, the number of clicks for a given keyword, is integral: for negative x (rare), round x up to zero. For positive x, round x up to $\lceil x \rceil$ with probability $x - \lfloor x \rfloor$, and down to $\lfloor x \rfloor$ otherwise.

3. A VPC ESTIMATION ALGORITHM

So far we have established the desirable properties of any VPC learning algorithm but we have yet to describe how this estimation process works. Before we do so, we review the information at hand. At the conclusion of each day t the advertiser has the following information:

- The number of paid clicks c_{it} on each keyword i that the advertiser actually received on day t.

- For every channel k, a third-party report of its daily total revenue r_j (recall that $j = th + k$.)

- A record of H_j, the keywords assigned daily to each channel.

Depending on whether we assign just one or multiple keywords per measurement we can use one of the following two methods to compute keyword VPCs.

3.1 One Keyword per Measurement: Weighted Averages

If we were to assign a single keyword per measurement we could arrive at a VPC estimate for that keyword by simply dividing the total value earned by a keyword with the total number of clicks it has received so far. Since each keyword appears in each measurement by itself we know for a fact that any value associated with the corresponding channel can be attributed to that specific keyword. Using our

	Weighted Averages	**Linear Regression**
Oblivious	Round-Robin	Regular-p Regular-p-FGLS
Adaptive	Adaptive-1	Adaptive-OLS Adaptive-FGLS

Table 1: A taxonomy of the strategies we employ to solve the channelization problem.

notation the VPC for keyword i on day t' would then be:

$$\hat{v}_{it} = \frac{\sum_j \mathbf{1}_{H_j(i)} r_j}{\sum_{t=1}^{t'} c_{it}}$$

Observe that this is a more efficient estimator than taking the average of daily computed VPCs even though both will eventually converge to the same answer.

3.2 Multiple Keywords per Measurement: Linear Regression

By assigning multiple keywords per channel and varying the assignments over time, the advertiser will eventually end up with an overdetermined system of equations that can be solved using a linear least squares method. The model that the linear regression has to estimate is the following:

$$\mathbf{r} = \mathbf{C}\mathbf{v} + \epsilon$$

The dependent variables of the model constitute the vector $\mathbf{r}$, an $(m \times 1)$ vector of measurements. The independent variables are contained in $\mathbf{C}$, an $(m \times n)$ matrix of clicks. The vector $\mathbf{v}$ is an $(n \times 1)$ vector of VPC estimators and ϵ is a vector of errors. Once enough measurements have been collected so that the system of equations is overdetermined, i.e., the rank of $\mathbf{C}$ is greater than n, we can apply a least-squares fit to obtain the unique solution which minimizes the sum of squared residuals $\sum_j (r_j - f(c_t, \hat{v}))^2$ where $f(c_t, \hat{v}) = \sum_i \hat{v}_{it} c_{it}$.

The question which naturally arises in this context is: to what extent can we exploit our freedom to assign keywords to channels to minimize RMSE? The following subsections develop the strategies we employ in detail.

3.3 Strategies

There are two main classes of strategies for the assignment problem that we consider. *Oblivious* strategies are those in which we compute a fixed, static assignment schedule *a priori* and execute it daily, gathering measurements along the way. On the other hand, *adaptive* strategies dynamically compute an assignment schedule for time period t based on the previous schedule and on the results of measurements up through time $t-1$. Furthermore, depending on whether they assign single or multiple keywords per channel the strategies are broken down in those that use weighted averages to obtain VPC estimates and those that employ regression. Table 1 provides a taxonomy. The specifics of the strategies we consider are detailed below.

3.3.1 Basic oblivious strategies

Round-Robin: we simply run through the keywords in round-robin order assigning one keyword per channel per day.

Regular-p: an oblivious strategy that attempts to take p measurements per keyword per day. For example if $p = 2$

then each keyword appears in exactly two channels per day. The channels to which the keyword is assigned are selected uniformly at random (without replacement) from the set of available channels. Note that when $p > 1$ we violate the restriction that each keyword appears in only one channel per day; we therefore employ such schemes only as comparison points. We also point out that instead of using randomization, we could have instead used deterministic designs from the theory of weighing designs [13]. However, such designs are somewhat complicated, and patterns for optimal deterministic designs are not known for all general parameters. Employing randomization appears to give up very little while simplifying the program for the strategy greatly.

