

# **A Model of Individual Keyword Performance in Paid Search Advertising**

Oliver J. Rutz

Randolph E. Bucklin\*

June 2007

Oliver J. Rutz is Assistant Professor of Marketing, Yale School of Organization and Management, 135 Prospect Street, New Haven, CT 06520, [oliver.rutz@yale.edu](mailto:oliver.rutz@yale.edu). Randolph E. Bucklin is Professor of Marketing, UCLA Anderson School, 110 Westwood Plaza, Los Angeles, CA 90095, [rbucklin@anderson.ucla.edu](mailto:rbucklin@anderson.ucla.edu). The authors wish to thank a collaborating firm for providing the data used in this study.

## **A Model of Individual Keyword Performance in Paid Search Advertising**

### **Abstract**

In paid search advertising on Internet search engines, advertisers bid for specific keywords, e.g. “Rental Cars LAX,” to display a text ad in the sponsored section of the search results page. The advertiser is charged when a user clicks on the ad. Many of the keywords in paid search campaigns generate few, if any, sales conversions – even over several months. This sparseness makes it difficult to assess the profit performance of individual keywords and has led to the practice of managing large groups of keywords together or relying on easy-to-calculate heuristics such as click-through rate (CTR). The authors develop a model of individual keyword conversion that addresses the sparseness problem. Conversion rates are estimated using a hierarchical Bayes binary choice model. This enables conversion to be based on both word-level covariates and shrinkage across keywords.

The model is applied to keyword-level paid search data containing daily information on impressions, clicks and reservations for a major lodging chain. The results show that including keyword-level covariates and heterogeneity significantly improves conversion estimates. A holdout comparison suggests that campaign management based on the model, i.e., estimated cost-per-sale on a keyword level, would outperform existing managerial strategies.

**Keywords: Internet, Advertising, Paid Search, Bayesian Methods**

## **Introduction**

Paid search advertising<sup>1</sup> allows companies to address consumers directly during their electronic search for products or services. When a consumer searches for products or services with the help of an Internet search engine, he uses keywords (a keyword can consist of multiple words) such as “Hotels Los Angeles” to start the search. A company in the lodging business can address this consumer directly via paid search. It “buys” specific keywords, e.g., “Hotels Los Angeles” and creates a text ad that will appear when a consumer searches for these keywords. The serving of a text ad is called an impression. Paid search ads are displayed in the clearly marked sponsored section of the search results page (see Figure 1). In most cases, multiple advertisers are interested in the same keywords. For example, a keyword such as “Hotels Los Angeles” is attractive to almost any lodging company in the Los Angeles area.

From an advertiser’s perspective, paid search ads are not like traditional ads. Advertisers cannot buy their place into the listings for a fixed dollar amount. Unlike banner ads, for example, advertisers cannot buy “web real estate.” In paid search, advertisers bid what they are willing to pay for a click on a paid search ad. An automated auction algorithm then determines placement and position, e.g., 1<sup>st</sup> or 3<sup>rd</sup>, of the ad in the sponsored listings. Paid search is pay-for-performance advertising. Companies do not pay for impressions served, but for actual clicks on their paid search ads. This may be one of the reasons for the rapid growth of paid search over the past several years.

Currently, the two major forms of online advertising are paid search and display (or banner) advertising. In display advertising consumers are exposed to ads that appear as banners on a web-page. Display advertising had been the preeminent form of online advertising until 2005 when paid search overtook it. The largest search engine, Google, generated more than \$4 billion

---

<sup>1</sup> We will use paid search instead of paid search advertising for the remainder of the paper.

in net ad revenue in 2005 with their paid search operation called AdWords (Business Week 2006).

Despite the rapid growth and scale of paid search, there has been little academic study of this advertising service, especially in the marketing literature. In paid search an advertiser faces four different decisions (or levers): (1) which keywords to select, (2) how much to bid for each keyword, (3) how to design the text ad and (4) how to design the landing page. Some existing research has focused on decision two: bidding. Game-theoretic in nature, this research does include limited empirical applications (Edelman et al. 2007, Chen and He 2006). To the best of our knowledge, no research has specifically investigated the design of the text ad and the landing page. On the other hand, designing ads and landing pages might not differ materially from traditional ad and web design, suggesting that findings from previous studies might apply.

This leaves the first question: which keywords to select? Much of current management practice is to use heuristics or to assess campaign performance on an aggregate level, i.e., group keywords together and manage the resulting groups. Because the mechanism of paid search (including data reporting by the search engines) operates at the keyword level, this raises the question how decisions should be made at that level. Measuring keyword performance is an often difficult problem in paid search because the data are sparse, i.e., for most keywords few or no conversions to purchase (or other desired online action) are observed. Often this is the case even over extended periods of time.

Our dataset contains daily information on paid search advertising from a major lodging chain on a keyword-level. A traditional approach to performance evaluation would be to calculate the marginal benefit of spending on each keyword, comparing advertising-related revenue-per-reservation with advertising-related cost-per-reservation<sup>2</sup> (CPR). If that difference is positive, a keyword is attractive and the company should retain it. If not, the company should drop this

---

<sup>2</sup> For the remainder of the paper we will use cost-per-reservation instead of advertising-related cost-per-reservation for the convenience of the reader.

keyword from the campaign. Alternatively, the company could also consider adjusting its bidding strategy. Search engines, however, do not provide their advertisers with data on the actual auction or competitive bids. We therefore base this work on the data that are made available to advertisers by search engines and used by campaign managers, leaving competitive behavior and auction issues as topics for future research.

At a first glance, assessing individual keyword performance on a CPR basis seems straightforward. CPR is given by

$$CPR = \frac{\text{Cost - per - Click (CPC)}}{\text{Conversion Rate (ConvRate)}}.$$

Most keywords have impressions and clicks (costs) associated with them on a daily basis. But while searches lead to impressions, and sometimes clicks, they do not always lead to sales. Indeed, average conversion rates in paid search are very low. In April 2004, for example, only 84 out of 301 keywords (27%) in our data set led to reservations (see Figure 2). Thus, on a daily, or even weekly or monthly, basis most keywords simply did not generate any sales. This observed conversion rate of zero precludes us from calculating CPR. Of course, the true, long-term conversion rate may not be equal to zero for these keywords. But because the rate might be very small, it could take quite a while for a sale to occur. In the meantime, the firm continues to pay for the clicks on its ad without any accompanying revenue to show for it.

The purpose of this paper is to develop a method to evaluate the performance of individual keywords based on cost-per-sale or, in our case, cost-per-reservation (CPR). We condition our analysis on a click. This gives us cost-per-click (CPC) information for each keyword. To calculate CPR we need the keyword's conversion rate. To obtain a model-based estimate for this, we conceptualize conversion as a binary choice decision conditional on a click (i.e., a user has clicked on the paid search ad and has been taken to the landing page on the company's website).

We use shrinkage procedures to infer conversion rates for keywords with no or very few sales (and to improve the estimated conversion rates for other keywords). Taking advantage of a keyword's similarity to other keywords, we can produce a shrinkage-based conversion rate estimate for each one. Our approach draws on commonly used methods in choice modeling and provides a statistically-sound way to address the sparseness problem. The estimated set of conversion rates allows us to evaluate the individual performance of keywords in terms of CPR, circumventing the inherent sparseness of the raw paid search conversion data.

From our model we find that position, click-through rate (CTR) and keyword characteristics are significant predictors of conversion rates for keywords. Our results show that the conversion rate depends on the underlying keyword and other factors that are related to the paid search text ad. We also find that when modeling individual keyword performance we need to take keyword heterogeneity into account.

We also assess the predictive power of our model in a holdout sample. The list of attractive keywords generated by our model outperforms commonly used model-free managerial strategies. Model-free strategies used by managers either focus on keyword category management or other easy to calculate performance measure such as click-through-rate. Category management approaches do not allow the advertiser to assess the performance of individual keywords and cannot be used to optimize the existing keyword list. Focusing on click-through rate or other easy to calculate measures such as number of impressions can be hazardous. These approaches almost always require the manager to arbitrarily pick a threshold, e.g., keep all keywords with  $CTR \geq 1\%$ . It is not clear whether this type of approach will result in profit maximization.

Our intended contribution is fourfold. First, we address the problem of individual keyword management in paid search. Second, we find that the underlying keyword is predictive of the conversion decision. Third, we show that when modeling keywords performance, keyword level heterogeneity and keyword characteristics need to be addressed. Fourth, we find that model-free managerial strategies are inferior to our model based approach.

While this study focuses on the performance of keywords at the individual level, we hope that the ability to better assess keyword performance will allow managers, going forward, to evaluate the impact of keyword selection, bidding strategies and ad copy/landing page design. In this paper, we will not study the impact of these levers, i.e., we will not investigate, for example, the optimal bid amount for a given keyword. The position of the keyword, however, is taken into account in our investigation of individual keyword performance.

The paper is structured as follows. First we give an introduction to paid search and summarize the relevant literature on online advertising. We then present our dataset, model and results. Next we discuss the implications of our findings and illustrate how to improve the performance of a paid search campaign by individual keyword management. We finish with a conclusion, the limitations of our approach, and discuss future research in the realm of search engines and marketing.

## **Background and Literature**

Given the very limited exposure of paid search advertising in the marketing literature, we first give a brief overview of the topic.

### **Paid Search Advertising**

How does paid search work from a company and consumer perspective? Consider the following scenario: a consumer searches using the keyword “Hotels LAX”. A company had previously selected this keyword and created a text ad. The search results page will display non-sponsored (so called organic) and sponsored search results, including the ad of the company (depending on the bid). From the company’s perspective, one impression has been generated which is attributed to the keyword “Hotels LAX”. Imagine the consumer likes the company’s text ad and decides to click on it. He is transferred to the landing page on the company’s website and the keyword “Hotels LAX” has now generated a click. The consumer then decides whether to reserve a hotel

room. If he does so, the keyword “Hotels LAX” becomes associated with a reservation from the company’s perspective (see Figure 3).

What might the company be willing to pay for this service? How is the position of the text ad determined in relation to all other companies that also have selected “Hotels LAX”? These two questions are intertwined and one cannot be answered without addressing the other. The search engines’ paid search revenue model does not follow traditional advertising practices – advertisers neither pay for “web real estate” nor for impressions served. In paid search, the advertiser bids the maximum dollar amount he is willing to pay for a click on a text ad served in response to a search for a keyword. The actual cost-per-click (CPC) and position of the text ad are then determined by an auction style method. Paid search is pay-for-performance advertising: advertisers are charged for clicks and not impressions.

What types of auction mechanisms are used? Google, on the leading edge of paid search, combines the bids with a proprietary estimate of click-through rate (CTR) and calculates expected revenue from the ads. The best position is given to the ad (company) with the highest expected revenue for Google. Other ads (companies) follow in order of expected revenue. Search engines on the other side of the spectrum, e.g., Yahoo!, used pure second price auctions to determine the positions of the paid search ads up to January 2007. Here the highest bidder is given the best position; all others follow in order of bids. BusinessWeek in 2006 reported that Google earns about 30% more than Yahoo! per ad impression with its expected revenue model. MSN adopted the Google model in fall 2005 and Yahoo announced that it is switching to more use of the click-through rate in placing search results.

The software provided to advertisers by the search engines allows managers to change the paid search strategy on a near real time basis. The selection of the keywords and the bid per keyword can be changed at any time during the day. The software, e.g., Google’s Adwords, allows advertisers to group keywords, e.g., Google’s Adgroups, and to manage these groups of keywords. Paid search strategies are commonly managed based on these groups.

To the best of our knowledge, no published research about paid search advertising has yet appeared in marketing literature. Research has so far focused on search engine performance. Telang et al. (2004) investigate search engine visits, whereas Bradlow and Schmittlein (2000), Spiteri (2000), and Jansen and Molina (2005) focus on the effectiveness of search engines in information search. Paid search auction mechanisms have recently become the topic of game theoretic papers in the economics literature (Edelman and Ostrovsky 2007, Edelman et al. 2007). Paid search advertising is also investigated from an information signaling viewpoint as a product differentiation game (Chen and He 2006).

Edelman and Ostrovsky (2007) examine data on paid search auctions and find evidence of strategic bidder behavior in these auctions. They estimate that Overture's revenue from sponsored search might have been higher if it had been able to prevent this strategic behavior. The authors present a specific alternative mechanism, a second price auction, which could reduce the amount of strategizing by bidders, raise search engines' revenue, and also increase the overall efficiency of the market.

Following up on their previous work Edelman et al. (2007) investigate the "generalized second price" auction (GSP), a new auction mechanism which is used by search engines to sell paid search advertising. GSP is not a true second price auction mechanism (Vickrey-Clarke-Groves, VCG) as claimed by search engines. Although GSP looks similar to VCG, the authors find that its properties are very different, and equilibrium behavior is far from straightforward. They find that naive buyers who (incorrectly) base their bidding behavior on a second price auction mechanism end up paying more than they need to. More savvy advertisers, however, do strategically bid lower than in a true second price auction.

Chen and He (2006) study a product-differentiation game where consumers are initially uncertain about the desirability of, and valuation for, different sellers' products, but can learn about a seller's product through a costly search. They find that a seller bids more for placement

when his product is more relevant for a given keyword. This results in efficient (sequential) search by consumers and increases total output.

### **Online advertising**

In marketing research, online advertising has been mostly synonymous with banner advertising. The revenue models of banner advertising have been driven by a pay-for-performance approach and click-through rates have been a key measure of online advertising performance (Hoffman and Novak 2000, Dahlen 2001). However, compared with paid search advertising consumers rarely click on banner ads. Click-through rates declined since the 1990s from 7 percent in 1996 to around 0.3% in 2002 (DoubleClick 2003). Dreze and Hussherr (2003) find that click-through rates are low because consumers actually avoid looking at banner ads during their online activities. If consumers avoid looking at banner ads it might imply processing of banners at the pre-attentive level. If such is the case, click-through rate is an ineffective measure of banner ad performance and traditional measures, such as brand awareness and brand recall, are more appropriate. This is in marked contrast with paid search advertising where consumers actively search for products or services and are therefore more likely to pay attention to ads.

Building on the above, Danaher and Mullarkey (2003) discover that user involvement plays a crucial role in understanding banner ad effectiveness. If users are in a goal-directed mode (i.e., actively searching for a product or service) they are much less likely to recall and recognize banner ads than users who are in a surf mode. Users in a goal-directed mode simply ignore “noise” in the form of unwanted banner ads. Cho and Choen (2004) investigate why this happens. They find that (1) perceived ad clutter, (2) prior negative experience and (3) perceived goal impediment are the major drivers of advertising avoidance. Perceived goal impediment is found to be the most significant antecedent explaining advertising avoidance on the Internet. They recommend that advertisers use highly customized context-congruent advertising messages to reduce perceived goal impediment. Moore et al. (2005) also investigate the importance of

congruity between the website and the ad and find that congruity has favorable effects on attitudes.

Paid search advertising addresses many of the issues raised about banner advertising. Click-through rate is a meaningful measure of performance, as consumers actually click on the text ads and advertisers are only billed for clicks. Paid search text ads are context-congruent with the current goal of the searcher and are therefore not perceived as goal impediments. Consumers are paying attention to ads they consider helpful and click on them. These factors may explain the faster growth in paid search relative to banner ads.

## **Data**

The data generated by paid search differ significantly from traditional advertising data. In paid search the advertiser bids for the search engine to display a text ad in the sponsored section of the results page. This text ad usually consists of a headline, a limited number of words and a link to the company's website. These words may describe the nature of the company, the content of the site or any current offers. The serving of a text ad in response to a search for a certain keyword is counted as an impression. If the searcher clicks on the text ad, he is transferred to the "landing page" on the company website. This is recorded as a click and companies are billed per click. If, based on this visit to the company site, a sale (in our case a reservation) occurs it is recorded in the data as a conversion (see Figure 3).

Our data represents a quarterly snapshot of the Google campaign for a major lodging chain. It begins in April 2004 and spans 3 months. The data consist of two parts – standard information that advertisers receive from Google and complementary, additional information purchased from a third party data provider. The standard information supplied by Google includes daily information on an individual keyword level. For each keyword (e.g., Hotels Los Angeles) we have daily information on cost (in \$), average position served (ranking, e.g., 2.3), and number

of impressions and clicks. This information allows us to assess the costs of the campaign. However, we can not address the issue of financial performance with these data. This is where the additional third party data come in, providing daily information on the number of reservations for each keyword. In combination, the two datasets allow us to investigate the financial performance of individual keywords.

The company bid on 405 keywords in April 2004. Of these, only 301 keywords led to at least one click. We focus our investigation on these 301 keywords for two reasons: First, in the absence of any clicks, the company does not incur any costs. No-click keywords are free advertising and the company has nothing to lose by continuing to bid on them. Second, and more important, Google's new bidding rules (implemented at the end of 2005) would not allow bidding on keywords that do not lead to any clicks over a certain period of time. Google runs its business on expected revenue – it estimates the expected revenue from each keyword based on the bid and the expected click-through rate. The position of the text ad in the listings is determined by expected revenue and not simply by the amount bid. One of Google's requirements is that the click-through rate of each keyword has to be above a minimum threshold (Google does not make the number public). If the click-through rate of a keyword has been below the threshold for a certain time Google will put the keyword on probation. If the keyword's performance does not improve, Google will disable the keyword and the company will not be able to bid on it anymore (Goodman 2005). Most other search engines have already implemented a similar auction mechanism (MSN) or are in the process of implementing it (Yahoo!).

We enhance the data by introducing semantic keyword characteristics. The keywords used have certain common characteristics that are specific to the lodging industry, e.g., a city or a holiday destination is included. We “decompose” each of the 301 keywords along the following set of characteristics:

1. *Branded*: Is the company brand name included? 99 keywords are branded.
2. *US*: Is the keyword for a US location? 223 keywords are for a US location.

3. *State*: Does the keyword include a state name? 52 keywords include a state name.
4. *City*: Does the keyword include a city name? 210 keywords include a city name.
5. *Hotel*: Does the keyword include the word hotel or other lodging related phrases such as accommodation, motel, or room? 222 keywords include hotel or other related phrases.

The correlation between keyword characteristics is very low with the following notable exception: branded and hotel are highly negatively correlated (-0.83). This is as expected because in the absence of the brand name phrases that indicate lodging are needed, e.g., “Hotel Los Angeles.”

In many respects keyword characteristics are analogous to demographics, so we label them *wordographics*. We believe that wordographics might be important for studying individual keyword performance and investigate whether wordographics allow us to better understand the performance of a keyword. Analogous wordographic decompositions should be possible for other keyword datasets from different industries.

We use the data for April 2004 as an estimation sample. In April 2004 the campaign (based on 301 keywords with at least one click) generated 2,281,023 impressions, 14,302 clicks and 518 reservations (see Table 1). The average position was 6.0 and the company spent \$5,106.74 on the campaign. The average click-through rate (percentage of impressions that led to a click, CTR) was 0.6% and the average conversion rate (percentage of clicks that led to a reservation, ConvRate) was 3.6%. The average cost-per-click (CPC) was \$0.36 and the average cost-per-reservation (CPR) was \$9.86 (see Table 1). We use the data from May and June 2004 as a hold-out sample. The performance of the paid search campaign in May and June is very similar to April (see Table 1): in terms of conversion and cost-per-reservation, there is little difference from the estimation sample.

Practitioners typically evaluate paid search using aggregate measures because data on most individual keywords are sparse. The sparseness is in the reservation data, as most of the

keywords in our dataset generate very few reservations over the entire period. On a daily, weekly and even monthly basis most keywords do not lead to reservations. Out of the 301 keywords only 84 “conversion” keywords led to reservations (conversion) in April 2004. While we do not have revenue or margin information, we know that the average price range for a room is between \$75 and \$100 per night. Assuming an average of 1.5 nights per trip and a profit margin of 30%, the cost-per-reservation should not exceed \$40. Only 4 keywords have a CPR that is higher than \$40.