3.3.2 Basic adaptive strategies

Our adaptive strategies are built on the observation that certain keywords are much more important to measure (and isolate) than others. For example, when a small number of keywords dominate the click distribution, we devote individual channels to those keywords to accurately measure their values. To motivate the first such strategy, we begin with a definition.

DEFINITION 2 (PERCEIVED ERROR). *The perceived error of a keyword is the width of the a'th confidence interval around the mean of its VPC measurements, as computed by $t(a, N-1) * s/\sqrt{N}$, where $t(a, N-1)$ is the critical value of the T distribution for $N-1$ degrees of freedom, a is the confidence (e.g., $a = 0.95$ for the 95% confidence interval), s is the sample VPC standard deviation and $N \geq 2$ is the number of measurements for the keyword.*

This leads us to define an adaptive algorithm similar to the round-robin algorithm, except with priorities based on the perceived error.

Adaptive-1: an adaptive strategy that first orders all keywords by $\text{Err}(i, t) * \text{clicks}(i, t)$, where $\text{Err}(i, t)$ is the computed perceived error for keyword i through time $t-1$ at confidence 0.95, and $\text{clicks}(i, t)$ is the average daily number of clicks for the keyword up through time $t-1$. The h channels then take isolated measurements of the h top-ranked keywords.

Our remaining and most advanced adaptive strategies draw from weighing designs, whereby we pack all keywords into channels, but do so adaptively and do so in order to maximize the amount of new click-weighted information that we can infer about keyword values from a given set of measurements. The two remaining approaches both start by packing keywords into channels while keeping the expected total value of each of the h measurements as even as possible. That is, we sort the keywords according to the product $\hat{v}_{it} * \text{clicks}(i, t)$, where $\hat{v}_{it}$ is the current estimate of the VPC, and again $\text{clicks}(i, t)$ is the average daily number of clicks for the keyword. We greedily pack the keywords in decreasing order into h bins, representing the channels for use the next day. We note that decreasing order is generally known to give better performance for greedy packing algorithms [6]. Experiments with various greedy packing strategies (least full, first fit, and best fit) did not yield significant differences, so we chose the simplest and best-performing: least full. Least full also has the intuitively beneficial property that the h keywords with the largest product are necessarily placed in separate bins.

Adaptive-OLS: The ADAPTIVE-OLS strategy starts off mimicking the REGULAR-1 strategy, by placing each keyword in a single channel each day. This continues until we end up with an overdetermined system of equations which we solve using Ordinary Least Squares (OLS) (see, e.g., [8]). The key point of the adaptive strategy is that we then use those VPC estimates to inform our assignments of keywords to channels the following day as stated above.

3.3.3 Handling heteroskedasticity

Our final two methods differ from the previous in the way they employ least-squares estimation to compute residual values. Ordinary least squares estimation relies on an assumption of *homoskedasticity*, which means that the error in each of the measurements should have the same variance. In our case this assumption is clearly violated, since the variance in the measurement's error depends not only on the number of keywords packed into that measurement but also the variances in VPC values and numbers of measured clicks across keywords. Both of these latter terms can vary significantly.

Feasible Generalized Least Squares (FGLS) (again, see e.g. [8]) is a standard statistical cure for heteroskedasticity which we employ. Intuitively what we wish to achieve is to discount the measurements (equations) that might contain significant error so that the regression does no try to overfit them. To achieve that we use a two-step approach. First, we run OLS as above and compute the residuals associated with each equation. We then reweigh the equations dividing them by the square root of their corresponding residuals and run OLS a second time. Effectively, we are using the residual associated with each measurement (equation) as an estimate of the variance of its error. More sophisticated statistical techniques for estimating the variances of measurement errors do exist [8], but we have not investigated them within the current scope of our work.

Adaptive-FGLS: The ADAPTIVE-FGLS strategy is identical to ADAPTIVE-OLS except for the use of FGLS to correct for heteroskedasticity.

We emphasize that asymptotically both OLS and FGLS will provide the same correct answers, but FGLS converges to the final answer more quickly, as we confirm empirically.

We point that once we have introduced the possibility of handling heteroskedasticity, it makes sense to consider using FGLS in the context of oblivious strategies as well. Hence we also suggest the following variation of REGULAR-p:

Regular-p-FGLS: This strategy is identical to REGULAR-p except for the use of FGLS to correct for heteroskedasticity.