How should the company manage its campaign? What is the conversion rate for the 207 “non-conversion” keywords without reservations? What is the value of these words in terms of CPR? Some of the 84 “conversion” keywords have very few reservations. Can we trust the conversion rate (point) estimates for these keywords when the confidence intervals are big? It seems that managing by observed conversion rates might not be such a good strategy. Consultants in the field often recommend managing by click-through-rate. However, it is not clear how a cut-off point can be established that is consistent with profit maximization. Managing a campaign in order to maximize profit requires a marginal cost-benefit analysis which normally fails because of the sparseness of the conversion data (see above). We address this problem with our modeling approach by estimating shrinkage based keyword-level conversion rates. This allows us to measure the performance of individual keywords from a profit perspective (CPR).

## **Model**

We conceptualize conversion as a binary choice conditional on click. A visitor is transferred to the company landing page after he has clicked on a paid search text ad. Does this click lead to a reservation? We investigate whether the conversion probability can be predicted based on the keyword that was used in the search. Our data do not include information on the visitor level, i.e.,

we do not have clickstream data<sup>3</sup> that would allow us to model a visitor’s decision. We do not investigate what drives visitors to make a reservation. Our study is focused on measuring keyword performance, i.e., the cost-per-reservation (CPR). We build our model with keywords as the unit of investigation and study conversion rates on a keyword-level. We seek to understand whether conversion rates differ across individual keywords and if so, whether we can explain those differences based on observable covariates.

We employ the binary logit model to investigate the probability of making a reservation conditional on a visitor reaching the company landing page via a click. That is, we model conversion conditional on click. The daily clicks for each individual keyword are used as choice occasions, whereas the daily reservations for each individual keyword represent the “successful” choices. We do not model how consumers may differ in their choices to make a reservation. Based on the binary logit model the conversion probability ( $PConv_{w,t}$ ) for keyword  $w$  at time  $t$  is given by

$$(1) \quad PConv_{w,t} = \left( \frac{\exp(v_{w,t})}{1 + \exp(v_{w,t})} \right)$$

where  $v_{w,t}$  represents a keyword’s latent value.

In a standard application of the logit model, the data contain one choice outcome (observation) per time period. Since we do not have clickstream data, we cannot link a specific click (choice occasion) to a specific reservation (successful choice). Thus, our data usually have more than one choice occasion per time period (daily). We observe the numbers of clicks and the number of reservations for each individual keyword on a daily basis. The likelihood function is given by:

---

<sup>3</sup> Clickstream data assigns each visitor to the site an individual ID (cookie). Clickstream data would allow researchers to connect a reservation to a specific click and that click, in turn, to a specific impressions and keyword. Google has not provided impression and keyword data on a user specific level in the past (smallest aggregation level is hourly) and experts in the field believe that Google has no intention of doing so in the future.

$$(2) \quad \text{Likelihood} = \prod_{t=1}^T \prod_{w=1}^W (P\text{Conv}_{w,t})^{\text{reservations}_{w,t}} * (1 - P\text{Conv}_{w,t})^{(\text{clicks}_{w,t} - \text{reservations}_{w,t})}$$

where  $t$  is time from April 1<sup>st</sup> until April 30<sup>th</sup> 2004 and  $w$  is keyword from 1 to 301.

### Latent Value of Keywords

What drives the latent value  $v_{w,t}$  of keywords? We frame our study of this question along two substantive issues:

- 1) *Do conversion rates differ systematically across keywords?* In other words, is knowing where (i.e., which keyword was used) a visitor to the site comes from predictive of conversion?
- 2) *Are campaign metrics “valuable” in predicting conversion rates?* Can we improve the predictive power of the model by including keyword-ad covariates (e.g. position)?

There is no existing research on the latent value of keywords in predicting conversion (to the best of our knowledge). This is an important empirical issue in paid search and our study provides a first opportunity to gauge if and how managers can build conversion models based on keyword information.

Our logit approach allows us to investigate the four cases (see Figure 4) in the same framework. We use shrinkage to address the sparseness in the data. This enables us to leverage information across similar keywords to better predict conversion rates for keywords with no or very few conversions. We study the performance of two different shrinkage procedures – homogeneous vs. heterogeneous shrinkage (Andrews et al., 2002). A homogeneous shrinkage procedure shrinks to the common mean (ML models) – a click is a click is a click. In other words, we assume that keywords are not predictive of conversion. We implement the homogeneous shrinkage procedure through Maximum Likelihood estimation. In heterogeneous shrinkage (or random effects/coefficients) we assume that an underlying prior distribution exists that allows for, in combination with the data, individual-level posterior estimates to be distributed around the

posterior mean. In other words, each keyword can have an individual, potentially different, conversion rate (Bayes and HB models). Here not every click has same the value (i.e., probability of conversion) to the firm and keywords help in predicting observed differences in conversion rates. We implement the heterogeneous shrinkage procedure by using Bayesian methods and estimate by Markov Chain Monte Carlo.

We first study substantive issue 1 – *are keywords predictive of conversion rates?* We compare a model that postulates that keywords are not predictive of conversion rates (Model I) with a model that allows conversion rates to systematically differ across keywords (Model II).

**Model I : ML<sup>Null</sup>.** One way to model individual-level conversion rates is to postulate that a click is a click is a click. In other words, the probability of conversion is independent of the underlying keyword (used by the visitor to find the site). The ML<sup>Null</sup> does not allow for evaluation of individual keywords on the basis of conversion rates, as all keywords are assigned the same conversion rate. Observed differences in conversion rates are based on other unobserved factors and sampling error only.

The ML<sup>Null</sup> model is given by:

$$(3) \quad v_{w,t} = \beta_0 + \varepsilon_{w,t}$$

where

$\beta_0$  is a common (homogeneous) intercept term,

$\varepsilon_{w,t}$  = Logit error for keyword  $w$  at time  $t$

and

$\beta_0$  is a parameters to be estimated by Maximum Likelihood (ML).

**Model II: Bayes<sup>Null</sup>.** The assumptions of the ML<sup>Null</sup> model seem overly restrictive. The data indicates that conversion rates differ systematically across keywords. Thus, keywords might be predictive of conversion rates. The Bayes<sup>Null</sup> model allows for conversion rates to differ across keywords – a pure random effects model. We could estimate each keyword’s conversion rate by

specifying a common distribution, e.g., Poisson, and estimate the parameters of this distribution. These types of stochastic modeling approaches were very popular in marketing up until the mid 80s (e.g., Morrison and Schmittlein 1981, Schmittlein et al. 1985). An equivalent approach within our binary logit framework is to use a heterogeneous keyword intercept.

The Bayes<sup>Null</sup> model is given by:

$$(4) \quad v_{w,t} = \beta_{0,w} + \varepsilon_{w,t}$$

where

$\beta_{0,w}$  is an individual-level (heterogeneous) intercept vector,

$\varepsilon_{w,t}$  = Logit error for keyword  $w$  at time  $t$

and

$\beta_{0,w}$  is a parameter vector to be estimated by Markov Chain Monte Carlo (MCMC).

In the next step we study the second substantive issue – *are campaign metrics “valuable” in predicting conversion rates?* Campaign metrics (e.g., position) could also explain the observed differences in conversion rates. We investigate the value of campaign metrics in our logit framework and again distinguish between the cases of keywords being predictive or not.

**Model III: ML<sup>CV</sup>.** Do observable covariates help us to better estimate individual conversion rates? In an extension of the ML<sup>Null</sup> model we assume, again, that keywords are not predictive of conversion rates. Differences in observed conversion rates, however, are not only driven by unobserved effects and sampling error as in the ML<sup>Null</sup> model. Conversion rates can be explained using observable covariates, e.g., position. This is analogous to traditional choice models: covariates such as price or display enter the latent utility function. We assume in the ML<sup>CV</sup> model that the effect of covariates is homogenous across keywords, e.g., the position sensitivity ( $\beta_i$ ) is the same across keywords.

The ML<sup>CV</sup> model reflects the spirit of Chen and He’s (2006) game-theoretic model. Their key finding is that a company should always bid the ad in the “market-relevant” position, i.e., if the product is the second most relevant product in the market the ad should be in position 2. Thus, from their perspective keywords do not matter and only position is relevant. Our nested modeling approach allows us to test their theoretic results in an empirical setting.

The ML<sup>CV</sup> model is given by:

$$(5) \quad v_{w,t} = \beta_0 + \beta_1 POS_{w,t} + \beta_2 CTR_{w,t} + \beta_3 COST_{w,t} + \varepsilon_{w,t}$$

where

$POS_{w,t}$  = Average position of keyword  $w$  at time  $t$

$CTR_{w,t}$  = Click-through rate of keyword  $w$  at time  $t$

$COST_{w,t}$  = Cost incurred for keyword  $w$  at time  $t$

$\varepsilon_{w,t}$  = Logit error for keyword  $w$  at time  $t$

and

$\beta_0, \beta_1, \beta_2,$  and  $\beta_3,$  are common parameters to be estimated by ML.

**Model IV: Bayes<sup>CV</sup>.** The Bayes<sup>CV</sup> model is an extension of the Bayes<sup>Null</sup> model. We assume that keywords are predictive of conversion rates and that the effect of covariates is heterogeneous across keywords – a random coefficients model. In other words, the sensitivity ( $\beta_i$ ) can differ across keywords. For example, a higher position may be of greater value for a highly competitive keyword such as “Hotels Los Angeles” than for a less competitive keyword such as “Santa Cruz Hotel”.

The Bayes<sup>CV</sup> model is given by:

$$(6) \quad v_{w,t} = \beta_{0,w} + \beta_{1,w} POS_{w,t} + \beta_{2,w} CTR_{w,t} + \beta_{3,w} COST_{w,t} + \varepsilon_{w,t}$$

where

$POS_{w,t}$  = Average position of keyword  $w$  at time  $t$

$CTR_{w,t}$  = Click-through rate of keyword  $w$  at time  $t$

$COST_{w,t}$  = Cost incurred for keyword  $w$  at time  $t$

$\varepsilon_{w,t}$  = Logit error for keyword  $w$  at time  $t$

and

$\beta_{0,w}$ ,  $\beta_{1,w}$ ,  $\beta_{2,w}$  and  $\beta_{3,w}$  are individual-level (heterogeneous) parameter vectors to be estimated by MCMC.

Assuming we are able to find that keywords are predictive of conversion rates, we can further explore keyword heterogeneity in a hierarchical approach. Can leveraging keyword characteristics, i.e., wordographics, help to better predict conversion rates?

**Model V: HB<sup>WG</sup>.** Can we better understand the individual-level keyword conversion rates  $\beta_{0,w}$ ? We decompose the keywords into a set of common characteristics, which we call wordographics in relation to demographics (see Data section). Following Bayesian literature in general and Steenburgh et al. (2003) in particular we propose to incorporate wordographics through a hierarchical approach (HB<sup>WG</sup> model). We also test in a rival approach whether wordographics are better treated as covariates and include them as indicators in equation 5 (ML<sup>CV&WG</sup> model) and equation 6 (Bayes<sup>CV&WG</sup> model).

We investigate whether individual intercepts,  $\beta_{0,w}$ , can be decomposed into a “standard” and a “keyword-effect” intercept (see Equation 7 and 8). The “standard” intercept captures individual keyword heterogeneity that cannot be explained by either covariates (analogous to a brand intercept in a brand choice model) or wordographics. The “keyword-effect” intercept enables us to use information on wordographics. This approach is analogous to traditional choice modeling: consumer demographics are often non-significant when included in the utility function as covariates. Using demographics in a hierarchical fashion, however, often improves model fit and leads to new insights about consumer behavior (Ainslie and Rossi 1998).

The HB<sup>WG</sup> model is an enhancement of the Bayes<sup>CV</sup> model and is given by

$$(7) \quad v_{w,t} = \underbrace{X_{w,t}^T \beta_w}_{\text{Covariates}} + \underbrace{\omega_w}_{\text{Keyword effect}} + \varepsilon_{w,t}$$

$$(8) \quad \omega_w = C_w \gamma + v_w \quad v_w \sim N(0, V_w)$$

where  $X_{w,t}^T \beta_w$  are the previously introduced covariates (see Equation 6),  $\omega_w$  is the individual keyword effect,  $C_w$  is a matrix of wordographics (without an intercept to ensure identification) and  $\gamma$  a parameter vector to be estimated by MCMC (for detailed estimation procedure see Appendix).

The wordographics covariates enter the model via a hierarchical regression which is given by

$$(6) \quad \omega_w = \gamma_1 GEN_w + \gamma_2 US_w + \gamma_3 STA_w + \gamma_4 CIT_w + \gamma_5 HOT_w + v_w$$

where

$GEN_w$  = Indicator for generic for keyword  $w$ ,

$US_w$  = Indicator for US for keyword  $w$ ,

$STA_w$  = Indicator for state for keyword  $w$ ,

$CIT_w$  = Indicator for city for keyword  $w$ ,

$HOT_w$  = Indicator for hotel or similar lodging-related expression for keyword  $w$ ,

$v_w \sim N(0, V_w)$ ,

and

$\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5$ , and  $V_w$  are parameters to be estimated by MCMC.

## Estimation Results

We estimate our model on the data for April 2004; May and June are set aside for a hold-out test.

Reviewing the paid search campaign on a monthly basis is a reasonable policy. Even niche

keywords can accumulate multiple observations over that time period. Shorter time periods result in a significant number of keywords that do not generate any clicks. Our model is conditional on click for the simple reason that without a click there cannot be a conversion. Without at least one click we cannot estimate a conversion rate and are not able to evaluate the performance of keywords on an individual basis.

We estimate seven separate models on the April data: three ML models ( $ML^{Null}$ ,  $ML^{CV}$ ,  $ML^{CV\&WG}$ ) and four Bayes models ( $Bayes^{Null}$ ,  $Bayes^{CV}$ ,  $Bayes^{CV\&WG}$ , and  $HB^{WG}$ ). We have 8,497 observations (daily) for 301 keywords, resulting in, on average, 28 observations per keyword (some keywords had zero clicks on certain days). Thus, nearly all keywords have observations for each day of the month. The 8,497 observations represent 14,302 clicks. As discussed before, our data can have multiple choice occasions (clicks) per observation. We address this issue when constructing the likelihood (see modeling section). The majority of clicks do not lead to a reservation. We observe 518 reservations (choices), or an average conversion rate of 3.6%.

We start our discussion with model selection. We apply the standard Bayesian procedure of hypothesis testing using Bayes factors (BF). Bayes factors are easily interpreted and do not require nested models. We use a numerical procedure (harmonic mean, Newton and Raftery, 1994) to calculate the log-marginal density based on the MCMC output from model estimation. We report log-marginal densities and the Bayes factors in Table 3. A Bayes factor can be understood as a likelihood ratio test for Bayesian estimation. A very small value ( $<2$ ) suggests no evidence for the superiority of one model, a value above 6 gives very strong evidence for one model vs. the other (see Newton and Raftery (1994) for the full set of values between 0 and 6).

Based on Bayes factors we find very strong evidence in favor of the  $HB^{WG}$  model over the  $Bayes^{CV}$  model (BF: 8.77), the  $Bayes^{Null}$  model (BF: 26.37), the  $ML^{CV}$  model (BF: 46.80) and the  $ML^{Null}$  model (BF: 150.90). We also find that using wordographics as covariates does not

improve the model fit:  $ML^{CV}$  vs.  $ML^{CV\&WG}$  (BF: 23.8) and  $Bayes^{CV}$  vs.  $Bayes^{CV\&WG}$  (BF: 2.05<sup>4</sup>). Thus, the HB Bayes model is clearly favored over the ML and Bayes models.

Our results shed light on two substantive issues. First, keywords are predictive of conversion rates. Differences in conversion rates across keywords can be explained, in part, by the underlying keyword. Thus, when modeling the performance of individual keywords we can harness the predictive power of keywords by appropriately modeling keyword heterogeneity. Our empirical results contradict Chen and He's (2006) theoretic findings. The position of the ad is not the only factor that drives the conversion rate (or the value of the click). Our findings show that the underlying keyword is a key predictor of conversion and a game-theoretic model should reflect the choice of keyword as a second strategic variable next to the choice of position (focus of Chen and He 2006).

Second, covariates are valuable. Observable campaign metrics such as position or CTR that are already used by managers in campaign evaluation are valuable in predicting conversion rates on a keyword level. Covariates allow us to establish a link between conversion rates and the auction which determines the position and CPR. The significant influence these factors have on conversion rates will allow, going forward, to connect our conversion model with an empirical auction model. Using such a combined model should enable the researcher to determine, empirically, which position (and resulting CPR) are optimal for a given keyword. This analysis hinges on a conversion type model like ours to allow the researcher to evaluate the different positions from a conversion rate perspective.

Third, exploring keyword heterogeneity using observable keyword characteristics increases the predictive power of the model. We decompose keywords using common *observable* characteristics to create a consumer demographic type data which we call wordographics. Our approach is flexible and allows, going forward, to include *latent* keyword characteristics. Next to

---

<sup>4</sup> A Bayes Factor of 2.05 provides no evidence that one model should be preferred over another one. As such the  $Bayes^{CV}$  model is the more parsimonious model because uses less parameters.

observable wordographics latent characteristics might also enhance the accuracy of prediction. One source for these is available for advertisers: search engines offer a keyword recommendation engine. This recommendation engine allows advertisers to input keywords from their campaign and get recommendations which other keywords are similar and should be used by the advertiser. The search engines do not make their recommendation system public, so from an advertiser's perspective these are latent similarities. A future study could investigate if these latent characteristics enhance the predictive power of our approach.

Next we discuss the coefficient estimates of the best model, the Bayes /HB models (Table 2). As expected, the intercept (or the "base" conversion rate) is strongly negative, reflecting the fact that the conversion rate (or probability of conversion) is very low. Position also has a negative effect on the conversion rate. Position is measured from 1 (high) to 10 (low), so that a lower position (worse) has a negative effect on conversion rate. An effect of position on conversion rate is somewhat unexpected. We model conversion conditional on click. The consumer is on the company webpage and has left the keyword and the position of the text ad behind. However, our results show that the position of the keyword has an effect on conversion rate. Potentially position could signal potential fit between the offer and the average consumer: a higher position indicates a more attractive keyword *from the firm's* perspective (Chen and He 2006).

Next, we investigate the effect of click-through rate (CTR). We find that CTR has a positive effect on the conversion rate. A higher CTR is a sign of a more attractive keyword *from a consumer's* perspective – i.e., more consumers click on the text ad after searching for the keyword. The attractiveness of the keyword is apparently carrying-through to conversion.

We find that cost is not significant in the Bayes/HB models. In paid search cost is a proxy for position and CTR. Holding the keyword constant, bidding a higher amount (more cost) results in a higher position. Comparing keywords along CTR we find that keywords with higher CTR are also more expensive, holding position constant. This can be explained by competition: a keyword

with higher CTR is a more attractive keyword and competition for that keyword in the auction is more intense. More intense competition leads to a higher cost holding, position constant. When we allow for individual-level differences in the effects of position and CTR, the effect of cost on conversion rate, as expected, is not significant.

Our findings suggest new topics for both quantitative and consumer behavior research. If, in the future, a visitor-centric panel dataset becomes available it can be investigated if and how different keywords attract distinct visitor segments. A potential project could study whether consumers' characteristics and observed search behavior allows us to determine to what extent factors like position or CTR that influence conversion can be explained by consumer heterogeneity. On the behavioral side we could imagine research that focuses on the behavioral constructs that explain why factors such as position carry-over into the final decision to make a reservation.

We now turn to the effect of wordographics. First, we use wordographics as indicator variables in the ML and the Bayes models. Adding wordographics in a non-hierarchical manner does not improve model fit and is rejected by model selection criteria (see above). When we use wordographics in a hierarchical framework (HB<sup>WG</sup> model) to explore keyword heterogeneity four out of the five keyword characteristics are significant and model fit is improved (see Tables 2 and 3). Our approach decomposes the intercept into two parts: one that can be explained by wordographics and a residual of other effects not captured by our model. Wordographics influence conversion rates as follows: a generic keyword has a lower conversion rate than a branded keyword. A branded keyword generally does better on all aspects of paid search due to the potentially more advanced stage of the decision process and a different competitive environment (Rutz and Bucklin 2006). Conversion rate is not significantly different if the search is for a hotel in the US vs. outside the US. If the keyword includes a state or a city the conversion rate is lower. Finally, if a keyword includes hotel or a similar lodging-related phrase the conversion rate is lower.