Again, we note that REGULAR-p itself does provide unbiased estimates of the keyword VPCs, but will not converge as quickly to the actual VPCs when the homoskedasticity assumption is violated.

4. EXPERIMENTAL RESULTS

Broadly speaking, there are two orthogonal axes along which our VPC estimation algorithms can be classified: adaptive versus oblivious; and algorithms which primarily employ measurements of single keywords as opposed to those which primarily employ *composite* measurements, i.e. placing multiple keywords in each channel. In our experimental section we explore this strategy space to elucidate the relative merits of each technique.

As a baseline, we selected the following choice of parame-
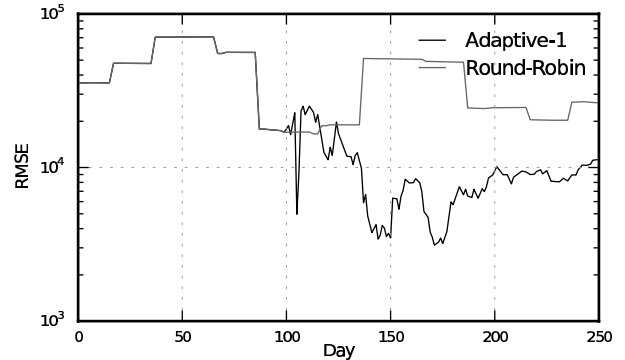
(a) High variance in VPC ($\sigma^2 = 2\mu$)
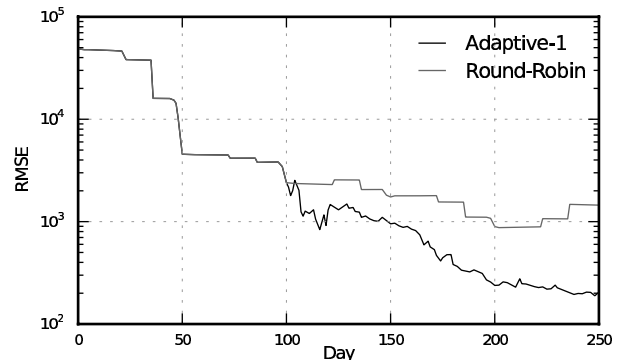
(b) Low variance in VPC ($\sigma^2 = 0.25\mu$)

Figure 1: RMSE convergence of single keyword per channel strategies.

ters as our default setting, with our primary considerations being the realism of the settings and the feasibility of conducting simulations. We recognize that the experiments depicted here provide a glimpse into only a portion of a vast set of possible settings; based on a much larger set of experiments we conducted, we believe our findings to be representative of a broader range of settings.

- 500 keywords and 10 channels.

- VPC and click means were sampled from a Zipf(1.8) distribution.

- The variance of the VPC normal variables was set to two times their mean while the variance of the click normal variables was set to a quarter of their respective means.

- The plots display the average performance over ten runs of each experiment.

4.1 Adaptive vs Oblivious: The one keyword per channel case

We begin our discussion by quantifying and explaining the effects of adaptivity on simple weighted average-based algorithms by comparing ADAPTIVE-1 with ROUND-ROBIN. Both algorithms assign a single keyword to each channel every day, but the former attempts to concentrate measurements on keywords with uncertain current estimates and

large potential click-weighted value and improve these estimates as early on as possible. In order to do so it begins by taking at least two measurements per keyword so that it can compute a perceived error for each. A comparison between these two basic approaches is depicted in Figure 1. As with all our subsequent plots, time appears on the x-axis, and the daily RMSE (as defined in Equation 1) is plotted on log-scale on the y-axis. The tradeoffs associated with adaptivity are clearly depicted in this plot. After day 100, ADAPTIVE-1 has enough information to calculate perceived errors and can focus its measurement efforts on keywords with large uncertainties in the current VPC estimates. In contrast, ROUND-ROBIN continues in a round-robin fashion, which fails to reduce the RMSE significantly.

Even though ADAPTIVE-1 outperforms its non-adaptive counterpart, it has some of the same limitations. First, the time overhead of gathering measurements to compute perceived errors is significant. In fact, to take the initial two measurements per keyword it needs to compute perceived errors ADAPTIVE-1 follows the same schedule as ROUND-ROBIN for the first 100 days, explaining their identical performance during that period. Ideally one would hope to exploit adaptivity earlier in the process — especially as the keyword-to-channel ratio increases.