How important is it to address keyword heterogeneity? We find that, when including wordographics, the sensitivity of conversion to position and CTR decreases. When we simulate the effect of a decrease in position by one (e.g., from 2.3 to 3.3) we find that the Bayes<sup>CV</sup> model predicts that conversion rates, on average, decrease by 24.5% (across all keywords). However, when we include wordographics (HB<sup>WG</sup> model) we find that conversion rates, on average, only decrease by 14.2%. This means that, in the Bayes<sup>CV</sup> model, the covariates incorrectly pick up variance that comes from the underlying keyword and not the covariates. The value of position and CTR is “overstated” in the Bayes<sup>CV</sup> model compared to the HB<sup>WG</sup> model. Thus, if, in our data, wordographics are erroneously excluded from the analysis a resulting campaign management strategy could lead to 1) overinvestment in keywords with high CTR and/or 2) overinvestment in position across keywords.

The wordographics results should not be understood as a toolbox to build keywords. Most keywords necessarily include more than one of the characteristics. The best keyword for the company, according to the model, is the keyword without all the characteristics, i.e., “BRAND NAME.” As can be expected, “BRAND NAME” has the highest conversion rate for the company. Wordographics should be understood as a way to better discriminate among keywords. They allow us to explain more of the observed variance in conversion rates and to better estimate conversion rates. Ultimately, adding wordographics improves our ability to evaluate keyword performance on an individual keyword-level.

So far we have determined model performance on in-sample statistics. However, one goal of our study is to design a method that allows us to improve the performance of a paid search campaign on an individual keyword level going forward. In the next section we introduce a way to utilize the estimated conversion rates to improve the paid search campaign by just measuring performance. We will also test the performance of our measurement-based approach against model-free, easy to implement, managerial approaches to paid search campaign management. In this study we are not providing a method on how to optimize a paid search campaign. This would

entail addressing and modeling the auction. We focus on measuring keyword performance, so that in a future study the impact of the auction on the paid search campaign can be measured.

## **Managing the Paid Search Campaign**

We focus on a first-stage approach on keyword management in this section. Using our findings and the remaining data for May-June 2004 we compare common model-free managerial strategies with our model. Again, we are not focusing on optimizing the paid search campaign by determining the optimal combination of bid amount and position on a keyword level. By measuring performance on a keyword-level alone we are able to discriminate between keywords that are attractive vs. unattractive from a profit perspective without changing the existing bidding strategy. We test whether this “measurement alone” strategy is already outperforming existing managerial approaches that rely on heuristic or aggregate measures.

We employ the various different strategies to generate keyword lists based on the notion of “attractive” vs. “unattractive” keywords (see below). The performance of these different keyword lists is tested in a holdout sample (May-June 2004). Our approach has two parts:

1. *Assess individual keyword performance:* Keep “attractive” keywords and drop “unattractive” keywords. Based on the data from April 2004, we use model-free and model strategies to determine the following:
  - Which keywords are “attractive” and should be kept in the campaign? The definition of “attractive” differs across strategies and is discussed in detail below. If a keyword is “attractive” the company is keeping it in the campaign. The bidding strategy for keywords in May-June will be the same as in April.
  - Which keywords are “unattractive” and should be dropped from the campaign in May-June 2004? If a keyword is “unattractive” the company is not going to bid on it in May-June, therefore effectively removing it from the campaign.

For each strategy  $a$ , potentially different, list of keywords is generated.

2. *Evaluate Holdout-Performance.* Using the data from May-June 2004 we can evaluate the performance of the different strategies in terms of profit. We assume a certain revenue-per-reservation and calculate the profit of the strategy based on
  - How many reservations did the strategy generate?
  - What was the cost-per-reservation (CPR)?

In the next section we discuss how attractive keywords are selected based on the different strategies. We then describe how the performance of the different strategies can be evaluated using a hold-out sample.

### **Generate Keyword List**

What is an “attractive” keyword? With full information we would be able to calculate profit based on contribution margin. However, our dataset does not include this. Without knowledge of the profit margin, we can still compare the performance of different approaches based on cost-per-reservation in the hold-out sample. We only need to select a CPR threshold ( $CPR_{\text{threshold}}$ ) which allows us to discriminate between attractive and unattractive keywords. We can then assess keyword performance and use different strategies to generate lists of the keywords to be kept. We test the performance of these different lists (strategies) based on the data for May-June 2004.

**Model-free Strategies.** We start our description of the campaign management approaches with several model-free, easy-to-implement managerial strategies in current practice:

- *Do Nothing:* Continue with the current keywords. This strategy does not need a performance evaluation on an individual keyword level. However, it does not shed any light on keyword performance. This is the strategy the company used for most of 2004.

- *Face Value*: Take the data from April 2004 at face value. Keep all keywords with  $CPR_{\text{monthly}} \leq CPR_{\text{threshold}}$ . Keywords with zero reservations in April 2004 have a  $CPR_{\text{monthly}}$  of  $\infty$  ( $\text{costs}_w > 0, \text{reservation}_w = 0$ ). This strategy results in retaining a very small number of keywords and has a strong bias against infrequent keywords with no or very few clicks.
- *CTR*: Pick a CTR threshold  $x$  and keep all words with  $CTR \geq x$ . Managing by CTR is popular in practice, in part because CTR is available for most keywords. CTR can be calculated based solely on the data available from search engines. No other data need to be collected. However, the CTR threshold  $x$  is chosen arbitrarily by management. Managing by CTR allows for individual word performance evaluation, but it is not clear whether this strategy is consistent with profit maximization.

We generate an “attractive” keyword list for each of these strategies. The “Do Nothing” strategy keeps all 301 original keywords. The “Face Value” and the “CTR” strategy select, potentially different, subsets of the 301 keywords.

**Model-based Strategies.** We estimate daily conversion probabilities for each keyword  $w$  for April 2004. However, daily conversion probabilities are only the first step in answering the question which keywords should be kept? To answer this question we need one monthly performance measure per keyword: average monthly cost-per-reservation. A monthly measure of conversion rate is necessary to calculate the average monthly cost-per-reservation ( $CPR_w^{\text{monthly}}$ ) for each individual keyword  $w$ . We construct average monthly conversion rate ( $ConvRate_w^{\text{monthly}}$ ) by weighting the daily conversion probability ( $PConv_{w,t}$ ) by the daily number of clicks.  $ConvRate_w^{\text{monthly}}$  is given by

$$ConvRate_w^{\text{monthly}} = \frac{\sum_T clicks_{w,t} PConv_{w,t}}{\sum_T clicks_{w,t}}.$$

Next we calculate average cost-per-click ( $CPC_w^{monthly}$ ) as the total amount spent in April 2004 divided by the total number of clicks generated in April 2004 for each keyword  $w$ .  $CPC_w^{monthly}$  in combination with  $ConvRate_w^{monthly}$  gives us the summary measure  $CPR_w^{monthly}$ :

$$CPR_w^{monthly} = CPC_w^{monthly} / ConvRate_w^{monthly}.$$

Our evaluation of individual keyword performance is subsequently based on  $CPR_w^{monthly}$ .

Our model allows us to assign each keyword an estimated monthly CPR. Without a model the data only allow us to calculate monthly CPR for 84 keywords, the remaining 217 keywords can only be seen as having an infinite monthly CPR. Our models (ML and Bayes/HB) result in different estimated  $CPR_w^{monthly}$ . For each model we rank order the keywords by estimated  $CPR_w^{monthly}$  and keep keywords for which estimated  $CPR_w^{monthly} \leq CPR_{threshold}$ . All other keywords are discarded. This leaves us with different keyword lists, one for each model.

### Evaluating Hold-out Performance

The company decided not to change the keyword list used in April for the remainder of the quarter. This allows us to test the performance of subsets of keywords that have been generated by the different strategies described above in May-June. The performance of the different strategies can be evaluated based on two criteria:

- *Number of Reservations* – How many reservations are generated using the strategy?
- *Cost-per-Reservation* – What is the cost-per-reservation using the strategy?

Again, we do not have a contribution margin and can only estimate profit based on a revenue assumption. However, we are focusing on how well the different strategies can identify attractive keywords. We base our analysis on strategy profit ( $\pi_{strategy}$ ):

$$\pi_{strategy} = \sum_{strategy} reservations * \left[ \frac{revenue}{reservation} - \frac{cost}{reservation} \right].$$

## Empirical Example

We do not have data on CPR thresholds that might be of interest to the company. Based on the average room rate, a CPR in the range of \$30 - \$50 is a reasonable assumption. We tested the models based on CPR thresholds ranging from \$20 - \$60 and find that the performance of the strategies is mostly independent from the CPR threshold. For an illustration, we set \$30 as the  $CPR_{\text{threshold}}$  and discuss the findings based on that assumption.

First we compare the results of the model-free strategies with the best fitting model, the  $HB^{WG}$  model (see Table 4). The “Do Nothing” uses all existing keywords (301) from April. This strategy generates the highest number of reservations. However, it also is the most expensive one in terms of CPR and total cost. The “Face Value” strategy uses the smallest number of keywords (74) as all keywords with no reservations in April are discarded. It generates the lowest CPR (\$5.87) of all strategies, but results in a very low number of reservations. The “CTR” strategy is controversial. It is not clear how a target CTR should be picked. We decided to keep all keywords that have a CTR above the average CTR. The “CTR” strategy uses a number of keywords similar to the best model, however does not discriminate very well between attractive and unattractive keywords and leads to high CPR and total cost. The best model ( $HB^{WG}$ ) generates is very similar to the “CTR” strategy in terms of keywords used and reservations generated. However, the model is significantly better in picking attractive keywords, as total cost and CPR are significantly lower than with the “CTR” strategy. If we compare these different model-free strategies with the best model on profits we find that our best model outperforms the model-free strategies by over \$1,200 in two months. In other words, based on the 150 keywords selected we have an improvement of \$8/keyword over two month or \$48/keyword over the period of a year. Our results show that managing by CTR, a popular strategy, is not profit maximizing and, in our case, even worse than just discarding all keywords with no reservations (“Face Value”).

As a second step we compare a subset of models ( $ML^{CV}$ ,  $Bayes^{CV}$ , and  $HB^{WG}$ ) in detail against each other. Earlier we found that based on model selection criteria, heterogeneity and

wordographics matter when modeling conversion rates. Is the best fitting model ( $HB^{WG}$  model) also the best model in terms of hold-out CPR performance?

We report the holdout-results of the three models in Table 5. Based on profit we again find that the  $HB^{WG}$  model outperforms the other two models. The  $ML^{CV}$  model seems to generally overestimate the conversion rates. It categorizes 173 (out of 301) keywords as attractive. In terms of profit it is even worse than the “Do Nothing” strategy. The  $Bayes^{CV}$  model, on the other hand, is conservative in estimating conversion rates and thus only selects a limited number of keywords (79). It is close to the “Face Value” strategy in terms of keywords selected and reservations generated. In terms of profit, however, the  $Bayes^{CV}$  outperforms all the model-free strategies. The  $HB^{WG}$  model provides a middle ground between the  $ML^{CV}$  and the  $Bayes^{CV}$  model and is able to better discriminate between attractive and unattractive keywords than model-free and all other model strategies.

Summing up, the  $HB^{WG}$  model is the best model based on profit hold-out performance (see Table 6). Compared to the  $Bayes^{CV}$  model it captures an additional 110 reservations with an average CPR of \$21.70, significantly below our assumed contribution margin of \$30. The key driver of the predictive power of the different models in hold-out appears to be addressing keyword heterogeneity. This is consistent with the in-sample findings: the biggest improvement in terms of log-marginal density occurs when we address keyword heterogeneity with a random coefficient approach. We find that, consistent with our in-sample results, test using wordographics in a hierarchical fashion to explore keyword heterogeneity improves the performance of our model in a hold-out.

## **Conclusion and Limitations**

We investigate paid search campaign management on an individual keyword level. We focus our analysis on two related questions: 1) how can we evaluate the performance of individual

keywords and 2) can this performance evaluation be used to improve the campaign on a keyword level? To the best of our knowledge our work is the first empirical investigation in marketing of paid search campaign management on a keyword level. Paid search is the key growth area of online advertising and is the one of the few profitable business models of Web 2.0 so far. Paid search campaigns have become a crucial part of the marketing budget of most firms. We hope that our work can help companies to manage campaigns based on costs and profits on a keyword level.

The performance of a paid search campaign can be evaluated by cost-per-reservation (CPR). However, most keywords do not lead to reservations on a frequent basis. In our sample only 84 keywords out of 301 led to reservations in April 2004. Thus, we can only calculate CPR for 84 keywords. Does that mean that the remaining keywords have a conversion rate of zero and a CPR of infinity? Should the company immediately drop these remaining keywords?

We think not and develop a model address this. Conversion can be conceptualized as a binary choice conditional on click. We estimate daily conversion probabilities for each keyword using a shrinkage approach and similarities across keywords. Differences in conversion probabilities can be explained by keyword performance measures (such as position and CTR) and keyword characteristics, which we call wordographics. We find that when modeling keywords on an individual level we should also incorporate keyword heterogeneity. Summing up, knowing where a visitor to the site comes from (i.e., which keyword was used to initiate the search) helps to predict the probability of conversion. We use the estimated daily conversion probabilities to calculate monthly CPR for each keyword.

We compare popular model-free campaign management strategies against our estimated CPR strategy in a hold-out test. We do not optimize the paid search campaign, i.e., we do not investigate the optimal bid amount on a keyword level. Before optimizing a campaign we need to be able to measure its performance as a first step. We provide a method to measure campaign performance on an individual keyword level, a non-trivial problem due to very sparse conversion

data. In our hold-out test we investigate whether the ability to measure performance alone helps improve the campaign compared with popular, model-free, management approaches.

First we determine which keywords are attractive according to each strategy. We create subsets of the 301 keywords for each strategy by keeping the attractive keywords and dropping the remainder. Based on the data from May-June 2004 we evaluate the performance of the different strategies by CPR and profit. We find that our best model outperforms the model-free strategies in terms profit. Our model based CPR strategy is the only strategy that we are aware of that enables the manager to measure performance of individual keywords and use this information when discriminating among keywords.

We base our model on data that can be made easily available to management. Our strategy can be implemented and the performance of a campaign measured based upon it. The ability to measure allows the manager to test different position/cost combinations and decide based on the measured outcomes which are the optimal ones. The data, however, is also one of the weaknesses of our study. First, we do not have data on competition. However, we know that our company did not change its bidding strategy over the period of the data. When we compare the estimation sample (April 2004) to the hold-out sample (May-June 2004) we do not find evidence for structural breaks in competitors' performance. Key figures such as average position or CPC that would be affected by changing competitive behavior do not change over time. The lack of competitive data also precludes us from modeling the actual auction. Without modeling the actual auction we can not determine bidding strategies. Yet, we are hard-pressed to imagine that such a dataset would be made available. Companies do not have access to their competitors bidding strategies and, from the perspective of the search engines, it seems reasonable to keep this confidential.

Second, we do not have clickstream data and can not model the consumer's choice process. If a clickstream dataset becomes available, we could investigate how consumer search behavior and characteristics, i.e., demographics, influence the conversion decision.

Finally, we do not investigate the role of ad copy and landing page design in paid search. These are two important drivers of paid search performance that have not yet been examined. A future research project could combine empirical and experimental work on a small subset of keywords in which the role of ad copy and landing page could be explored.

## REFERENCES

- Ainslie, A. & Rossi, P.E. (1998). Similarities in Choice Behavior Across Product Categories, *Marketing Science*, 17 (2), 91-106
- Allenby, G & Rossi, P. (2003). Bayesian Statistics and Marketing, *Marketing Science*, 22, 304-328
- Andrews, R., Ainslie, A. & Currim, I.S. (2002). An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity, *Journal of Marketing Research*, Vol. 34, 479-487
- Bhat, S., Bevans, M. & Sengupta, S. (2002). Measuring the Users' Web Activity to Evaluate and Enhance Advertising Effectiveness, *Journal of Advertising*, Vol. 31 (Fall 2002), 97-106
- Bradlow, E.T. & Schmittlein, D.C. (2000). The Little Engines that could: Modeling the Performance of World Wide Web Search Engines, *Marketing Science*, Vol. 19, No.1, 43-62
- Broadbent, S. (1984). Modeling with Ad Stock, *Journal of the Market Research Society*, 16, 295-312
- Burke, R. & Srull, T.K. (1988). Competitive Inference and Consumer Memory for Advertising, *Journal of Consumer Research*, 15, 55-67
- BusinessWeek (2006), The Counterattack on Google, May 8
- Chatterjee, P., Hoffman, D.L. & Novak, T.P. (2003). Modeling the Clickstream: Implications for Web-Based Advertising Efforts, *Marketing Science*, 22(4), 520-541
- Chen, Y. & He, C. (2006). Paid Placement: Advertising and Search on the Internet, *Working Paper*
- Cho, C & Cheon, J. (2004). Why do People avoid Advertising on the Internet?, *Journal of Advertising*, Vol.33 (Winter 2004), 89-97
- Dahlen, M. (2001). Banner Advertisements through a New Lens, *Journal of Advertising Research*, August 2001, 23-30
- Dahlen, M., Rasch, A. & Rosengren, S. (2003). Love at first Sight? A Study of Website Advertising Effectiveness, *Journal of Advertising Research*, March 2003, 25-33
- Danaher, P. & Mullarkey, G. (2003). Factors affecting Online Advertising Recall: A Study of Students, *Journal of Advertising Research*, September 2003, 252-267
- Dreze, X. & Hussherr, F-X. (2004). Internet Advertising: Is Anybody Watching?, *Journal of Interactive Marketing*, Vol.17 (Autumn 2003), 8-23
- Edelman, B. & Ostrovsky, M. (2007). Strategic Bidder Behavior in Sponsored Search Auctions, *Decision Support Systems*, v. 43(1), February 2007, pp. 192-198

- Edelman, B., Ostrovsky, M. & Schwarz, M. (2007). Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords, *American Economic Review*, forthcoming March 2007
- Gelfand, A.E. & Smith, A.F.M. (1990). Sampling-based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Associations*, 85, 997-985
- Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian Data Analysis*, Chapman&Hall/CRC, Boca Raton, FL
- Goodman, A. (2005). *Winning Results with Google AdWords*, McGraw-Hill/Osborne, Emeryville, CA
- Havlena, W. & Graham, J. (2004). Decay Effects in Online Advertising: Quantifying the Impact of Time since last Exposure on Branding Effectiveness, *Journal of Advertising Research*, December 2004, 327-332
- Hofacker, C. & Murphy, J. (2000). Clickable World Wide Web Banner ads and Content Sites, *Journal of Interactive Marketing*, Vol. 14 (Winter 2000), 49-59
- Hoffman, D., Kalsbeek, W. & Novak, T. (1996). Internet and Web Use in the US, *Communications of the ACM*, Vol. 39 (December 1996), 36-46
- Jansen, B.J. & Molina, P.R. (2005). The Effectiveness of Web Search Engines for Retrieving Relevant Ecommerce Links, *Information Processing and Management*, 42, 1075–1098
- Janiszewski, C. (1993). Preattentive Mere Exposure Effects, *Journal of Consumer Research*, 20, 376-392
- Johnson, E. J., Moe, W., Fader, P., Bellman, S. & Lohse, G. (2004). On the Depth and the Dynamics of Online Search Behavior, *Management Science*, Vol. 50, No.3, 299-308
- Lynch, J.G. Jr. & Ariely, D. (2000). Wine Online: Search costs affect competition an price, quality and distribution, *Marketing Science*, Vol.19, No.1, 83-103
- Manchanda, P., Dube, J.-P., Goh, Y. & Chintagunta, P. (2006). The Effects of Banner Advertising on Internet Purchasing, *Journal of Marketing Research*, forthcoming
- McCulloch, R.E., Rossi, P.E. & Allenby, G.M. (1996). The Value of Purchase History Data in Target Marketing. *Marketing Science*, 15(4):321–340
- Montgomery, A., Li, S., Srinivasan, K. & Liechty, J.C. (2004). Modeling online browsing and path analysis using Clickstream data, *Marketing Science*, Vol. 23 (Fall 2004), 579-595
- Moore, S., Stammerjohan, C. & Coulter, R. (2005). Banner Advertiser-Web Site Congruity and Color Effects on Attention and Attitudes, *Journal of Advertising*, Vol. 34, 71-84
- Morrison, D. G. & Schmittlein, D.C. (1981), Predicting Future Random Events based on Past Performance, *Management Science*, Vol. 27, No.9