Second, neither strategy exhibits a desirable monotonic decrease in RMSE over time in the presence of high variance in daily VPC measurements, as can be seen by the wide oscillations in RMSE in Figure 1(a). When the variance in VPC is much lower, as shown in Figure 1(b), convergence becomes smoother. Of course, lower variance can also be realized in the former scenario by conducting more measurements, and thus in both cases given enough time and enough measurements both strategies will converge to the same, correct answers.

Third, we reiterate a previously mentioned shortcoming common to any strategy that places only a single keyword on some channels: should such a keyword receive no clicks on that day, then that measurement yields no information. Even though ROUND-ROBIN is much more susceptible to this problem, ADAPTIVE-1 may also attempt to measure keywords that receive no clicks but have high VPCs. In practice, large-scale advertising campaigns that attempt to mine rare keywords within the long tail of possible search terms may include many terms that will likely not obtain a click on any given day. In our particular setting we varied τ^2, the variance of the normal distribution from which click frequencies are sampled. For the smaller setting, where τ^2 was 1/4 the mean, on average about 6% of our measurements yielded zero clicks, while when τ^2 was set to half the mean, that quantity rose to about 12%.

Lastly, we mention a shortcoming particular to our current implementation of ADAPTIVE-1. Recall that at least two measurements per keyword are required to compute a perceived error. To get these, ADAPTIVE-1 devotes the first $2n/h$ days measuring keywords individually in a round-robin fashion. But what happens should either or both of these measurements receive zero clicks? In the former case the VPC estimate of the keyword will equal that of the successful measurement, and in the latter case the VPC estimate will be zero (which is our prior for all VPCs). Moreover, in our implementation these keywords will then not participate in the adaptive phase of our algorithm, as they have no perceived error. Fortunately, we confirmed experimentally that

this effect is minimal, as estimation errors on these tail keywords rarely contribute significantly to the RMSE. Similar experiments with different rounding so that each keyword receives at least one click do not significantly change the results.

4.2 Adaptive vs Oblivious: The composite measurement case

We similarly consider the benefits of adaptivity when using schemes that conduct composite measurements that combine several keywords in a channel each day, coupled with daily regression-based valuation. Our regression-based methods are compared in Figure 2. In particular we compare two oblivious methods, REGULAR-1 and REGULAR-5, with ADAPTIVE-OLS. The most immediate improvement common to all regression-based techniques is that the RMSE is nearly monotonically decreasing with respect to elapsed time.
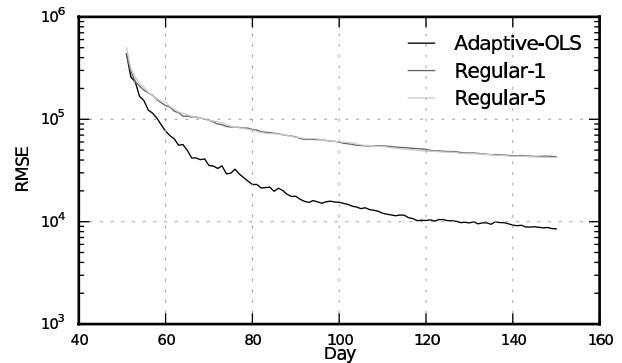

Figure 2: REGULAR-p vs ADAPTIVE-OLS: **using regression to combine adaptivity with multiple keywords per channel.**

As for the relative performance of the adaptive methods against the oblivious ones, the answer is similar to that of the single keyword case: adaptivity results in significant gains almost approaching an order of magnitude in RMSE improvement. A final observation is that placing a keyword in more channels per day as is the case with REGULAR-5 does not yield any improvement as compared with placing each keyword in a single channel. Indeed, the two curves are essentially indistinguishable on the logscale plot. The point is that it is not simply more measurements including a keyword that makes a difference; placing each keyword in five channels per day increases the number of variables per equation by a factor of five as well, increasing the noise and offering no gain.