- Novak, T.P. & D.L. Hoffman (2000), "Advertising and Pricing Models for the Web," in *Internet Publishing and Beyond*, Cambridge: MIT Press
- Ratchford, B. T., Lee, M. & Talukdar, D. (2003). The impact of the internet on information search for automobiles, *Journal of Marketing Research*, Vol. XL (May 2003), 193-209
- Rutz, O.J. & Bucklin, R. E. (2006), From Generic to Branded: A Model of Spillover Dynamics in Paid Search Advertising, *Working Paper*
- Schlosser, A., Shavitt, S. & Kanfer, A. (1999). Survey of Internet Users' Attitudes toward Internet Advertising, *Journal of Interactive Marketing*, Vol. 13 (Summer 1999), 34-54
- Schmittlein, D. C., Bemmaor, A. C. & Morrison, D. G. (1985), Why does the NBD Model work? Robustness in Representing Product Purchases, Brand Purchases and Imperfectly Recorded Purchases, *Marketing Science*, Vol. 4, No.3
- Shen, F. (2002). Banner Advertisement Pricing, Measurement, and Pretesting Practices: Perspectives from Interactive Agencies, *Journal of Advertising*, Vol. 31 (Fall 2002), 59-67
- Spiteri, L.F. (2000). Access to Electronic Commerce Sites on the World Wide Web: An Analysis of the Effectiveness of Six Internet Search Engines, *Journal of Information Science*, 26 (3) 2000, pp. 173–183
- Steenburgh, T.J., Ainslie, A. & Engebretson, P.H. (2003). Massively Categorical Variables: Revealing the Information in ZIP Codes, *Marketing Science*, Vol. 22, No. 1, Winter 2003
- Telang, R., Boatwright, P. & Mukhopadhyay, T. (2004). A Mixture Model for Internet Search Engine Visits, *Journal of Marketing Research*, Vol. XLI (May 2004), 206-214

Table 1: Sample Statistics

	Impressions	Clicks	Reservations	Cost	Average Position
<b>April 04</b>	2,281,023	14,302	518	\$ 5,107	6.0
<b>May-June 04</b>	2,983,085	38,878	1,348	\$ 12,548	6.3

	Conversion Rate	Cost/Click	Cost/Reservation
<b>April 04</b>	3.62%	\$ 0.36	\$ 9.86
<b>May-June 04</b>	3.47%	\$ 0.32	\$ 9.31

Table 2: Estimation Results

			Coefficient Estimates		
			Bayes <sup>Null</sup>	Bayes <sup>CV</sup>	HB <sup>WG</sup>
<b>Performance</b>	<i>Intercept</i>	$\beta_0$	-4.22 (-4.39, -4.06) <sup>1</sup>	-4.10 (-4.34, -3.86)	-2.32 (-2.91, -1.93)
	<i>Position</i>	$\beta_1$		-0.29 (-0.37, -0.20)	-0.16 (-0.28, -0.03)
	<i>CTR</i>	$\beta_2$		1.57 (1.20, 1.94)	0.49 (0.14, 1.16)
	<i>Cost</i>	$\beta_3$		- <sup>2</sup>	-
<b>Wordo-graphics</b>	<i>Generic</i>	$\gamma_1$			-0.72 (-1.30, -0.27)
	<i>US</i>	$\gamma_2$			-
	<i>State</i>	$\gamma_3$			-1.59 (-1.92, -1.20)
	<i>City</i>	$\gamma_4$			-0.76 (-1.08, -0.43)
	<i>Hotel</i>	$\gamma_5$			-0.93 (-1.24, -0.68)

<sup>1</sup> 95% coverage interval

<sup>2</sup> We use Bayes factors for model selection. We only report the best model and exclude any non-significant covariates.

Table 3: Model Selection

	Information		Fit		Log-Bayes Factor
	Covariates <sup>1</sup>	Wordographics <sup>2</sup>	Log-Likelihood	Log-Marginal Density	HB Bayes vs. ...
<b>ML<sup>Null</sup></b>			-2,227.39	-2,231.90 <sup>3</sup>	150.90
<b>ML<sup>CV</sup></b>	✓		-2,109.22	-2,127.80 <sup>3</sup>	46.80
<b>ML<sup>CV&amp;WG</sup></b>	✓	✓	-2,106.33	-2,151.60 <sup>3</sup>	115.84
-----					
<b>Bayes<sup>Null</sup></b>				-2,107.37	26.37
<b>Bayes<sup>CV</sup></b>	✓			-2,089.77	8.77
<b>Bayes<sup>CV&amp;WG</sup></b>	✓	✓		-2,091.82	10.82
-----					
<b>HB<sup>WG</sup></b>	✓	✓		-2,081.00	

<sup>1</sup> Covariates information include position, click-through rate (CTR) and Cost.

<sup>2</sup> Wordographics are keyword characteristics and include indicators for generic, US, state, city and hotel.

<sup>3</sup> We use BIC as an approximation of Log-Marginal Density.

Table 4: Hold-out Performance: Model-free vs. Model Strategies

	# Key-words	# Reservations	Cost	CPR	Profit
<b>Do Nothing</b>					<b>\$ 27,891.36</b>
<i>All Keywords selected</i>	301	1,348	\$ 12,548.64	\$ 9.31	
<b>Face Value</b>					<b>\$ 27,239.11</b>
<i>Keywords selected</i>	74	1129	\$ 6,630.89	\$ 5.87	
<i>Keywords not selected</i>	227	219	\$ 5,917.75	\$ 27.02	
<b>CTR</b>					<b>\$ 27,879.12</b>
<i>Keywords selected</i>	158	1,242	\$ 9,607.47	\$ 7.55	
<i>Keywords not selected</i>	143	106	\$ 2,941.17	\$ 38.70	
<b>Best Model</b>					<b>\$ 29,063.95</b>
<i>Keywords selected</i>	150	1,238	\$ 8,076.05	\$ 6.52	
<i>Keywords not selected</i>	151	110	\$ 4,472.59	\$ 40.66	

Table 5: Hold-out Performance: Comparison of Model Strategies

	# Key-words	# Reser-vations	Cost	CPR	Profit
<b>Do Nothing</b>					<b>\$ 27,891.36</b>
<i>All Keywords selected</i>	301	1,348	\$ 12,548.64	\$ 9.31	
<b>ML<sup>CV</sup></b>					<b>\$ 27,859.00</b>
<i>Keywords selected</i>	173	1,228	\$ 8,981.00	\$ 7.31	
<i>Keywords not selected</i>	128	120	\$ 3,567.64	\$ 29.73	
<b>Bayes<sup>CV</sup></b>					<b>\$ 28,150.64</b>
<i>Keywords selected</i>	79	1,128	\$ 5,689.36	\$ 5.04	
<i>Keywords not selected</i>	222	220	\$ 6,859.28	\$ 31.18	
<b>HB<sup>WG</sup></b>					<b>\$ 29,063.95</b>
<i>Keywords selected</i>	150	1,238	\$ 8,076.05	\$ 6.52	
<i>Keywords not selected</i>	151	110	\$ 4,472.59	\$ 40.66	

Table 6: Hold-out Performance: Profit

		Profit
<b>Model-free</b>	<b>Do Nothing</b>	\$ 27,891.36
	<b>Face Value</b>	\$ 27,239.11
	<b>CTR</b>	\$ 27,879.12
<b>Model-based</b>	<b>ML<sup>CV</sup></b>	\$ 27,859.00
	<b>Bayes<sup>CV</sup></b>	\$ 28,150.64
	<b>HB<sup>WG</sup></b>	<b>\$ 29,063.95</b>

Figure 1: Search Results Page Example for Google

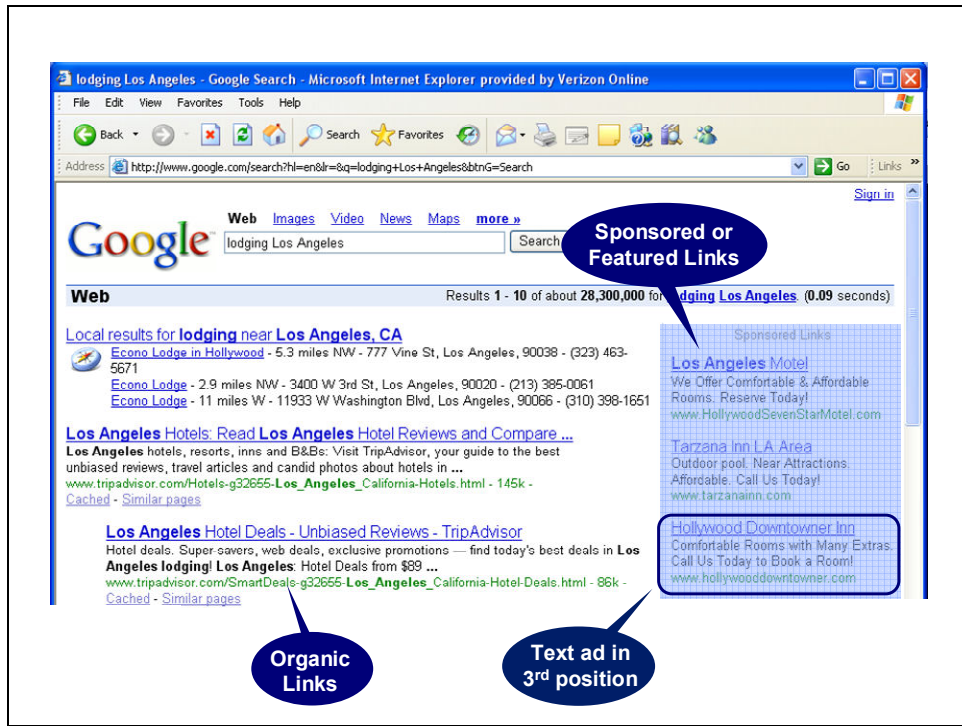


Figure 3: Data Sparseness

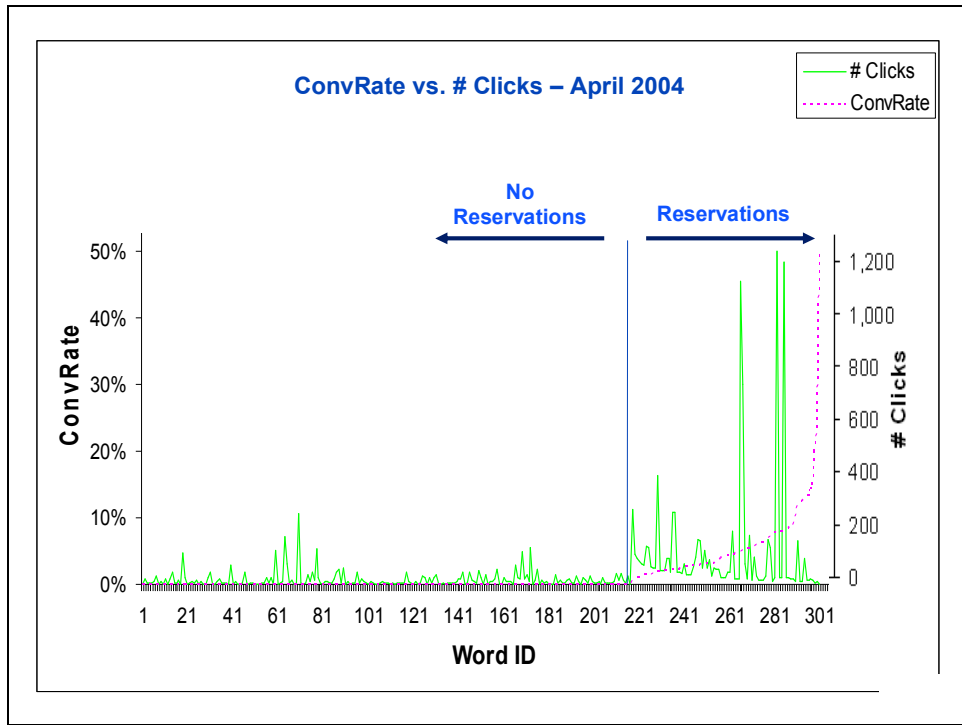


Figure 3: Paid Search from the Company's and Consumer's Perspective

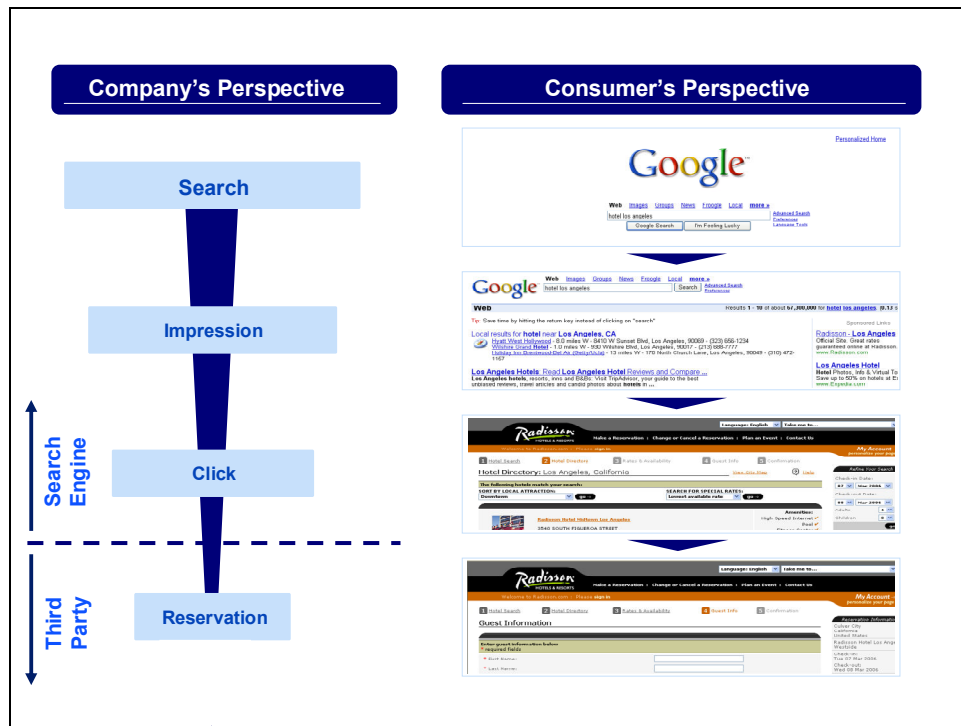
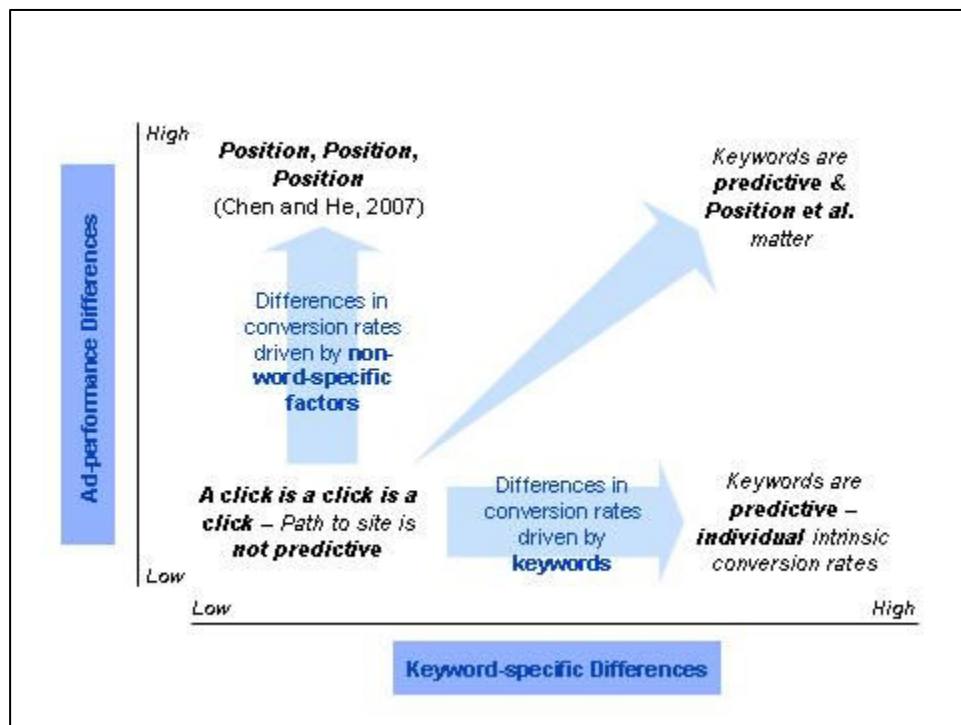


Figure 4: Substantive Issues – Modeling Strategy



## Appendix: Bayesian Estimation

We include keyword characteristics in a hierarchical fashion following Steenburgh et al. (2003).

$$(1) \quad v_{w,t} = X_{w,t}^T \beta_w + \omega_w + \varepsilon_{w,t} \quad \text{for } w = 1, \dots, n \text{ keywords}$$

$$(2) \quad \omega_w = C_w \gamma + v_w \quad \text{for } w = 1, \dots, n \text{ keywords}$$

where

$X_{w,t}^T$  is a vector consisting of an intercept and  $m-1$  keyword measures that can change over time (position, click-through rate and cost)

$\beta_w$  is a vector of coefficients to be estimated

$\omega_w$  is the keyword effect

$C_w$  is a matrix of  $c$  keyword characteristics or wordographics that do not change over time (indicators for generic, US, state, city and hotel or a similar lodging related phrase)

$\gamma$  is a vector of coefficients to be estimated

$\varepsilon_{w,t}$  = Logit error for keyword  $w$  at time  $t$

$v_w \sim N(0, V_w)$ .

For identification purposes there can be no intercept in  $C_w$ .

### Prior Distributions

$$1) \quad \beta \sim N_m(\mu_\beta, V_\beta), \text{ with hyper-priors } \mu_\beta \sim N(\mu_b, V_b) \text{ and } V_\beta \sim \text{Wishard}(v_1, v_2).$$

Standard values apply for parameters of hyper-priors.

$$2) \quad \gamma \sim N_c(\mu_\gamma, V_\gamma), \text{ where } \mu_\gamma = 0_c \text{ and } V_\gamma = 10^6 I_c.$$

$$3) \quad V_w \sim \text{gamma}(k/2, m/2), \text{ where } k = 4 \text{ and } m = 4.$$

## Sampler

- 1) Draw  $\beta_w^{new}$  and use Metropolis-Hastings step to accept/reject  $\beta_w^{new}$  for all w.
- 2) Draw  $\omega_w^{new}$  and use Metropolis-Hastings step to accept/reject  $\omega_w^{new}$  for all w.
- 3) Draw new  $\mu_\beta$  and  $V_\beta$  using Gibbs sampling
  - $\mu_\beta \sim N(\tilde{\mu}_\beta, \tilde{V}_\beta)$ , where  $\tilde{V}_\beta = [nV_\beta^{-1} + V_b^{-1}]^{-1}$  and  $\tilde{\mu}_\beta = \tilde{V}_\beta [nV_\beta^{-1}\bar{\beta} + V_b b]$
  - $V_\beta^{-1} \sim Wishart\left(v_1 + n, \left[v_2 + \sum_{w=1}^n (\beta_w - \mu_b)(\beta_w - \mu_b)'\right]^{-1}\right)$ .
- 4) Draw new  $\gamma$  and  $V_w$  using Gibbs sampling
  - $\gamma \sim N(\tilde{\mu}_\gamma, \tilde{V}_\gamma)$ , where  $\tilde{V}_\gamma = [C'V_w^{-1}C + V_\gamma^{-1}]^{-1}$  and  
 $\mu_g = \tilde{V}_\gamma [C'V_w^{-1}\omega_w + V_\gamma\mu_\gamma]$ .
  - $V_\gamma^{-1} \sim gamma\left(\frac{n+k}{2}, \frac{m + (\omega_w - C\gamma)'(\omega_w - C\gamma)}{2}\right)$ .

Steps 2 and 4 do not apply for the Bayes model without wordographics.