4.2.1 Accounting for heteroskedasticity

As we have mentioned earlier, Ordinary Least Squares does not provide the fastest converging VPC estimates, because of the problem of heteroskedasticity (unequal variances of the errors in our measurements). The substantial gains provided by using Feasible Generalized Least Squares with the method labeled ADAPTIVE-FGLS can be seen in Figure 3(a). It is also worth highlighting that in a setting where the measurement error is larger, the benefits of using FGLS can become even more pronounced.
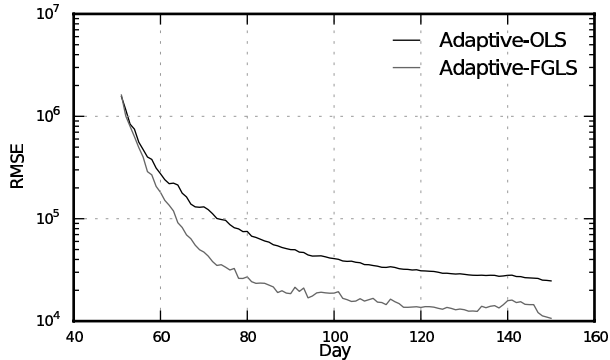
(a) Lower measurement error: $\sigma^2 = 2\mu$, $\tau^2 = \frac{1}{4}\nu$
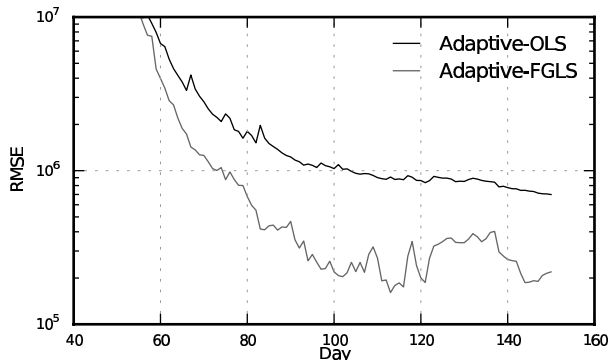
(b) Higher measurement error: $\sigma^2 = 4\mu$, $\tau^2 = \frac{1}{2}\nu$

Figure 3: Accounting for heteroskedasticity in the presence of various degrees of measurement error.

Figure 3(b) considers a setting where the sample variances for both the VPC distribution and click frequency distribution are double that depicted in Figure 3(a). Yet, while the gains in RMSE are apparent, the noisier environment also makes the performance of ADAPTIVE-FGLS less predictable, as seen in the period between day 120 and day 140 in Figure 3(b), where we see a sudden increase in RMSE. This is to be expected: Feasible Generalized Least Squares depends on knowing the true variance of error each measurement. Since this information is not available to us, we use the measurement residual as a proxy, and occasionally, this heuristic yields erroneous results.

4.3 Regression vs Weighted Averages

While we have seen the power of adaptivity, the question remains whether assigning multiple keywords to each channel and using regression to compute VPC estimates provides any real benefit over using just a single keyword on each channel. We present some results comparing these two strategies in Figure 4.

We first observe that, in the case of oblivious strategies, regression as employed by REGULAR-1 does not outperform the naïve ROUND-ROBIN technique. With 500 keywords, ROUND-ROBIN performs better in the early stages of optimization and REGULAR-1 catches up only towards the end as demonstrated in Figure 4. However, when we correct for heteroskedasticity, using REGULAR-1-FGLS, we can improve on
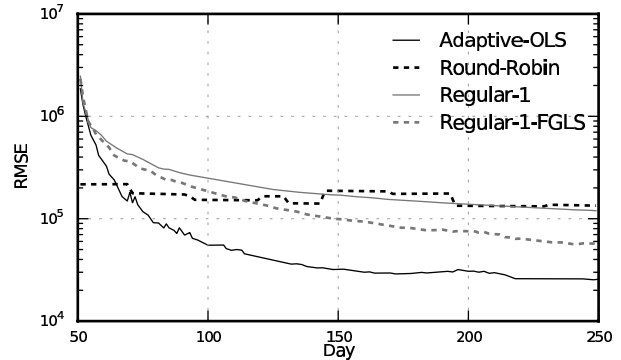
Figure 4: A comparison of multiple-keyword-per-channel regression methods with single-keyword-per-channel ones.

the performance of REGULAR-1 by a margin large enough to beat ROUND-ROBIN. We note the relative performance of ROUND-ROBIN deteriorates further with higher keyword-to-channel ratios, which decreases the frequency with which each keyword is measured, and when the variation in clicks per day is sufficient that the likelihood of a keyword obtaining zero clicks is larger. However, overall in the oblivious setting our experiments do not provide a compelling reason to prefer composite measurements over single-keyword measurements.

We now consider the more interesting setting of adaptive strategies, and consider the value of composite measurements. Figure 5 sheds some light on the question. Figures 5(a) and 5(b) demonstrate that under certain assumptions a simple adaptive technique such as ADAPTIVE-1 can outperform more sophisticated methods. The assumptions in this case are that the ratio of keywords to channels is not very high and far more crucially that the variance in the keyword VPCs is very low. Again, when the ratio of keywords to channels is low, each individual keyword is measured more frequently, improving the relative performance of ADAPTIVE-1. Similarly, when the variance in keyword VPCs is smaller, each single-keyword measurement has less noise which facilitates faster convergence to the true VPCs. While this effect holds for all schemes, it appears more pronounced in the case of single-keyword measurements. One final point with regard to the low-variance setting is that ADAPTIVE-FLGS doesn't improve much on ADAPTIVE-OLS. Again this is supported by the theory: lower variance implies lower heteroskedasticity and hence less of a benefit from employing FGLS.

On the other hand, Figures 5(c) and 5(d) showcase the conditions under which ADAPTIVE-OLS and ADAPTIVE-FGLS outperform the naïve techniques. We believe that high variance in VPCs and a large keyword to channel ratio are more indicative of real world advertising campaigns, hence making those methods an attractive choice.
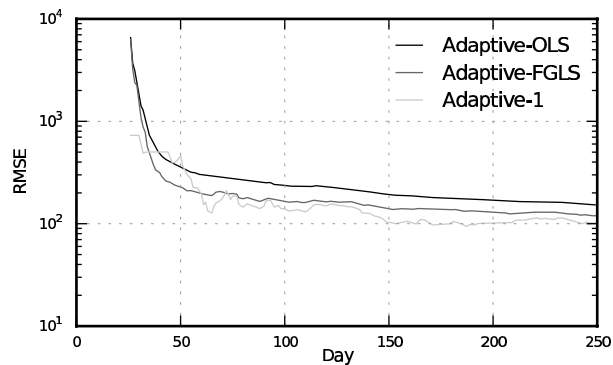
5. CONCLUSIONS AND FUTURE WORK

Our work opens the study of a key technical challenge facing online advertisers: that of valuing individual keywords when monetization data provided by third parties
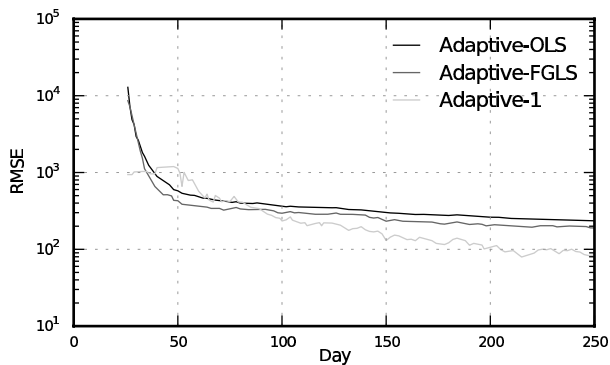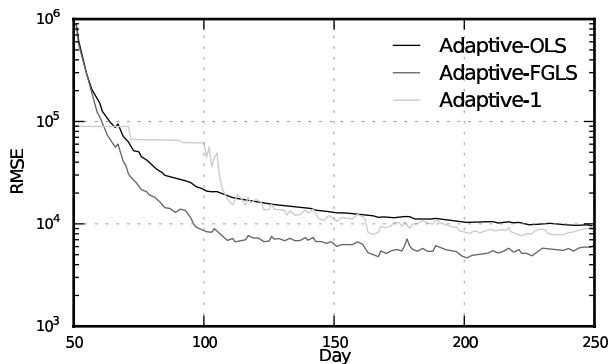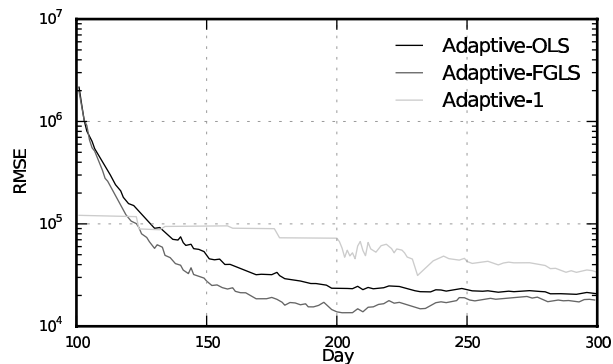
(a) 250 keywords, 10 channels, $\sigma^2 = \frac{1}{10}\mu$, $\tau^2 = \frac{1}{10}\nu$

(b) 500 keywords, 10 channels, $\sigma^2 = \frac{1}{10}\mu$, $\tau^2 = \frac{1}{10}\nu$

(c) 500 keywords, 10 channels, $\sigma^2 = 2\mu$, $\tau^2 = \frac{1}{4}\nu$

(d) 1000 keywords, 10 channels, $\sigma^2 = 2\mu$, $\tau^2 = \frac{1}{4}\nu$

Figure 5: A comparison of all adaptive strategies in various settings which showcase their relative merits.

is highly aggregated. The channelization problem we formulated necessitates several careful modeling choices, especially the distribution of VPCs and clicks across keywords and over time. With the high variability present in the realistic distributions we modeled, quickly identifying the most relevant keywords is central to minimizing our objective function, RMSE. For this reason, classical oblivious weighing designs underperform. More subtly, adaptive designs which focus on weighing individual keywords are also not ideal, as highly concentrated measurements fail to glean insights about the large set of unmeasured keywords. The challenges become more acute either as the ratio of keywords to channels increases, or as the variance in daily VPC measurements increase. The best designs we have identified exploit both adaptivity and composite keyword weighings followed by regression-based valuation. Also of interest is the importance of employing appropriate statistical methods in generalized least-squares estimation to address the significant heteroskedasticity that arises in our experimental datasets. We believe the question of finding adaptive design algorithms, especially in complex settings similar to ours where both clicks and VPC vary on each measurement, is a subject worthy of further attention and may find further applications.

Our future work focuses on improving the automation of our designs to learn the (initially unknown) click and VPC distributions and adjust the density of composite weighings accordingly. We are also interested in methods to speed up the learning phase associated with adaptive designs to reduce start-up overhead, and in studying settings of much larger scale when the keyword-to-channel ratio is potentially orders of magnitude greater. Finally, for simplicity we have chosen to model clicks and VPCs as random variables with distributions that are not time-dependent. Considering stochastically varying click or VPC distributions in the context of weighing designs appears to be an open statistical challenge.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] G. Aggarwal, J. Feldman, and S. Muthukrishnan. Bidding to the top: VCG and equilibria of position-based auctions. *Lecture Notes in Computer Science*, 4368:15, 2006.

[2] 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. ACM New York, NY, USA, 2006.

[3] K. Banerjee. *Weighing designs for chemistry, medicine, economics, operations research, statistics.* Marcel Dekker, Inc. New York, NY, 1975.

[4] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 559–566. ACM New York, NY, USA, 2007.

[5] M. Cary, A. Das, B. Edelman, I. Giotis, K. Heimerl, A. Karlin, C. Mathieu, and M. Schwarz. Greedy bidding strategies for keyword auctions. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 262–271. ACM New York, NY, USA, 2007.

[6] E. G. Coffman, Jr., M. R. Garey, and D. S. Johnson. Approximation algorithms for bin packing: a survey. pages 46–93, 1997.

[7] 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):242–259, 2007.

[8] W. Greene. *Econometric analysis*. Prentice Hall Upper Saddle River, NJ, 2003.

[9] H. Hotelling. Some improvements in weighing and other experimental techniques. *The Annals of Mathematical Statistics*, pages 297–306, 1944.

[10] B. Kitts and B. Leblanc. Optimal bidding on keyword auctions. *Electronic Markets*, 14(3):186–201, 2004.

[11] A. Mood. On Hotelling's weighing problem. *The Annals of Mathematical Statistics*, pages 432–446, 1946.

[12] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 403–410. ACM New York, NY, USA, 2008.

[13] A. Street and D. Street. *Combinatorics of experimental design*. Oxford University Press, Inc. New York, NY, 1986.

[14] H. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.

[15] F. Yates. Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, pages 181–247, 1935.